Solvation Thermodynamics and Free Energy Surfaces of Intrinsically Disordered Proteins (IDPs) in Aqueous Solutions

by

Sthitadhi Maiti

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2024 by the
Graduate Supervisory Committee:

Matthias Heyden, Chair
S. Banu Ozkan
Petr Šulc

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Contrary to the traditional structure-function paradigm for proteins, intrinsically disordered proteins (IDPs) and regions (IDRs) are highly disordered sequences that lack a fixed crystal structure yet perform various biological activities such as cell signaling, regulation, and recognition. The interactions of these disordered regions with water molecules are essential in the conformational distribution. Hence, exploring their solvation thermodynamics is crucial for understanding their functions, which are challenging to study experimentally. In this thesis, classical Molecular Dynamics (MD), 3D-Two Phase Thermodynamics (3D-2PT), and umbrella sampling have been employed to gain insights into the behaviors of intrinsically disordered proteins (IDPs) and water.

In the first project, local and total solvation thermodynamics around the K-18 domain of the intrinsically disordered protein Tau were compared, and simulated with four pairs of modified and standard force fields. In empirical force fields, an imbalance between intramolecular protein interactions and protein-water interactions often leads to collapsed IDP structures in simulations. To counter this, various methods have been devised to refine protein-water interaction models. This research applied both standard and adapted force fields in simulations, scrutinizing the effects of each adjustment on solvation free energy.

In the second project, the MD-based 3D-2PT analysis was utilized to examine variations in local entropy and number density of bulk water in response to an electric field, focusing on the vicinity of reference water molecules.

In the third project, various peptide sequences were examined to quantify the free energy involved when specific sequences, known as alpha-MoRFs (alpha-Molecular Recognition Features), transition from intrinsically disordered states to structured secondary motifs like the alpha-helix. The low folding free energy penalty of these sequences can be exploited to design peptide-based or small-molecule drugs. Upon binding to alpha-MoRFs, these drugs can stabilize the helix structure through a binding-induced folding mechanism.

Alpha-MoRFs were juxtaposed with entirely disordered sequences from known proteins, with findings benchmarked against leading structure prediction models. Additionally, the binding free energies of various alpha-MoRFs in their folded conformation were assessed to discern if experimental binding free energies reflect the separate contributions of folding and binding, as obtained from umbrella sampling simulations.

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my academic advisor, Dr. Matthias Heyden, for his support, kindness, and exemplary leadership during my Ph.D. study. His extended knowledge and passion for science have always encouraged me in my academic research and daily life. My gratitude extends to Dr. Banu Ozkan and Dr. Petr Sulc for being a part of my Ph.D. journey and providing me with plenty of valuable advice as a part of my graduate committee.

My deepest thanks go to all the colleagues with whom I've had the honor of working over the last five years. Their contributions have been instrumental to my journey, particularly the members of Dr. Heyden's group: Brandon Neff, Michael Sauer, Aleksandar Lazaric, Holly Hemsath, Edgar Manriquez, Sylvia Adomako, and Heekun Cho. A profound debt of gratitude is owed to Dr. Souvik Mondal, the present postdoctoral researcher in the Heyden group and a cherished friend, for his unwavering assistance and sage counsel during times of challenge. I must also express my substantial gratitude to Dr. Viren Pattni for his extensive guidance on essential programming concepts during the initial few years of my studies.

I would also like to thank everyone else who has helped me along the way. I would like to extend my heartfelt appreciation to my roommate Jyotirish Das, whose support during challenging times has been invaluable. I extend my gratitude to my dear friend Julian Danciu, from whom I have learned important coding skills that have been crucial to my progress.

I would like to express my gratitude to David Nutt, Graduate Programs Coordinator at Arizona State University, for his gracious aid in my application process and the smooth transition to ASU.

I would love to thank my family and friends for supporting me all these years in school. In particular, I would like to extend my gratitude to my mother, Kumkum Maiti, whose

boundless support and empathy have been pivotal in my pursuit of personal and academic achievements. My dad has been a constant inspiration for me to achieve my goals. His advice and the invaluable life lessons he imparted have been instrumental in sculpting my path to success. In addition, I would like to give a sincere appreciation to my friends Bhaskar Rana, Saayak Halder, Deblin Jana, Kiron K. Ghosh, Amritendu S. Koley, Saumyadip Nemu, Aniruddha Chattaraj, Indranil Samanta and a few others for never giving up on me. They have constantly given me encouragement, and support during my difficult years of life and my academic journey.

I wish to convey my profound gratitude to my extended family—my cousins and relatives, both near and far—for their unwavering support and endless optimism, which have been a constant source of strength throughout my academic pursuits. Their belief in my abilities and their words of encouragement have been a beacon of hope, guiding me through challenging times and celebrating with me in moments of triumph. The network of support they have provided has been indispensable, not only in my career but in shaping the person I have become. Their influence extends beyond occasional gestures; it is the accumulation of countless acts of kindness, motivational conversations, and genuine enthusiasm for my success that I carry with me every day. I am eternally thankful for their love, which has been as much a part of my professional journey as any academic achievement.

Finally, I would like to acknowledge the assistance of the AI tool ChatGPT-4.0. It was used solely for language editing, proofreading, and providing suggestions for this manuscript (Chapters 1-4, 7, & 8), and not for content creation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRINSICALLY DISORDERED PROTEINS AND SOLVATION THERMODYNAMICS

1.1 Intrinsically Disordered Proteins/Regions

Proteins are essential biomolecules that play critical roles in various biological processes. They are the foundation of life, giving structure to cells and organs and facilitating communication between various parts of the body. Until the dawn of the 21st century, it was widely believed that proteins could only perform their biological functions when properly folded into specific structures [6]. This notion was grounded in the central dogma of molecular biology [7, 8], which posits that DNA encodes genetic information, which is then transcribed into messenger RNA and translated into an amino acid sequence. This sequence is thought to fold into a functional protein. The biophysical mechanisms that dictate the folding of an amino acid sequence into a precise three-dimensional protein structure remain incompletely understood [6]. For over a century, the conventional understanding of protein structure has been widely accepted in the scientific community. Despite this established paradigm, the modern era of protein science, particularly during the genomic era that began at the end of the 20th century, has revealed the existence of intrinsically dynamic, flexible, and biologically active proteins. With access to complete genome sequences [9], scientists observed that some amino acid sequences did not conform to the expectation of folding into globular protein structures [10]. Concurrently, experimental research started to identify examples of essential proteins and domains that were partially structured or entirely disordered in solution but still retained their biological functionality [9]. As a result,

the term ”intrinsically disordered proteins” was introduced at the turn of the century [10]. The significant hypothesis put forward was that intrinsically disordered proteins (IDPs) are not rare exceptions but rather a distinct and extensive class of proteins [10]. The contemporary understanding of disorder encompasses protein regions or entire proteins that, while biologically active, exhibit dynamic flexibility in their conformational ensembles at the secondary and/or tertiary structure levels [11–13]. Numerous proteins, alongside IDPs, possess significant intrinsically disordered regions (IDRs) within their structured domains, crucial for their biological roles [14, 15]. IDPs and intrinsically disordered protein regions (IDPRs or IDRs) exist as dynamic ensembles, often likened to ”protein clouds,” in which the positions of atoms and backbone Ramachandran angles fluctuate significantly over time, without settling into specific equilibrium values [16–18]. IDPs are characterized by a distinctive amino acid composition, with a notable bias towards a greater fraction of polar and charged residues and a relatively lower presence of hydrophobic and aromatic residues [18–20]. The DisProt database, which continuously updates its collection of identified IDPs and IDPRs, currently houses about 2,900 unambiguous IDPs and approximately 6,560 IDPRs, reflecting the growing recognition of these protein types [21].

1.2 Functions of IDPs/IDRs

By the mid-1990s, both experimental studies on regulatory proteins and bioinformatics analyses of newly available genome sequences revealed that disordered regions are prevalent in eukaryotic proteins, highlighting the abundance and functional importance of IDPs in these organisms [22, 23]. In reality, about 33% of the eukaryotic proteome and 40% of the human proteome are thought to be disordered or to possess extensive disordered regions (more than 100 residues), with a larger proportion featuring disordered segments of up to 30 residues [24–28]. It is common to find intrinsically disordered regions (IDRs) exceeding 50 amino acids in length within functional proteins [29]. The presence of functional

intrinsically disordered proteins (IDPs), like polypeptide hormones, has been recognized for years, and IDPs have been identified in intact cells using nuclear magnetic resonance (NMR) experiments [30]. Functions of IDPs encompass the regulation of transcription and translation [31], cellular signal transduction [23],molecular recognition[32], storage of small molecules, protein phosphorylation, and the regulation of self-assembly processes, such as in ribosomes and bacterial flagella [33]. IDRs can act as molecular chaperones for RNA and other proteins. Their function involves binding to either misfolded or properly folded proteins or RNA molecules [34]. This binding capability is related to their role as recognition elements and their potential to unfold or loosen intermediates that are kinetically trapped in the folding process. The biological function of certain intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) is associated with folding or partial folding, particularly upon binding to precise targets [35, 36]. However, it has been observed that even fully unfolded IDPs can form high-affinity complexes with their binding partners [37].

1.3 Importance of IDP/IDR Solvation Effects

IDPs/IDRs are implicated in various diseases, including cancer, amyloidosis, and neurodegenerative disorders [38–43]. To understand the pathological behavior of IDPs, such as the propensity for liquid-liquid phase separation [44–47], dysregulated binding [23], or aggregation [39], it is essential to explore the relationship between their sequence, function, and characteristic features. To understand this relationship, it is essential to determine how the conformational ensembles of intrinsically disordered proteins in solution are affected by their interactions with the surrounding molecular environment. In principle, the exploration of the structural and dynamic properties of conformational ensembles in IDPs can be characterized through techniques such as small-angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR), and Förster resonance energy transfer (FRET) experi-

ments [48–50]. Concurrently, atomistic molecular dynamics (MD) simulations provide a crucial tool for gaining a detailed microscopic insight into these properties [51–53]. Standard empirical force field models used in molecular dynamics (MD) simulations often inaccurately predict overly compact conformations for solvated intrinsically disordered proteins (IDPs) [51, 52, 54]. To address this issue, IDP-specific protein force field parameters have been developed [55]. These parameters adjust the weighting of dihedral angles in the simulation, modifying the predicted conformational ensembles of IDPs to better align with experimental observations. The limitations of standard force field simulations in accurately describing intrinsically disordered proteins (IDPs) are often attributed to an imbalance between non-covalent intraprotein interactions, which favor compact states, and protein-solvent interactions, which favor extended conformations with larger solvent-accessible surface areas [51, 52, 54]. This is shown in Fig. 1.1

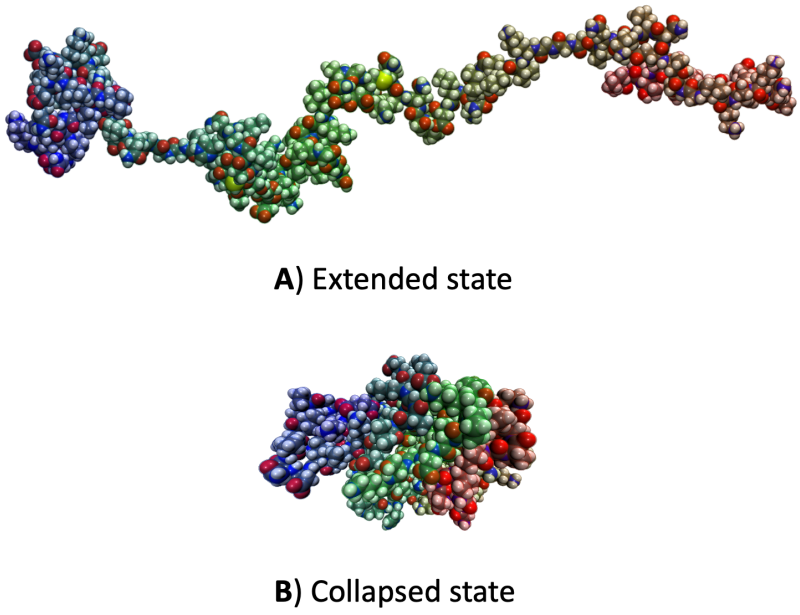

Figure 1.1: Extended and collapsed forms of the K-18 Tau IDP from its ensemble of conformations generated by a 100 ns atomistic, classical MD simulation with Amber03ws-TIP4P/2005 combination of force field-water model. A is the extended state while B is the compact or collapsed structure

To design new force fields with enhanced protein-water interactions, or to distinguish

between them (discussed in Ch. 5), it is essential to address the topic of structure of water, water models and hydration/solvention thermodynamics in the following sections.

1.4 Structure of Water

Being vital for life, water is essential for the structural integrity, dynamic behavior, energetic stability, and overall functionality of biomolecules. Its structure is determined by the sp3 hybridized oxygen atoms, creating a tetrahedral coordination geometry. This configuration allows for the preferred coordination with four other water molecules, a characteristic observed in both the solid and liquid phases of water [56]. Hydrogen bonds between water molecules create a 3D network structure responsible for water's unique properties, including its density maximum at 4°C [56]. In ice, this network is tetrahedral, similar to diamond, but in liquid water, higher temperatures cause some hydrogen bonds to break, disrupting the perfect tetrahedral arrangement. Numerous experiments over the past few decades have used hydrogen/deuterium substitution and neutron diffraction to study water's structure [57]. Combined with X-ray diffraction data, these methods have provided a more precise characterization of water, revealing a hydrogen bond length of 2.8 Å in liquid water [58]. The strength of these bonds is 20-25 kJ/mol, with a lifetime of about 1 ps [56].

Soper *et al*. [57] utilized data from diffraction experiments on liquid water to develop empirical potentials, enabling detailed computer simulations to further investigate water properties. Throughout the years, numerous water models have been created for molecular dynamics simulations, where each individual water molecule can be depicted as either rigid or flexible.

Rigid water models used typically include TIP3P (transferable intermolecular potential 3P) [59], SPC (simple point charge) [60], and SPC/E (extended simple point charge) [61]. These models employ the SHAKE constraint algorithm [62–64] to maintain fixed OH and HH bond lengths, ensuring a constant HOH angle. The GROMACS software package [65],

which employs the SETTLE algorithm [63] for analytically solving constraints in water molecules, was used for all simulations. The rigid water models utilized lack bonded potentials and are defined solely by non-bonded pair potentials, comprising Lennard-Jones (LJ) and Coulombic terms. These water models are characterized by non-bonded pair potentials consisting of Lennard-Jones (LJ) and Coulombic terms, as detailed in Eq. 1.1. Each model has three interaction sites, mirroring the three atoms of a water molecule, with the oxygen atom's site having additional LJ parameters alongside a point charge that is common to all sites. The primary distinctions between the models are their geometrical configurations and their specific LJ and electrostatic parameters, which are unique to each model. [64] The potential for these models is given by [59]:

$$E_{\mathrm{cd}} = \sum_{i}^{on\ c} \sum_{j}^{on\ d} \frac{q_i q_j e^2}{\mathbf{r_{ij}}} + \frac{A}{\mathbf{r_{oo}^{12}}} + \frac{B}{\mathbf{r_{oo}^{6}}} \tag{1.1}$$

The equation involves partial charges q_i and q_j, representing the charge on each atom or interaction site. The term $\mathbf{r_{ij}}$ denotes the distance between any two atoms or charged sites. Additionally, A and B are parameters in the Lennard-Jones potential, and $\mathbf{r_{oo}}$ signifies the distance between oxygen atom sites. Apart from the SPC model, which adopts the ideal tetrahedral HOH angle of 109.5° instead of the actual 104.5°, the rigid structure of these models closely mirrors the geometry of a real water molecule. Additionally, the SPC/E model introduces a polarization correction to the potential energy function outlined in Eq. 1.1:

$$E_{\mathrm{pol}} = \frac{1}{2} \sum_{i} \frac{(\mu - \mu^0)^2}{\alpha_i} \tag{1.2}$$

In this equation, μ represents the electric dipole moment of the effectively polarized water molecule, μ_0 is an isolated water molecule's dipole moment, and α_i is the isotropic polarizability constant. The modification in the SPC/E model leads to better predictions for the density and diffusion constant of liquid water at room temperature and ambient pressure (1 bar) compared to the SPC model. In much of my research, I utilized 3-point water mod-

els like TIP3P and TIPS3P [66], the latter being used alongside the CHARMM36 force field [67], as well as 4-point water models such as TIP4P/2005 [68] and their modifications, which will be discussed in more detail shortly. TIPS3P is a slight modification of the standard TIP3P, with the addition of Lennard-Jones sites on the hydrogen atoms in addition to the oxygen atom. There exists a flexible variant of the SPC model, which was developed by Toukan *et al* [69]. It permits anharmonic stretching of OH bonds. This model is noted for providing the most accurate representation of water's density and dielectric permittivity among non-polarizable three-point water models [70].

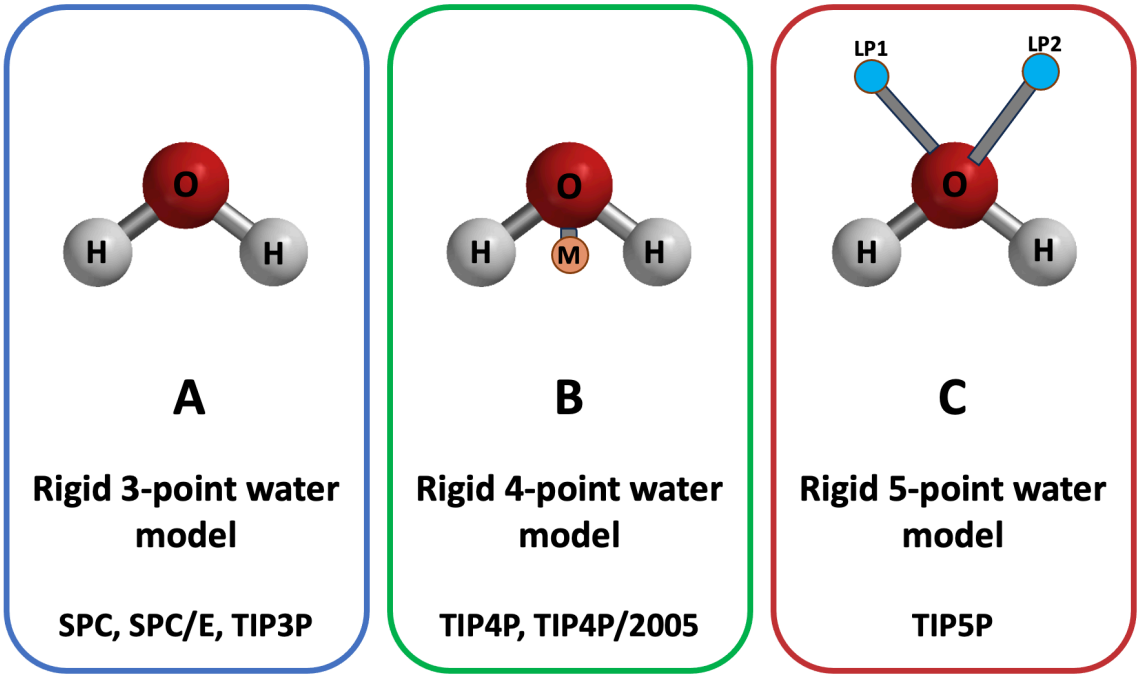

Figure 1.2: Depiction of three main types of rigid water model structures. The oxygen atoms are colored red, and the hydrogen atoms are white. In the 4-point models, the offset partial charge on oxygen, denoted as M, is colored orange. The lone pairs in the 5-point model, referred to as LPs, are colored cyan.

Here in Fig. 1.2, all standard rigid water models are represented. The three-site water models are the most commonly used, but there are also models with four, five, or six interaction sites initially developed to explore water/ice systems [71]. The first four-site model was the Bernal–Fowler model introduced in 1933 [72]. However, the TIP4P model [59]

and its variants, such as TIP4P-D [73] and TIP4P/2005 [68], are more frequently used in popular force fields like AMBER [74–77].

Similar to 3-point water model TIP3P, TIP4P is rigid, with the addition of an extra site, often called the ”M site” or ”dummy site,” to represent the negative charge distribution near the oxygen atom (Fig. 1.2). This model maintains three physical sites for the two hydrogen atoms and the oxygen atom, but the negative charge (electrostatic point) is positioned on the dummy atom M along the bisector of the HOH angle, slightly distant from the oxygen atom. The M site's inclusion is intended to more accurately reflect the tetrahedral arrangement of water molecules in ice and liquid water. The TIP4P model was reparametrized to align with modern molecular dynamics settings and better replicate experimental properties, leading to the development of the TIP4P/2005 variant. This rigid four-site model, with three fixed point charges and one Lennard-Jones center on the O atom, offers enhanced accuracy in representing thermodynamic properties, phase diagrams, melting and vaporization properties, dielectric constant, pair distribution function, and self-diffusion coefficient. [68].

The TIP4P-D model, developed by Piana *et al*. [73], is a rigid four-site water model that builds on the geometry of the TIP4P/2005 model. It enhances the dispersion coefficient C6 of the oxygen atom by 50%. This adjustment considerably enhances the alignment between simulations of intrinsically disordered proteins and experimental findings [78].

The TIP5P model [79] enhances the representation of water's charge distribution by including two sites for the lone pairs of electrons on the oxygen atom. This five-site model aims to more accurately depict hydrogen bonding and angular distribution in liquid water. Compared to earlier models, TIP5P improves the geometry of the water dimer, provides a more ”tetrahedral” water structure, and better matches experimental data, including the temperature of maximal density of water.

1.5 Hydration / Solvation Thermodynamics

Water plays a critical role in biomolecular reactions, acting as both a solvent and a participant in biochemical processes [80]. Its unique properties, such as high polarity, the ability to form hydrogen bonds, and a high dielectric constant, make it an ideal medium for the solvation and interaction of biomolecules. Water facilitates the folding of proteins into their functional three-dimensional structures, stabilizes the structures of nucleic acids, and participates directly in the catalysis of enzymatic reactions. Additionally, water's ability to act as both an acid and a base enables it to participate in proton transfer reactions, which are fundamental to many biochemical processes. By mediating interactions between biomolecules, water significantly influences the specificity and efficiency of biochemical reactions, playing a key role in sustaining life.

A significant part of my thesis utilizes the 2PT formalism [81, 82], which is outlined in section 3.3, and its grid-based extension, 3D-2PT [83]. These methods leverage information about intermolecular vibrations of water molecules to estimate entropies and ultimately assist in calculating the solvation free energy of proteins (Ch. 5). The hydrophobic or hydrophilic nature of a solute influences the H-bond network of water, in terms of its structure and dynamics. The most noticeable changes in liquid water's structure occur in the area closest to the solute, termed the first hydration shell.

Atomistic simulations can offer valuable insights into key biological processes [84], such as drug development [85, 86], protein-protein interactions [87], and protein folding [88]. These processes occur in water, making hydrophobic and hydrophilic interactions crucial in solvation free energy. Understanding water's role in binding processes is gaining importance in drug design and protein research [89–93]. However, modeling water's influence around a binding site on ligand affinity is challenging. Water can moderate attractive interactions or form repulsive barriers, depending on the solvation preferences of

the binding interface or ligand [94].

Water molecules surrounding a biomolecule can be classified into three distinct groups: a) Internal water, which is integrated into the structure of the biomolecule (e.g., forming water bridges), b) Hydration water, which temporarily resides in the hydration shell around a solute molecule, and c) Bulk water, which is situated at a marked distance (usually about 10 Å) from the solute molecule and remains unaffected by it. [95].

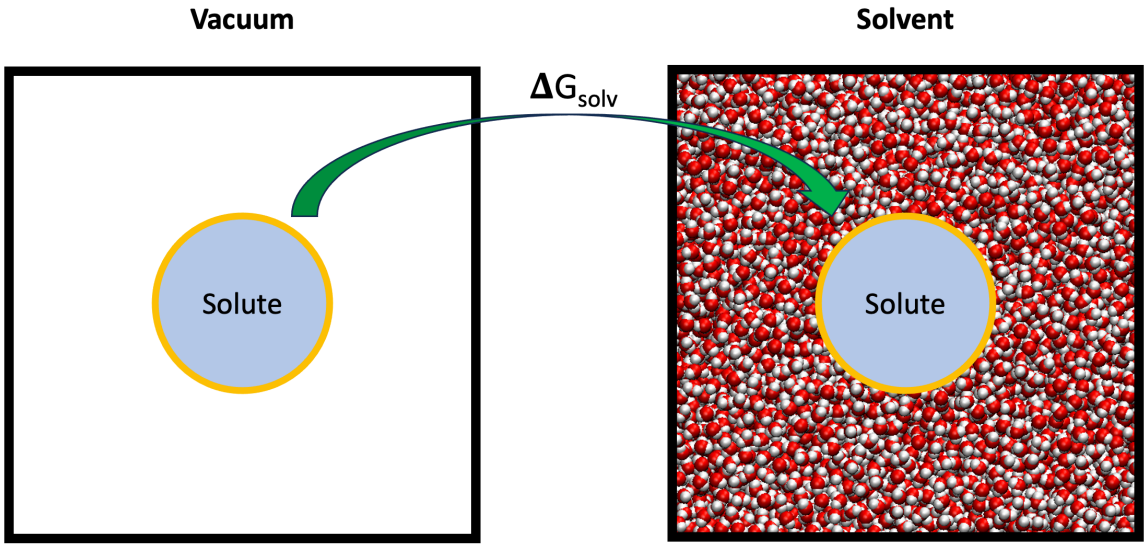

Figure 1.3: The solvation free energy, denoted as ΔG_{solv}, represents the change in free energy when a particle is moved from a vacuum (as shown in the left box) to a solvent environment (as depicted in the right box).

Hydration water is distinct from internal water in that it surrounds and interacts with solvable molecules like proteins but is not part of their molecular structure. It can diffuse more easily than internal water and often exchanges places with bulk water and occasionally with internal water. Although not as free as bulk water, hydration water's interactions with the solute make it crucial in biological processes such as protein folding[96–98].

The solvation free energy is the change in free energy associated with transferring a solute from its standard state in an ideal gas or vacuum to a solution, as illustrated in Fig. 1.3. Solute-solvent interactions are quantified by the solvation free energy, and changes in these interactions during a process involving solvated molecules result in solvent-mediated inter-

actions. It consists of enthalpic and entropic contributions from solute-solvent interactions (hydration water) and additional solvent reorganization terms. However, according to Ben-Naim's cancellation theorem (Fig. 1.4), these solvent reorganization terms counterbalance each other and do not affect the total solvation free energy [1].

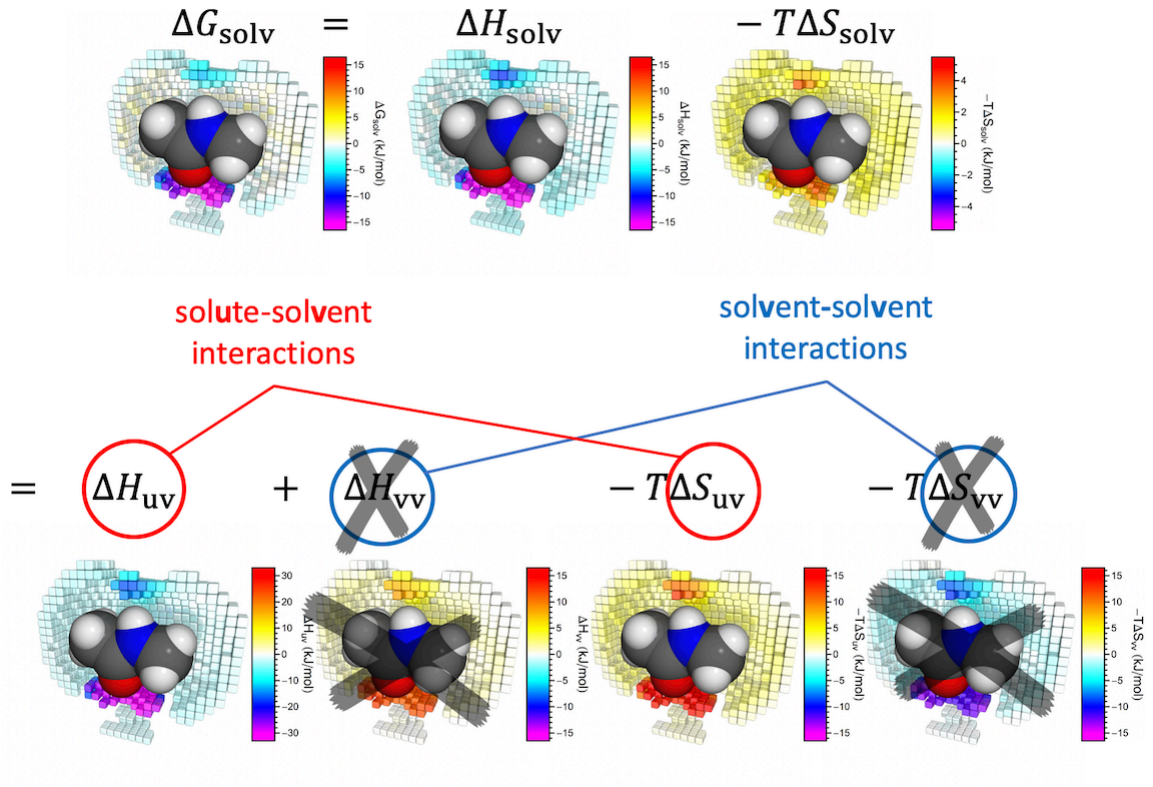

Figure 1.4: Solvation Free Energy Decomposition into Solute-Solvent and Solvent-Solvent Interaction Contributions. Ben-Naim's Cancellation Theorem [1] proves mathematically that enthalpic and entropic terms from solvent-solvent interactions cancel each other out, enabling the calculation of ΔG_{solv} based solely on solute-solvent interactions. Image courtesy *M. Heyden*.

The free energy, being a state function, facilitates the use of thermodynamic cycles to describe solvent-mediated interactions in biological processes like protein-ligand binding. The solvation free energy of binding can be determined by contrasting the free energies of binding in vacuum and solvent, or by comparing the solvation free energies of the complex

with those of the individual binding partners. This can be depicted in Fig. 1.5.

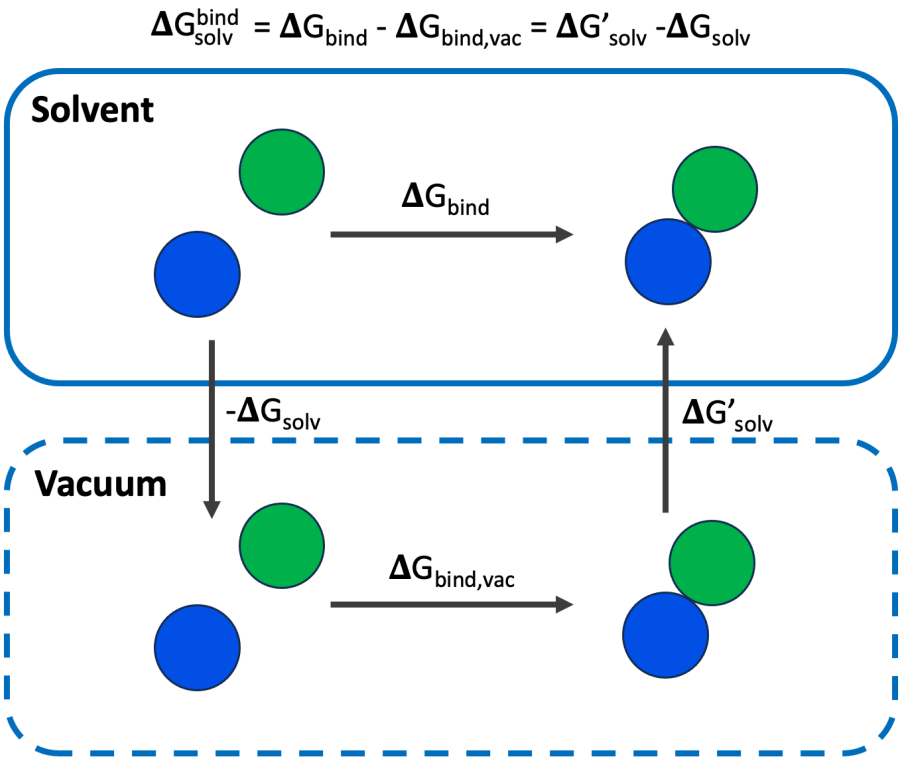

Figure 1.5: By comparing the binding free energies in vacuum and solvent, or by comparing the solvation free energies of the complex and individual binding partners, the thermodynamic cycle allows for the quantification of solvent-mediated interactions.

My first two research work chapters (Ch. 5 and Ch. 6) extensively use this Cancellation Theorem of Ben-Naim [1, 99] (Fig. 1.4) together with the 3D-2PT formalism [83]. Using intrinsically disordered proteins (IDPs) (Ch. 5) as a solute, these are applied to calculate solvation free energy, arising solely from protein-water enthalpic and entropic terms. For the second work (Ch. 6), it is used to compute the water entropy of bulk water in present of different strengths of electric fields. Hence, solvation analysis is extremely significant in studying biomolecular interactions and is one of the central themes of my thesis.

Chapter 2

MOLECULAR DYNAMICS AND UMBRELLA SAMPLING SIMULATIONS

2.1 Introduction

Computer simulations have become an indispensable tool across various domains of the physical sciences. Theoretical models are the basis of these simulations. The outcomes of a simulation can be utilized to evaluate the underlying theory and to compare with experimental data, thereby testing the validity of the model. An effective model can provide deep insights into the chemical process being simulated, offering details on atomic interactions across different lengths, and time scales that might be challenging or even impossible to obtain through experimental methods. Moreover, simulations offer the flexibility to explore a wide range of temperature and pressure conditions, including extreme environments that would be difficult to replicate in a laboratory setting.

Proteins are composed of anywhere from several dozen to several hundred amino acid residues, resulting in structures that contain thousands of atoms and exhibit a wide range of dynamics. These dynamics span from bond vibrations occurring on the femtosecond timescale to large-scale conformational changes that can take place over microseconds or even longer periods. Molecular Dynamics (MD) simulations represent one of the few established simulation techniques capable of handling the extensive system size and intricate complexity of protein structures while also capturing the relevant timescales. This section briefly outlines the foundational principles and approximations underlying MD simulations, with a focus on Classical MD simulations. It specifically addresses those aspects of MD simulations that were particularly relevant to this thesis, without aiming for an exhaustive overview. In addition, we explore the subject of enhanced sampling techniques, with a par-

ticular emphasis on umbrella sampling, which is employed in one of the research topics discussed in this thesis.

Classical Molecular Dynamics (MD) simulations utilize Newton's equations of motion to model the behavior of molecular systems computationally. The inception of molecular dynamics calculations can be traced back to the late 1950s, with the work of Alder *et al*. [100], who conducted simulations on simple gases represented as hard spheres. A pivotal moment in the evolution of MD came in the late 1970s with the publication of "Dynamics of folded proteins" by McCammon *et al*. [101], which marked a significant step forward in the application of MD to the study of biological molecules. With the progression of technology, Molecular Dynamics (MD) has grown in significance, offering the capability to dynamically simulate larger systems over continually extending timescales, thanks to improvements in computational capabilities. Contemporary simulations are now able to routinely capture the dynamics of systems comprising 10^5 to 10^6 atoms over timescales of microseconds. Furthermore, advancements in specialized hardware have extended the reach of MD simulations to cover timescales as long as milliseconds[102–105]. The proficiency of contemporary MD simulations in modeling the dynamics of large atomic assemblies has rendered them a crucial computational instrument that augments experimental studies in biomolecular systems. Molecular Dynamics (MD) simulations are capable of capturing various key biomolecular phenomena such as ligand binding, protein folding, and conformational changes in proteins. These simulations compute the trajectories of interacting atoms and molecules by numerically solving Newton's equations of motion. The forces and potential energies between particles are determined using empirical potentials, commonly referred to as force fields. Given that an analytical solution is unattainable for systems with more than three interacting particles, MD simulations employ a numerical integration scheme to solve the equations of motion for systems comprising several million interacting atoms. The time step chosen for this numerical integration is typically on the

order of femtoseconds (10^{-15} seconds), which is sufficient to capture the fastest molecular vibrations. However, since many biological processes occur over timescales of nanoseconds or longer, the number of iterations (simulated time steps) needed to adequately sample these processes often exceeds millions or even billions.

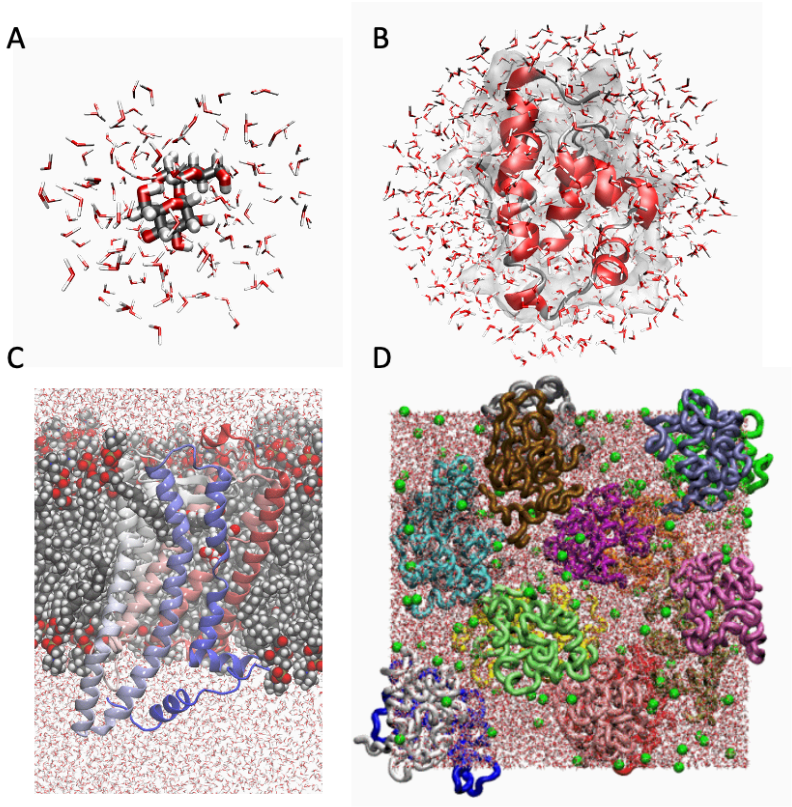

Figure 2.1: The potential of Molecular Dynamics to simulate systems from the minute to the vast, up to 10^6 atoms: (A) An individual molecule, (B) A protein molecule, (C) A fragment of a cell membrane, and (D) A multi-protein assembly. Image Courtesy: *M. Heyden*.

The accumulation of errors from a large number of discrete integration steps, which aim to approximate the true solution, can lead to significant deviations, such as in the conservation of energy in micro-canonical (NVE = constant number of particles, constant volume, constant energy) ensembles. However, for most practical uses of MD simulations in canonical (NVT = constant number of particles, constant volume, constant temperature) ensembles, these cumulative integration errors are dampened by the simulated heat exchange

with an infinite heat bath. This damping is effective as long as the cumulative error over time (drift of the corresponding conserved quantity) remains small compared to the thermal fluctuations.

In designing Molecular Dynamics (MD) simulations, carefully considering the simulation box size is essential to avoid artifacts from boundary conditions. Periodic boundary conditions (PBC) are commonly used to mimic an infinite system by repeating the simulation box in all directions. However, only the properties of the original box are tracked. PBC permits particles to move freely, with particles that "exit" the box reentering on the opposite side. The minimum image convention, employed alongside PBC, ensures that interactions are limited to the closest images of the particles. While PBC eliminates interfaces with a vacuum, insufficient box size can lead to artifacts, as it may amplify correlations between particles due to their periodic images. Additionally, selecting appropriate input parameters, like the integration timestep and total simulation duration, is crucial for achieving a balance between accuracy and computational efficiency. The integration timestep, which is the time interval between consecutive updates of potential energies and, notably, forces, must be sufficiently short to accurately capture the system's fastest (highest frequency) vibrations. Consequently, timesteps in MD simulations typically fall in the femtosecond (fs) range. However, some tools, such as the SHAKE [62] or LINCS [106] constraint algorithms, can fix bond lengths and/or angles, thereby reducing the degrees of freedom associated with high-frequency vibrations. This reduction permits the use of larger integration timesteps, such as 2 fs, in all-atom MD simulations. Moreover, the implementation of multiple timestep algorithms, like the reversible reference system propagator algorithms (RESPA) [107], can further enhance simulation speed by updating forces from different components of the potential energy at varying intervals. Running a longer simulation increases the computational cost. Currently, the total length of simulation trajectories can extend up to hundreds of milliseconds on specialized molecular dynamics supercomputers like

Anton 3 [104, 105, 108]. Regardless of the computational platform used—be it a personal computer, a computer cluster, or a specialized supercomputer—parallelization algorithms are employed to distribute the computational load across multiple CPUs or GPUs simultaneously. The primary computational challenge in MD simulations lies in the time-consuming evaluation of non-bonded Lennard-Jones and electrostatic potentials. Since Lennard-Jones potentials diminish with distance, a cutoff can be implemented for interactions beyond a certain range without introducing significant errors. Analytical corrections can then be applied to adjust for truncation errors [109] in the system's energy and pressure.

In contrast, electrostatic interactions are long-ranged, making a simple distance cut-off generally impractical. Instead, the particle mesh Ewald method (PME) [110] is often employed as one of the most common algorithms to incorporate long-ranged electrostatic interactions with minimal computational costs. In MD simulations involving solvents, the representation of the solvent can be explicit, with the simulation of individual solvent molecules, or implicit, using a continuum solvation model. Explicit solvent models, though computationally intensive, usually incorporate a significant number of solvent molecules that play a role in force evaluations. Water is often the solvent of choice, and a variety of water models, as outlined in Chapter 1, may be selected depending on the specific system and properties under investigation. On the contrary, implicit solvation models do not account for individual solvent particles; rather, they model the solvent as a uniform medium through a mean-field approach.

For all the research projects mentioned in this thesis, explicit water has been used. The previous chapter (Ch. 1) covers it in detail.

2.2 Approximation I: Born-Oppenheimer Approximation

Classical MD simulation involves three main approximations. The first one is the Born-Oppenheimer Approximation. The dynamics of a system is governed by the time-dependent

Schrödinger equation.

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

where H represents the Hamiltonian, which is the sum of the potential and kinetic energy, Ψ is the complete wave function of the system, and $\hbar$ is the reduced Planck's constant, given by $\hbar = h/2\pi$, where h is Planck's constant. The total wave function includes all particles in the system, meaning it's a function of the coordinates and momenta of both nuclei and electrons. Since electrons have a much lower mass and, consequently, a much higher velocity compared to nuclei, they are often considered to instantaneously adjust to nuclear motion. Hence, in the Born-Oppenheimer approximation [111, 112], the total wave function, Ψ_{tot}, is divided into the nuclear wave function, ψ_n, and the electronic wave function, ψ_e,

$$\Psi_{tot}(R, r) = \psi_n(R)\,\psi_e(r|R) \tag{2.2}$$

where R and r are the coordinates and momenta of the nuclei and electrons, respectively. Hence, the electronic wave function ψ_e is dependent on the position of the nuclei in a parametric manner, without being influenced by their dynamics. Due to this approximation, Eq. 2.2 splits into two separate equations: a time-dependent Schrödinger equation governing the motion of the nuclei, and a time-independent Schrödinger equation describing the electronic dynamics.

2.3 Approximation II: Force Fields

2.3.1 Definition of a Force Field

The Born-Oppenheimer approximation facilitates treating the electronic wave function of a system purely based on the nuclear coordinates. Within this approximation, electronic energies are determined by solving the time-independent Schrödinger equation for electrons. However, for proteins with their extensive number of electrons, this approach is

impractical. As a result, an alternative approximation is adopted, which involves using a classical force field to calculate the system's potential energy based on nuclear positions. This force field is tailored to align with the quantum-mechanical ground state energy and is built on a fundamental 'ball-and-spring' model for atoms linked by bonds [113, 114]. Detailed overview and recent developments in this field, have been reviewed by Jørgensen *et al*. [115].

A molecular mechanics force field is a mathematical model or formulation that expresses a molecular system's potential energy in terms of the positions of its particles. This model provides a simplified, empirical representation of the system's potential energy, using a parameterized analytic formula that is easy to calculate. Instead of delving into the physical details of the electronic properties and degrees of freedom, which would require quantum mechanical analysis, the force field captures the electronic energy of the molecules implicitly. Therefore, the interactions between particles and the energies within molecules are described in a purely phenomenological fashion [116, 117]. The parameters of a force field can be derived either from theoretical methods, such as ab-initio or semi-empirical quantum mechanical calculations, or by fitting to experimental data, such as neutron, X-ray, electron diffraction, NMR, and various spectroscopies like infrared, Raman, and neutron spectroscopy [118]. In a typical force field, molecules are modeled as atoms linked by spring-like harmonic oscillators, adhering to a fixed molecular topology that encompasses bonds, bond angles, and proper and improper dihedrals. The harmonic oscillator model for chemical bonds prevents the breaking of bonds, as the potential energy of the oscillator becomes infinite with increasing atom separation. To improve upon this model, Morse oscillators can be used to replace the parabolic potential, allowing for the possibility of breaking chemical bonds between particles given sufficient energy. Nonetheless, reactive force fields that permit a changeable molecular topology, like ReaxFF [119], are not frequently used in biomolecular systems.

The expression for potential energy in a force field can be segmented into contributions from bonded interactions and non-bonded interactions.

$$U(\mathbf{r}) = U_{\text{bonded}}(\mathbf{r}) + U_{\text{non-bonded}}(\mathbf{r}) \tag{2.3}$$

2.3.2 Bonded Interactions

Bonded interactions are those that rely on intramolecular degrees of freedom, such as bond lengths, angles between chemical bonds, dihedral angles between chemical bonds separated by a linking bond, and the deviation from optimal geometries in and out of the plane, which can be characterized by improper dihedral angles [116]. The total of all intramolecular interactions then constitutes the bonded contribution to the potential energy:

$$\begin{aligned} U_{\text{bonded}}(\mathbf{r}) &= \sum_{\text{bonds}} k_i^{\text{bond}}(\mathbf{r}_i - \mathbf{r}_0)^2 + \sum_{\text{angles}} k_i^{\text{angle}}(\theta_i - \theta_0)^2 \\ &+ \sum_{\text{dihedrals}} k_i^{\text{dihe}}[1 + cos(n_i\phi_i + \delta_i)] \\ &+ \sum_{\text{improper}} \frac{k_i^{\text{imp}}}{2}(\xi_i - \xi_0)^2 \end{aligned} \tag{2.4}$$

Here, k represents a force constant, $\mathbf{r}$ denotes particle coordinates, θ symbolizes bond angles, ϕ signifies dihedral angles, and ξ represents improper dihedral angles. The periodic dihedral potential includes a multiplicity factor n and a phase constant δ. In Eq. 2.4, the stretching of bonds is depicted as a simple harmonic oscillator that regulates the length of covalent bonds. As previously noted, the potentials in this model rely on empirical parameters, including parameters like the equilibrium bond length ($\mathbf{r}_0$) of chemical bonds, which can be obtained from methods such as X-ray diffraction or vibrational spectroscopy. Simultaneously, the spring constant may be determined from infrared or Raman spectra [118]. Although the harmonic potential may not be an ideal approximation due to its implication that bonds cannot break, it often serves as a reasonable compromise between computational

expense and accuracy. Angle potentials are typically represented as harmonic; however, two additional expressions can be utilized. One is a trigonometric potential, and the other is the Urey-Bradley potential $\sum k_i^{\mathrm{UB}}(\mathbf{s}_i - \mathbf{s}_0)^2$, where s represents the distance between the two outer atoms forming the angle [118]. For more complex molecules, dihedral potentials, which represent torsional rotation around the central bond, must be included. The energy barriers for dihedral rotations are usually low enough to permit conformational transitions at room temperature, making accurate parameterization of dihedral potentials essential for replicating the conformational dynamics in complex molecules, including biopolymers like proteins and nucleic acids [118]. Dihedral potentials are often expressed using cosine functions, where a combination of multiple terms allows for the creation of periodic potentials with multiple distinct minima. The parameters for dihedral potentials are typically derived from ab initio calculations using potential energy scans along the corresponding dihedral angle. The final term in Eq. 2.4 accounts for so-called improper dihedral angles, which help maintain the planarity of certain groups whose description described by the dihedral term isn't possible. It represents the potential for particles moving in and out of the plane, with ξ denoting the improper angle that measures deviation from planar geometry.

2.3.3 Non-bonded Interactions

Apart from bonded-interactions, the non-bonded interactions account for all interactions between atoms within the same molecule that are separated by three or more chemical bonds, as well as interactions between atoms of different molecules. Interactions between atoms separated by one or two bonds (1-2 and 1-3 pairs) are excluded because their interaction is already encapsulated within bonded potential terms. Atoms separated by three bonds can still partake in non-bonded intramolecular interactions, but these are typically addressed as a distinct case using scaling factors [120]. In many empirical force fields, non-bonded potentials are formulated as the sum of pairwise additive Lennard-Jones and

electrostatic potentials. The electrostatic potentials are modeled using Coulomb's law, describing interactions between empirical point charges located either on individual atoms or specific dummy sites defined by the internal coordinates of each molecule, such as in the TIP4P [68] and TIP5P [121] water models.

$$U_{\text{non-bonded}}(\mathbf{r}) = \sum_i \sum_{i \neq j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{\mathbf{r}_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{\mathbf{r}_{ij}} \right)^{6} \right] + \frac{1}{4\pi\epsilon_0} \sum_i \sum_{i \neq j} \frac{q_i q_j}{\mathbf{r}_{ij}} \tag{2.5}$$

where ϵ_{ij} represents the well depth of the Lennard-Jones potential between atoms i and j, σ_{ij} is the distance at which the potential energy is zero, q_i is the fixed partial charge of atom i, and ϵ_0 is the permittivity of the vacuum. As depicted in Eq. 2.5, individual interactions are expressed as an aggregate of pairwise additive potentials for all interacting particles in the given system. The Lennard-Jones potential uses a term inversely proportional to the twelfth power of the distance ($\frac{1}{r^{12}}$) i.e. Pauli repulsion term to approximate repulsive interactions, chosen mainly for computational efficiency. On the other hand, the $\frac{1}{r^6}$ term represents the attractive interactions, aligning with the expected behavior of London dispersion forces. The parameters σ_{ij} and ϵ_{ij}, which denote the zero-energy distance and potential well depth respectively between atoms i and j, can be individually defined in the force field for each atom pair. However, to simplify, many force fields assign individual atomic parameters (σ_i and ϵ_i) and use rules for combining them. Typically, the geometric mean, $\epsilon_{ij} = (\epsilon_i \epsilon_j)^2$, is used for combining ε values, while the arithmetic mean, $\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$, is used for σ values. The choice of combining rule varies with the force field. In my research, I utilized the Amber and CHARMM force fields, both of which adopt the arithmetic mean for σij and geometric mean for ϵ_{ij}.

The second term in Eq. 2.5 represents the electrostatic potential, reflecting the Coulomb interaction between atoms carrying a non-zero partial charge [116].

Partial charges q are typically determined by fitting the electrostatic potential around a molecule to match the potential obtained from *ab initio* electronic structure calculations [118].

For small molecules, adjustments can be made to partial atomic charges alongside other parameters to fit experimental data predicted from simulations of the molecule in its liquid state. This fitting process is commonly employed in force field models for water (such as TIP3P, SPC, SPC/E, TIP4P, and TIP5P) and other solvent molecules. The calculation of non-bonded interactions in the force field is computationally intensive because it involves evaluating all pairwise interactions in the system. This becomes especially challenging for large systems, leading to the adoption of various methods to reduce computational costs.

The Lennard-Jones term in Eq. 2.5, used to represent van der Waals (vdW) interactions, is effective only over short ranges. Consequently, interactions between atoms separated by considerable distances have a negligible impact on the overall potential. This feature enables the establishment of a cutoff distance $\mathbf{r}_{\mathrm{vdW}}$, which requires careful ascertainment to avoid excluding relevant interactions [122]. Interactions between atoms beyond this cutoff distance are disregarded. There can be slight artificial discontinuities in the potential energy at the cutoff point. To circumvent this issue, the interaction energy can be gradually reduced near the cutoff distance $\mathbf{r}_{\mathrm{vdW}}$ through sigmoidal functions that ensure a smooth transition from the residual interaction energy near $\mathbf{r}_{\mathrm{vdW}}$ to zero as the distance increases, instead of a sharp cutoff [122]. This gradual reduction technique near the actual cutoff distance is utilized for larger systems. For smaller systems, where the cutoff distance exceeds half the box size, the cutoff distance needs to be adjusted to prevent self-interaction between atoms and their periodic images [123].

The second part of Equation 2.5 outlines the electrostatic forces at play between atomic pairs, which extend significantly further than van der Waals forces. To address the calculation of long-range electrostatic contributions, the Ewald summation technique is used, which breaks down the interactions into short-range and long-range components [124]. By using the convergence function $\varphi(r)$, the original aggregate can be divided into two distinct

sums, allowing the potential to be expressed accordingly.

$$\begin{aligned} U_n &= \frac{1}{2}\sum_{l}\sum_{ij}^{N}\frac{B_{ij}}{|\mathbf{r}_{ij}+\mathbf{l}|^n} \\ &= \frac{1}{2}\sum_{l}\sum_{ij}^{N}\frac{B_{ij}\phi(\mathbf{r}_{ij})}{|\mathbf{r}_{ij}+\mathbf{l}|^n} + \frac{1}{2}\sum_{l}\sum_{ij}^{N}\frac{B_{ij}[1-\phi(\mathbf{r}_{ij})]}{|\mathbf{r}_{ij}+\mathbf{l}|^n} \end{aligned} \tag{2.6}$$

In the above equation, n specifies the power to which the potential is raised, B serves as the scaling factor for the potential, $\mathbf{r}_{ij}$ are the minimum distances between atoms, and $\mathbf{l}$ are the lattice vectors [124]. The initial part of Eq. 2.6 targets short-range interactions and is fairly straightforward to evaluate. However, the following term necessitates a Fourier transform for its calculation. The convergence function's values are expected to be smooth for shorter $\mathbf{r}_{ij}$ values and should swiftly decline to zero as $\mathbf{r}_{ij}$ increases [124]. The Ewald summation approach is highly efficient when the charge distribution is initially projected onto a discrete grid in real space, enabling the use of three-dimensional fast Fourier transform (FFT) algorithms [116]. FFTW is one such algorithm utilized for conversion into and out of reciprocal space [125]. This approach is adopted in the particle mesh Ewald (PME) algorithm [126], which has been applied to all the MD simulations in this body of work. PME refines the efficiency of the reciprocal sum by interpolating charges onto a grid rather than tallying wave vectors directly. A Fourier transform through a 3D FFT algorithm then computes the reciprocal energy term by a single summation over the k-space grid. The inverse 3D FFT computes the potential at these grid points and, with interpolation factors, facilitates the calculation of forces on individual atoms [109]. For medium to large-sized systems, PME is markedly more efficient than traditional Ewald summation, with computational complexity scaling as $N\log(N)$. In contrast, for smaller systems, the preparatory steps of grid setup and Fourier transformation in PME may consume more computational time than a simple Ewald summation.

2.4 Approximation III: Classical Dynamics for Nuclei

2.4.1 Newton's Equations of Motion

The complex size of proteins and large biomolecules precludes solving the time-dependent Schrödinger equation for nuclear motion. Consequently, as a third approximation, the dynamics of the nuclei are described classically using Newton's second law. Hence, classical molecular dynamics (MD) simulations rely on Newton's equations of motion to model a system of N interacting particles, necessitating numerical solutions since analytical ones are not feasible for systems with three or more interacting particles.

$$-\nabla_i U(\mathbf{r}(t)) = m_i \frac{d^2\mathbf{r}_i}{d^2t}\ , \quad \text{or} \tag{2.7}$$

$$F_i = m_i\, a_i \tag{2.8}$$

where ∇_i represents the gradient $\left(\frac{\partial}{\partial \mathbf{r}_{\mathrm{i,x}}}, \frac{\partial}{\partial \mathbf{r}_{\mathrm{i,y}}}, \frac{\partial}{\partial \mathbf{r}_{\mathrm{i,z}}}\right)$ [116].

In Equation 2.4, the force experienced by atom i is represented as the negative gradient of the potential energy concerning the atom's coordinates. These equations are solved numerically for every atom using discrete time steps, typically on the order of femtoseconds. During the simulation, the system's progress is tracked, with the positions of the atoms recorded at every nth step. This process creates a trajectory, $\mathbf{r}(t)$, which maps out the changes in the system's coordinates as a function of time. To initiate any MD simulation, it is crucial to have the initial coordinates and velocities of each atom at time $t = 0$. These initial velocities can be randomly assigned in accordance with the Maxwell-Boltzmann velocity distribution, which governs the distribution of speeds for particles in a gas.

$$p(\mathbf{v}_{\mathrm{i}}) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left(-\frac{m_i \mathbf{v}_i^2}{2k_B T}\right) \tag{2.9}$$

here, k_B is the Boltzmann's constant. To begin MD simulations, the initial parameters outlined earlier are essential for numerical integration methods. With each new time step

Δt, the method updates the coordinates and velocities based on the previously determined set, using the forces calculated for that step. One of the most straightforward numerical integration techniques employed is the Euler algorithm.

$$\begin{aligned} \mathbf{r}\,(t+\Delta t) &= \mathbf{r}\,(t) + \mathbf{v}\,(t)\,\Delta t + \frac{\mathbf{f}\,(t)}{2m}\Delta t^2 \\ \mathbf{v}(t+\Delta t) &= \mathbf{v}(t) + \frac{f(t)}{m}\Delta t \end{aligned} \tag{2.10}$$

In this equation, $\mathbf{r}(t)$ denotes the coordinates of atoms at time t, $\mathbf{v}(t)$ represents the velocity at time t, $\mathbf{f}$ is the force, m is the particle mass, and Δt is the time step. However, the Euler algorithm, described in Eq. 2.10, suffers from not being time-reversible and introduces local truncation errors of order (Δt^2) and global errors of order (Δt) [127]. To overcome these limitations, MD simulations commonly employ the velocity-Verlet and leap-frog algorithms. These methods strike a balance between Euler's simplicity and the precision of higher-order techniques. A crucial difference between Euler's method and these two algorithms is their second-order accuracy, making them symplectic [128] and time-reversible. The velocity-Verlet and leap-frog methods, although they treat velocities in slightly varied manners, are algebraically similar. In the velocity-Verlet algorithm, coordinates and velocities are updated simultaneously, leading to enhanced local errors at each time step compared to Euler's method, with errors of order (Δt^4) for coordinates and (Δt^2) for velocities. As for global errors, both coordinates and velocities in the velocity-Verlet algorithm exhibit errors of order (Δt^2) [127]. Expressing the velocity - Verlet algorithm as:

$$\begin{aligned} \mathbf{r}\,(t+\Delta t) &= \mathbf{r}\,(t) + v\,(t)\,\Delta t + \frac{f\,(t)}{2m}\Delta t^2 \\ v(t+\Delta t) &= v(t) + \frac{f(t)+f(t+\Delta t)}{2m}\Delta t \end{aligned} \tag{2.11}$$

In contrast to the velocity-Verlet algorithm, the leap-frog algorithm staggers the updates of coordinates and velocities by a half-time step, as demonstrated in the subsequent expression.

The leap-frog algorithm is represented as:

$$v\left(t+\frac{\Delta t}{2}\right) = v\left(t-\frac{\Delta t}{2}\right)+\frac{f(t)}{m}\Delta t \tag{2.12}$$

$$\mathbf{r}\,(t+\Delta t) = \mathbf{r}\,(t)+v\left(t+\frac{\Delta t}{2}\right)\Delta t \tag{2.13}$$

In reviewing Eq. 2.11 and Eq. 2.13, it may not be immediately obvious that the velocity-Verlet and leap-frog algorithms are equivalent, given their distinct approaches to updating coordinates and velocities. Additionally, while the velocity-Verlet algorithm requires initial coordinates and velocities to begin a new simulation, the leap-frog algorithm necessitates the calculation of velocities at $\frac{\Delta t}{2}$ for initialization. This can be achieved by reversing Euler's method with $\frac{\Delta t}{2} = -\frac{1}{2}$. Such an adjustment allows for a reformulation of the leap-frog method that facilitates the simultaneous evaluation of positions and velocities. In my research, all simulations were conducted using the leap-frog integrator, as implemented in the GROMACS software package [109].

The force is a critical component that must be evaluated both rapidly and precisely. Errors in the coordinates, obtained during the numerical integration of the equations of motion, can particularly impact the repulsive components of Lennard-Jones interactions, which scale with $\left(\frac{1}{\mathbf{r}^{12}}\right)$, resulting in forces that scale with $\left(\frac{1}{\mathbf{r}^{13}}\right)$. The same issue appears with strong covalent bonds possessing large force constants, as minor errors in integrating coordinates can result in substantial forces if bond constraints are not applied. If the time step is too large to accurately capture these interactions and movements, it can lead to the artificial generation of potential energy. This increase in potential energy often arises from fast particles being too close to the steep rise of a repulsive potential term, which may subsequently transform into kinetic energy during the simulation, amplifying the error in updating positions and velocities. Fast stretching vibrations of OH or CH bonds, with a period of about 10 fs, are particularly sensitive to this issue. To resolve these vibrations accurately, it is advised to use a time step of 0.5 fs or smaller. Specialized constraint algorithms, like

LINCS, which reset bonds to their equilibrium lengths following an unconstrained update, effectively eliminate these degrees of freedom from the simulation, thereby enabling larger time steps. However, in scenarios where the time step is too large, such as in extended simulations within a microcanonical NVE ensemble, significant errors in total energy can accumulate due to the artificial generation of potential energy, unless corrective measures for energy are involved.

2.5 Temperature and Pressure Coupling

2.5.1 Temperature Coupling

Newton's equations of motion, typically in the NVE (constant number of particles, constant volume, constant energy) ensemble. However, the canonical (NVT) ensemble [77], where particle count, volume, and temperature are kept constant, is more commonly used in various applications. To maintain a constant temperature during simulations, several temperature control algorithms are employed. In this thesis, all simulations were carried out using the GROMACS software package, where I utilized the Berendsen [129] and Nosé-Hoover [130] temperature control schemes. Maintaining temperature in simulations offers additional advantages, such as eliminating total energy drifts caused by force truncation and integration errors [109]. This helps prevent the artificial accumulation or build-up of energy due to the imperfect methods described in previous sections. In the simulations, temperature regulation is accomplished by linking the simulated system to an external heat bath through the use of thermostats. Within this thesis, two distinct temperature control strategies were implemented. The first is the Berendsen algorithm, which facilitates weak temperature coupling to an external bath. It operates by adjusting the atomic velocities to rectify any discrepancies between the system's temperature and the target temperature T_0,

as per the algorithm described below:

$$\frac{\partial T}{\partial t} = \frac{T_0 - T}{\tau_T} \tag{2.14}$$

This suggests that any deviation in temperature decays exponentially with a time constant τ_T [109]. The Berendsen thermostat's advantage lies in its ability to adjust the coupling strength according to the user's needs. However, it does not produce a proper canonical ensemble as it dampens fluctuations in kinetic energy [131]. The error associated with this thermostat scales with $\frac{1}{N}$, meaning that most ensemble averages will not be significantly impacted for sufficiently large systems. Nonetheless, the kinetic energy distribution will remain inaccurate. For system equilibration, the Berendsen thermostat is commonly used due to its rapid and effective temperature adjustment. After achieving equilibration, it is suggested to transition from the Berendsen to a thermostat such as Nosé-Hoover, which can accurately produce a canonical ensemble. The Nosé-Hoover thermostat adds an extra degree of freedom through a unique equation of motion, in which a friction coefficient $\xi(t)$ influences the atomic velocities $\mathbf{v}_i(t)$ and is linked to the system being simulated.

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = -\nabla_i U(\mathbf{r}(t)) - \xi(t)\frac{d\mathbf{r}_i}{dt} \tag{2.15}$$

The friction constant $\xi(t)$ is determined by the difference between the current temperature and the target temperature T_0, and it follows its distinct equation of motion:

$$\xi(t) = \frac{1}{Q}(T - T_0) \tag{2.16}$$

Here, Q represents a mass parameter that is inversely proportional to the square of the oscillation frequency [116]. As illustrated in Eq. 2.15, the Nosé-Hoover thermostat facilitates an oscillatory exchange of kinetic energy between the simulated system and the heat bath. Consequently, the temperature oscillates around the reference temperature, ultimately producing a correct canonical ensemble in this scenario.

2.5.2 Pressure Coupling

Just as with temperature coupling, the system can be coupled to a pressure bath. During simulations using GROMACS, I employed the Berendsen barostat for the equilibration of the NPT (constant number of particles, constant pressure, constant temperature) ensemble. This barostat is highly effective in relaxing the system, much like the Berendsen thermostat. It modifies the coordinates and box vectors at every step, leading to a first-order kinetic relaxation of the pressure towards a target reference pressure P_0 [132].

$$\frac{\partial P}{\partial t} = \frac{P_0 - P}{\tau_p} \tag{2.17}$$

While the Berendsen barostat is effective for the equilibration phase, it should be avoided for production simulations as it does not accurately probe the true NPT ensemble, owing to its quelling of pressure fluctuations.

For constant pressure, the extended ensemble is implemented via the Parrinello-Rahman method, similar to the Nosé-Hoover approach, which also follows its specific equation of motion. In this barostat, the matrix **b**, representing the box vectors, is governed by the matrix equation of motion [133, 134]:

$$\frac{d^2\mathbf{b}}{dt^2} = V\mathbf{W}^{-1}b'^{-1}(P - P_0) \tag{2.18}$$

In this expression, V denotes the volume of the simulation box, while P and P_0 are the matrices of the current and reference pressures, respectively. The term $\mathbf{W}^{-1}$ represents the inverse of a mass parameter that dictates the coupling strength and characterizes the box's deformability. The strength of the coupling is influenced by the dimensions of the simulation box L, the isothermal compressibilities β, and the pressure-time constant τ_p.

$$\left(\mathbf{W}^{-1}\right)_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L} \tag{2.19}$$

If the initial pressure in the simulation is significantly off from equilibrium, the Parrinello-Rahman barostat can lead to large oscillations in the box, potentially causing the simulation

to fail. As previously mentioned, it's more effective to first employ the weak-coupling method to attain the desired pressure and then switch to the Parrinello-Rahman barostat to accurately maintain a true NPT ensemble.

2.6 Enhanced Sampling Methods

Classical molecular dynamics (MD) simulations are a powerful tool for studying the dynamics of physical, chemical and biochemical systems at the atomic level [135–138]. However, they often face limitations when it comes to sampling long timescale events, such as protein folding, conformational changes in large biomolecules, or rare events like chemical reactions. These processes can occur on timescales much longer than those accessible by standard MD simulations, which typically span nanoseconds to microseconds. To overcome these limitations, enhanced sampling techniques [139–142] have been developed to accelerate the exploration of the conformational space and access longer timescale events.

Generally, these methods fall into two categories: one category utilizes bias potentials along specific collective variables (CVs) of the system (termed as CV-based methods), whereas the methods in the other category operate without the need for predefined CVs (referred to as CV-free methods).

CV-based sampling methods significantly enhance the convergence of free energy calculations when the collective variables (CVs) are appropriately chosen. Techniques such as Umbrella Sampling (US) [143, 144], Metadynamics (MetaD) [139, 145, 146], Steered Molecular Dynamics (SMD) [147], Adaptive Biasing Force (ABF) [148], Conformational Flooding [149], the String Method [150], and Temperature Accelerated Molecular Dynamics (TAMD) [151] are commonly used methods in this category of enhanced sampling. However, identifying suitable CVs for many systems can be a challenging task.

In cases where selecting suitable CVs is challenging, CV-free sampling methods prove

beneficial as they enable the exploration of biomolecular configurations without the need for prior knowledge of the system. These methods are particularly useful for investigating potential structural transition pathways in biomolecules and identifying unknown intermediate states. Notable methods in this second class of sampling techniques include Replica Exchange (RE) [152] or Parallel Tempering (PT), Self-Guided Molecular Langevin Dynamics (SGLD) [153, 154], and Accelerated Molecular Dynamics (aMD) [155]. For more complex or larger systems, employing a combination of enhanced sampling methods, such as Replica Exchange Umbrella Sampling (REUS) [156, 157] or Bias-Exchange Metadynamics (BE-MetaD) [158], can provide a robust approach for efficiently sampling conformational transitions.

A few of the common techniques are discussed here. One of the common approaches, developed by Parinello *et al*., is metadynamics [139, 145, 146], which introduces a history-dependent biasing potential to discourage the system from revisiting previously explored regions of the phase space.

Another popular method is replica exchange molecular dynamics (REMD) [152], where multiple replicas of the system are simulated at different temperatures, allowing for exchanges between replicas and facilitating the sampling of high-energy states.

Temperature-accelerated molecular dynamics (TAMD) [151] and accelerated molecular dynamics (aMD) [155] are techniques that modify the potential energy surface to lower energy barriers and speed up the sampling process. These methods are particularly useful for studying processes with high activation barriers.

Umbrella sampling [143, 144] is a technique that focuses on sampling specific regions of the reaction coordinate by applying a biasing potential. It is often used to calculate free energy profiles of chemical reactions or conformational changes. In umbrella sampling, the reaction coordinate is divided into overlapping windows, and separate simulations are performed for each window. The resulting data are then combined to reconstruct the free

energy landscape.

Overall, enhanced sampling techniques provide a way to study processes that are otherwise inaccessible to classical MD simulations, offering insights into the mechanisms of complex chemical and biochemical phenomena.

In the following discussion, I will concentrate on the CV-based technique of umbrella sampling, which I utilized in Chapter 7 for calculating the folding and binding free energies of disordered proteins.

2.6.1 Umbrella Sampling

Umbrella sampling is an equilibrium-enhanced sampling technique used to calculate the free energy profile along one or more reaction coordinates, or collective variables (CVs). Developed by Torrie and Valleau [143] in 1977, this method facilitates the sampling of conformational dynamics along selected CVs, allowing for the determination of relative free energies of various states along the reaction coordinates. A reaction coordinate (ξ) is commonly a single variable, characterized as a continuous parameter based on atomic coordinates. It streamlines the description of a high-dimensional system by minimizing the dimensionality of its parameters, offering a more manageable representation of the system's behavior.

The selection of a collective variable is crucial for differentiating between distinct states and calculating their free energy differences by biasing the system along the reaction coordinate. It is a critical step in accurately capturing the system's behavior and ensuring that the free energy landscape is well-sampled. This approach is particularly effective in surmounting energy barriers that are often encountered in free energy calculations, making it a widely favored method for exploring complex energy landscapes and understanding the thermodynamics of various processes.

A schematic representation of umbrella sampling is illustrated in Fig. 2.2. In this ap-

proach, a series of independent windows are employed along a selected collective variable (CV) to model the transition of conformation within the system. Each window focuses on a specific region of the CV, allowing for detailed sampling of the energy landscape and facilitating the reconstruction of the free energy profile associated with the conformational transition.

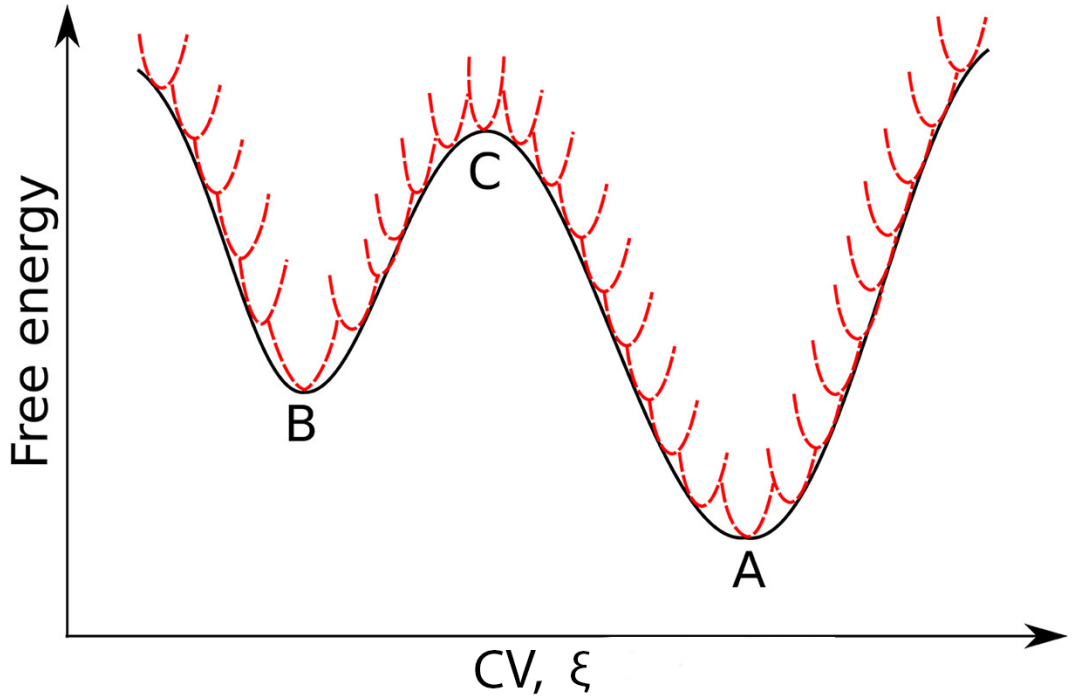

Figure 2.2: The potential of mean force is determined with the help of a series of overlapping umbrella potential windows along the chosen collective variable ξ. Image courtesy. Liao *et al.* [2]

In umbrella sampling, each window is centered at a specific value of the collective variable (CV). To facilitate sampling at that point, a simple harmonic bias potential, $\omega_i(\xi)$, is added to each window. The strength of this bias potential is determined by its force constant, k. This biasing helps overcome energy barriers and enhances the exploration of the conformational space around the targeted CV value. In some instances [144], an adaptive bias function can also be employed to dynamically adjust the biasing potential based on the sampling progress, further improving the efficiency of the sampling process.

When incorporating a bias potential ωi(ξ) that depends on the reaction coordinate into the system within a specific window, the total biased energy can be represented as follows:

$$E^b = E^u + \omega_i(\xi) \tag{2.20}$$

In this expression, *i* denotes the *i*th window in the umbrella sampling system. The subscripts *u* and *b* indicate unbiased and biased simulations, respectively. For this discussion, we are using *E* as the notation for the potential energy of the system, which is equivalent to other common notations such as *V* or *U*. The harmonic bias potential added to this system is depicted as:

$$\omega_i(\xi) = \frac{1}{2}k(\xi - \xi_i)^2 \tag{2.21}$$

In this context, ξ_i serves as the reference coordinate point for a single umbrella window. During the simulation, if the system deviates from the reaction coordinate, the added bias potential exerts a force that nudges the system back towards the mean position, ensuring that the sampling remains focused around the intended region of the reaction coordinate. To determine the unbiased free energy from the biased simulations, it is essential to obtain the unbiased distribution of the reaction coordinate. This involves correcting for the bias introduced by the umbrella sampling potential, allowing for the reconstruction of the true free energy landscape along the reaction coordinate.

In the canonical ensemble (NVT), the Helmholtz free energy can be calculated from the probability distribution as follows:

$$A^u(\xi) = -k_B T\ lnP^u(\xi) = -\frac{1}{\beta}\ lnP^u(\xi) \tag{2.22}$$

For isothermal-isobaric ensemble (NPT), the Gibbs free energy is estimated from:

$$G^u(\xi) = -k_B T\ lnP^u(\xi) = -\frac{1}{\beta}\ lnP^u(\xi) \tag{2.23}$$

The probability distribution along the reaction coordinate, ξ, is determined by integrating out all degrees of freedom except for ξ in the Boltzmann distribution. This process

involves averaging over all possible configurations of the system at each point along the reaction coordinate. By doing so, we capture the probability of the system being in a particular state for each value of ξ. This integration step is crucial as it enables us to derive the probability distribution along the reaction coordinate, providing insights into the relative populations of different states and the free energy landscape associated with the coordinate:

$$P^u(\xi) = \frac{\int \delta[\xi(r) - \xi]\, e^{-\beta E^u(r)}\, dN_r}{\int e^{-\beta E^u(r)}\, dN_r} \tag{2.24}$$

In the expression, dN_r represents the fraction of elements or points in the phase space, which encompasses both the configuration space and momentum space. The term $d\xi$ signifies an interval along the reaction coordinate ξ. The Kronecker delta is utilized to denote the specific instances where the reaction coordinate of a particular configuration falls within the interval $d\xi$. Hence the Kronecker delta is employed to represent $\delta[\xi(r) - \xi] = 1$ when $\xi(r) = \xi$, and $\delta[\xi(r) - \xi] = 0$ when $\xi(r) \neq \xi$.

This delta function ensures that only those configurations contributing to the probability distribution within the specified interval of the reaction coordinate are counted. This approach allows for a precise calculation of the probability distribution by considering the contributions from different regions of the phase space along the reaction coordinate.

To determine the unbiased free energy $G^u(\xi)$, we must first calculate the unbiased probability distribution $P^u(\xi)$. However, what we obtain directly from our umbrella sampling (US) simulation is the biased probability distribution $P^b(\xi)$. This biased distribution reflects the influence of the added biasing potentials in each umbrella sampling window. To extract meaningful thermodynamic information, we need to correct for this bias and recover the true, unbiased probability distribution of the system along the reaction coordinate ξ. This unbiased distribution is essential for accurately computing the free energy landscape and understanding the thermodynamics of the system under study.

To establish a relationship between the biased probability distribution $P^b(\xi)$ and the

unbiased probability distribution $P^u(\xi)$, we can compute $P^b(\xi)$ in a manner similar to $P^u(\xi)$ by incorporating the bias potential $\omega(\xi)$ from Eq. 2.21 into Eq. 2.24. This involves adjusting the probability distribution to account for the biasing effects introduced during the umbrella sampling simulation. By correctly incorporating the bias potential, we can derive the biased probability distribution from the simulation data, which can then be used to recover the unbiased probability distribution and, subsequently, the unbiased free energy profile.

$$P^b(\xi) = \frac{\int \delta[\xi(r) - \xi]\, e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r}{\int e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r} \tag{2.25}$$

or,

$$P^b(\xi) = \frac{\int \delta[\xi(r) - \xi]\, e^{-\beta E^u(r)}\, e^{-\beta\omega(\xi(r))}\, dN_r}{\int e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r} \tag{2.26}$$

Given that the bias potential solely depends on the reaction coordinate ξ, we can rewrite Eq. 2.26 to reflect this dependence and simplify the expression for the biased probability distribution. This modification allows us to explicitly incorporate the effect of the bias potential on the probability distribution along the reaction coordinate.

$$P^b(\xi) = \frac{e^{-\beta\omega(\xi)} \int \delta[\xi(r) - \xi]\, e^{-\beta E^u(r)}\, dN_r}{\int e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r} \tag{2.27}$$

By utilizing Equations 2.24 and 2.27, we can calculate the ratio of the unbiased probability distribution $P^u(\xi)$ to the biased probability distribution $P^b(\xi)$ *i.e.* $P^u(\xi)/P^b(\xi)$ as follows:

$$\frac{P^u(\xi)}{P^b(\xi)} = \frac{\int \delta[\xi(r) - \xi]\, e^{-\beta E^u(r)}\, dN_r \quad \int e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r}{\int e^{-\beta E^u(r)}\, dN_r \quad e^{-\beta\omega(\xi)} \int \delta[\xi(r) - \xi]\, e^{-\beta E^u(r)}\, dN_r} \tag{2.28}$$

After canceling out two similar terms from numerator and denominator, we have:

$$\frac{P^u(\xi)}{P^b(\xi)} = e^{\beta\omega(\xi)}\, \frac{\int e^{-\beta[E^u(r)+\omega(\xi(r))]}\, dN_r}{\int e^{-\beta E^u(r)}\, dN_r} \tag{2.29}$$

or,

$$P^u(\xi) = P^b(\xi)\, e^{\beta\omega(\xi)}\, \frac{\int e^{-\beta\omega(\xi(r))}\, e^{-\beta E^u(r)}\, dN_r}{\int e^{-\beta E^u(r)}\, dN_r} \tag{2.30}$$

In Eq. 2.30, we represent the unbiased probability distribution $P^u(\xi)$ in terms of the biased distribution $P^b(\xi)$ Here, the ensemble average of $e^{-\beta\omega(\xi(\mathbf{r}))} = A$ represents the average value of the exponential of the negative bias potential energy, weighted by the Boltzmann factor, over all possible configurations of the system. This quantity is a crucial component in reweighting the biased probability distribution to obtain the unbiased distribution.

$$< A > = \frac{\int A(r)\, e^{-\beta E^u(r)}\, dN_r}{\int e^{-\beta E^u(r)}\, dN_r} \tag{2.31}$$

Substituting Eq. 2.31 to Eq. 2.30, we can rewrite Eq. 2.30 as:

$$P^u(\xi) = P^b(\xi)\, e^{\beta\omega(\xi)}\ < e^{-\beta\omega(\xi)} > \tag{2.32}$$

So, this equation 2.32 essentially states that the unbiased probability distribution can be obtained by reweighting the biased probability distribution with the exponential of the bias potential and normalizing by the ensemble average of the exponential of the negative bias potential. This reweighting procedure corrects for the bias introduced during the simulation and recovers the true free energy landscape of the system.

By substituting Eq. 2.32 into Eq. 3.27, we can derive the unbiased free energy as follows:

$$G^u(\xi) = -\frac{1}{\beta}\, lnP^u(\xi) = -\frac{1}{\beta}\, ln[P^b(\xi)\, e^{\beta\omega(\xi)}\ < e^{-\beta\omega(\xi)} >] \tag{2.33}$$

or,

$$G^u(\xi) = -\frac{1}{\beta} lnP^b(\xi)\ - \frac{1}{\beta} lne^{\beta\omega(\xi)}\ - \frac{1}{\beta} ln < e^{-\beta\omega(\xi)} > \tag{2.34}$$

or by further simplification, we obtain,

$$G^u(\xi) = -\frac{1}{\beta} lnP^b(\xi)\ - \omega(\xi)\ - \frac{1}{\beta} ln < e^{-\beta\omega(\xi)} > \tag{2.35}$$

In this expression, we define the term $\frac{1}{\beta} \ln e^{\beta\omega(\xi)}$ as $\omega(\xi)$ which is the bias potential applied during the simulation. The biased free energy is denoted as $G^b(\xi) = -k_B T \ln P^b(\xi)$,

and the term $-k_B T \ln\langle e^{-\beta\omega(\xi)}\rangle$ represents the reweighting factor F required to adjust for the bias introduced by the umbrella sampling potential. By incorporating these definitions, we can rewrite the final expression as follows:

$$G^u(\xi) = G^b(\xi) - \omega(\xi) + F \tag{2.36}$$

Therefore, Equation 2.36 provides a method for calculating the unbiased free energy of the reaction, $G^u(\xi)$, using the biased probability distribution $P^b(\xi)$ acquired from umbrella sampling simulations. The calculation of $G^u(\xi)$ also involves considering the artificial bias potential $\omega(\xi)$ applied during the simulations and the factor F, which accounts for the reweighting necessary to correct for the bias. This approach enables the extraction of accurate thermodynamic information from simulations that have been artificially biased to enhance sampling efficiency.

Now, we can apply Equation 2.36 to each window, denoted by the index i, in the umbrella sampling simulation. By doing so, we can calculate the free energy $G_i^u(\xi)$ for each individual window. This approach allows us to obtain a set of free energy values corresponding to the different segments of the reaction coordinate covered by each window. By piecing together these individual free energy values, we can construct the overall free energy profile along the reaction coordinate, providing a detailed understanding of the thermodynamic landscape of the system under study.

$$G_i^u(\xi) = G_i^b(\xi) - \omega_i(\xi) + F_i \tag{2.37}$$

To derive the global unbiased free energy profile $G^u(\xi)$ by combining the individual free energy values $G_i^u(\xi)$ from each window, it is necessary to calculate the reweighting factor F_i for each window. This factor is crucial for accurately integrating the contributions from each window to construct the overall free energy landscape.

2.6.2 Weighted Histogram Analysis Method (WHAM)

Various methods exist for calculating the reweighting factor F_i and the potential of mean force $G^u(\xi)$, but the Weighted Histogram Analysis Method (WHAM) [159–161] is the most commonly used approach. WHAM is specifically designed to amalgamate the results from different windows in umbrella sampling simulations. Its primary objective is to minimize the statistical error in the unbiased probability distribution $P^u(\xi)$ by optimally combining the biased probability distributions from each window. This process ensures a more accurate and reliable reconstruction of the free energy landscape across the entire range of the reaction coordinate.

From the definition of F_i in Eq. 2.35:

$$\begin{aligned} F_i &= -\frac{1}{\beta} ln < e^{-\beta\omega(\xi)} > \\ or, -\beta F_i &= ln < e^{-\beta\omega(\xi)} > \end{aligned} \tag{2.38}$$

$$\begin{aligned} e^{-\beta F_i} &= < e^{-\beta\omega(\xi)} > \\ &= \frac{\int e^{-\beta\omega(\xi)}\, e^{-\beta E_i^u}\, d\xi}{\int e^{-\beta E_i^u}\, d\xi} \\ &= \int P^u(\xi)\, e^{-\beta\omega(\xi)}\, d\xi \end{aligned} \tag{2.39}$$

Equation 2.39 reaches its final form when the definition of the ensemble average is applied. This involves integrating over all possible configurations of the system, weighted by their Boltzmann factors, to obtain a statistically meaningful average value for the quantity of interest. By incorporating the ensemble average, the equation becomes a comprehensive representation that accounts for the contributions of all relevant states in the system.

The global unbiased probability distribution $P^u(\xi)$ is calculated as a weighted average of the probability distributions from each individual window in the umbrella sampling sim-

ulation. This weighted average takes into account the contributions of each window to the overall distribution, ensuring that the final $P^u(\xi)$ accurately reflects the thermodynamic properties of the system across the entire range of the reaction coordinate.

$$P^u(\xi) = \sum_{i}^{N_{windows}} p_i(\xi)\, P_i^u(\xi) \tag{2.40}$$

The weights p_i are chosen to minimize the statistical error of the unbiased probability distribution $P^u(\xi)$. By carefully selecting these weights, the combined probability distribution from all windows can be optimized to provide the most accurate representation of the system's thermodynamic behavior along the reaction coordinate. This optimization is crucial for ensuring that the final free energy profile is as precise as possible.

$$\frac{\partial \sigma^2(P^u(\xi))}{\partial p_i} = 0 \tag{2.41}$$

under the second applied normalizing condition of:

$$P^u(\xi) = \sum_{i}^{N_{windows}} p_i(\xi) = 1 \tag{2.42}$$

Combining Eq. 2.41 and 2.42 leads to[159]:

$$p_i = \frac{a_i}{\sum_j^{N_{windows}} a_j(\xi)} \tag{2.43}$$

Each of these individual weighting factors $a_i(\xi)$ are expressed as follows:

$$a_i(\xi) = N_i\, e^{-\beta\omega_i(\xi)+\beta F_i} \tag{2.44}$$

The Weighted Histogram Analysis Method (WHAM) operates by determining the reweighting factor F_i, which is essential for computing the unbiased free energy. In this process, $a_i(\xi)$ is a weighting factor for each window i at a particular value of the reaction coordinate

ξ, and N_i represents the total number of steps sampled for each window i. The primary objective is to calculate the unbiased probability distribution $P^u(\xi)$, which can be formulated as a linear combination of the probability distributions from each of the umbrella sampling windows, as depicted in Equation 2.40).

To achieve this, the weights in Equation 2.40 are computed using Equations 2.43 and 2.44, with the aim of minimizing the statistical error in the unbiased probability distribution. These weights play a crucial role in accurately combining the biased probability distributions from different windows into a coherent global distribution.

The iterative nature of the WHAM algorithm requires solving the equations repeatedly until the results converge to a stable solution. This iterative process ensures that the final unbiased free energy profile is both accurate and reliable, providing valuable insights into the thermodynamic properties of the system under study.

Chapter 3

3D-2PT

3.1 Introduction

In solutions, molecular interactions are driven by solvation free energies, which play a critical role in determining the conformational stability of biomolecules and the formation of complexes. These energies are central to solvent-mediated interactions, such as the hydrogen bonding, and hydrophobic effect in water. Nevertheless, accurately capturing solvation free energies, especially in aqueous environments, poses a considerable challenge in computer simulations when the solvent is not explicitly represented. To enhance existing models, a detailed understanding of the microscopic interactions between the solute and solvent is required, which is often not readily available from experimental data. This emphasizes the importance of employing advanced computational approaches to accurately model the complex interactions involved in solvation processes and their impact on molecular systems. In recent years, computational tools have been developed to derive insights into solvation free energies and solute-solvent interactions from simulations involving explicit solvents. These tools encompass spatial decompositions of solvation enthalpy and entropy, which hold the potential to enhance our theoretical comprehension of solvation thermodynamics. Several computational techniques have been developed to provide a detailed spatial understanding of solvent thermodynamics. These include inhomogeneous solvation theory (IST) [162, 163] and its three-dimensional (3D) grid-based variant (GIST) [164–166], Two-phase thermodynamics (2PT) [81, 82] and its spatially resolved counterpart, 3D-2PT [167, 168], and cell theory [169, 170] and grid cell theory (GCT) [171]. Moreover, multiscale cell correlation (MCC) [172–174] combines information across various

scales to offer a comprehensive view of solvent thermodynamics. In the course of my research, I have employed the 3D-2PT method to examine the spatially resolved solvation thermodynamics around flexible, unstructured proteins (IDPs) and to analyze the solvation environment of bulk water in the presence of different strengths of electric fields. In this section, we will delve into the methodologies behind the 3D-2PT technique.

3.2 Vibrational Densities of States (VDoS)

Classical Molecular Dynamics (MD) simulations, which utilize Newtonian mechanics to depict the system, enable the computation of diverse time correlation functions and their corresponding spectral densities. These computations are derived from trajectories generated by MD simulations and can be associated with various dynamical events [175]. The velocities of water molecules obtained from Molecular Dynamics (MD) simulations are instrumental in calculating the vibrational density of states (VDoS). This data is critical for understanding the properties of water molecules that are dynamically coupled with biomolecules, such as proteins. The analysis of water velocities near the surface of a biomolecule and in the bulk region provides insights into the behavior of these coupled water molecules. Such water molecules play a pivotal role in influencing the structure, dynamics, and function of the associated proteins.

The examination of the water's vibrational density of states (VDoS) can offer an estimate of the system's thermodynamic properties [176], such as entropy. This renders the VDoS a potent instrument for investigating the thermodynamics of biological systems through Molecular Dynamics (MD) simulations. The vibrational density of states (VDoS) of water can be calculated by applying the Fourier transform to the velocity autocorrelation function (VACF) of water molecules, denoted as (C_v):

$$C_v(\tau) = \langle \mathbf{v}_i(t+\tau)\, \mathbf{v}_i(\tau) \rangle \tag{3.1}$$

where $\mathbf{v}_i$ represents the velocity vector of atom i of a water molecule at time t. In Eq. 3.1, the angular brackets denote the process of averaging across all atoms of a given type in the system and over different initial reference times t. This equation is the conventional method for calculating the vibrational density of states (VDoS) [175]. By performing a Fourier transform on the velocity autocorrelation function (VACF), we can derive the VDoS:

$$VDoS_i(\nu) = \frac{2}{k_B T} \int_{-\infty}^{\infty} e^{i\,2\pi\nu\tau} m_i C_v(\tau)\,\mathrm{d}\tau \tag{3.2}$$

In this expression, k_B represents Boltzmann's constant, and T denotes the temperature of the system. The mass of atom i, represented by m_i, serves as a weighting factor in Eq. 3.2. This equation provides a description of how the kinetic energy associated with the normal modes is distributed across different frequencies in the frequency domain [3]. Additionally, the prefactor $\frac{2}{k_B T}$ plays a crucial role in converting the classical system into the VDoS. This conversion allows the VDoS to be used as a tool to represent the distribution of degrees of freedom across various vibrational frequencies, offering a deeper understanding of the system's vibrational properties. For unconstrained molecules, properties obtained from the VDoS can be easily separated into atomic contributions. This is possible because the overall VDoS is derived from the sum of the spectral densities of individual atoms [177]. For an unconstrained bulk water molecule, the total VDoS, as depicted in Eq. 3.2, is typically determined by aggregating the average contributions of the velocities of oxygen and hydrogen atoms. Regardless, in scenarios where simulated water molecules have constrained bond lengths and angles, as in the TIP3P or TIP4P/2005 water models used in this thesis's research projects, (except for the second work in Ch. 6 where 3D-2PT was modified to include the vibrational component as flexible TIP4P/2005 water was used. [178]) it is not feasible to accurately calculate the total VDoS as separate average contributions from the velocities of oxygen and hydrogen atoms. To capture the contributions to the VDoS, one can alternatively analyze the translational and rotational degrees of freedom of a water

molecule, as well as intramolecular vibrations, if present. The general case with translational and rotational degrees of freedom is illustrated in Fig. 3.1, B. The translational degrees are linked to the velocity of the water molecule's center of mass, while the rotational degrees are connected to the angular velocities ω_{rot}^{k} around the three distinct axes. The mass of the water molecule is used for the translational contributions, and the moments of inertia, I^{k}, are employed for the rotational contributions around each axis.

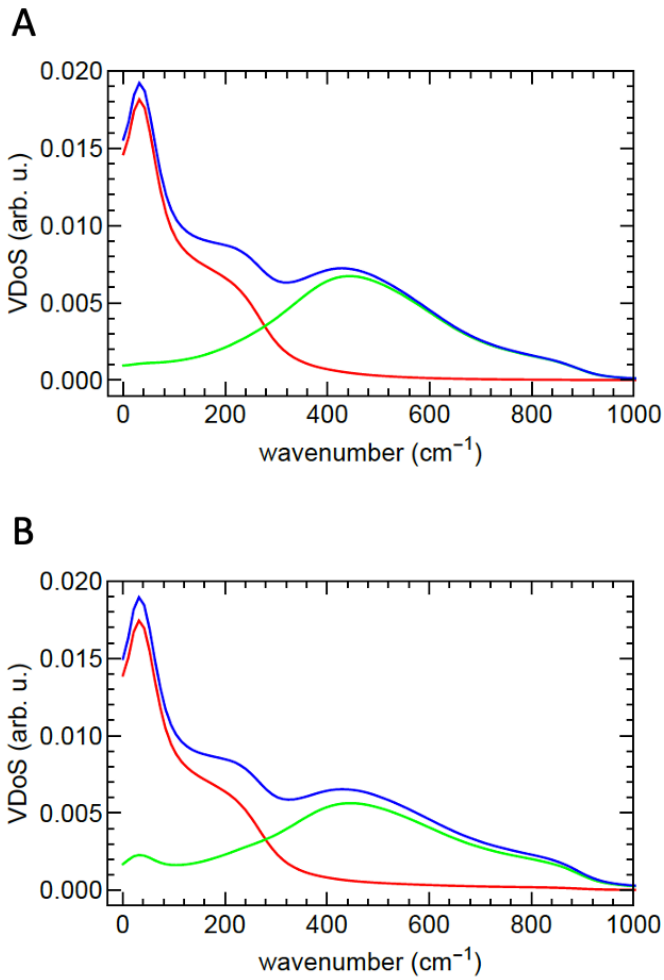

Figure 3.1: Vibrational Density of States (VDoS) of Bulk Water Decomposed into Different Contributions: A) The blue line represents the VDoS of a water molecule, while the red and green lines illustrate the contributions of oxygen and hydrogen atoms to the VDoS of a water molecule, respectively. B) The blue line indicates the VDoS of a water molecule, while the red and green lines depict the translational and rotational contributions to the VDoS of a water molecule, respectively. Image courtesy *V. Pattni*. [3]

In the upcoming sections of this chapter, we will discuss a method known as the two-phase thermodynamics (2PT) method [81, 82], along with its spatially resolved extension. This approach enables us to apply the 2PT formalism to water molecules distributed on a three-dimensional grid of voxels centered around a solute. This technique is instrumental in analyzing the thermodynamic properties associated with the process of solvation. Specifically, we will explore the 3D-2PT method developed by *Persson et al*. [83], which provides a detailed understanding of solvation thermodynamics. From the VDoS, it is possible to derive the system's total number of degrees of freedom. By integrating the area under the blue curve in Fig. **??**, the total degrees of freedom for a water molecule can be determined. Additionally, the integration of the red and green curves separately yields the translational and rotational degrees of freedom, respectively, each amounting to 3 [83]. It is important to note that when the VDoS is normalized, the total degrees of freedom for water remain constant, whether the water molecule is situated in the bulk or within the first hydration shell [175].

The vibrational density of states (VDoS) at zero frequency is directly associated with the self-diffusion coefficient D in pure fluids. The diffusion coefficient D can be determined from the velocity autocorrelation function using the following expression:

$$D = \frac{1}{3}\int_0^\infty C_v(t)dt = \frac{1}{6}\int_{-\infty}^{\infty} C_v(t)dt = \frac{1}{6mN}\int_{-\infty}^{\infty} C_v(t)dt \tag{3.3}$$

In this equation, m represents the mass of the particles, and N is the total number of particles. When zero frequency is substituted into Equation 3.2, the resulting expression becomes:

$$VDoS\,(0) = \frac{2}{k_B T}\int_{-\infty}^{\infty} C_v(t)dt = \frac{12mND}{k_B T}D = \frac{VDoS\,(0)k_B T}{12mN} \tag{3.4}$$

where the expression for diffusivity is:

$$D = \frac{VDoS(0)k_BT}{12mN} \tag{3.5}$$

The diffusivity of water molecules in bulk is greater than the ones interacting with a solute. This is indicated by the diffusion coefficient D. This suggests that the vibrational density of states (VDoS) at zero frequency for hydration water is lower than that of bulk water [179]. The property of the vibrational density of states (VDoS) enables the calculation of the total VDoS by incorporating the translational and rotational contributions from constrained molecules. This allows for the determination of the diffusion coefficient D for individual atoms. Additionally, the ability to differentiate between contributions to the total VDoS provides a systematic way to dissect thermodynamic properties such as entropy. This approach affords a more detailed and powerful tool for molecular analysis compared to earlier methods [177].

3.3 Two-Phase Thermodynamics Model

Developed by *Lin et al*. [81, 82], the Two-Phase Thermodynamics Model (2PT) leverages the VDoS of a liquid to determine molecular entropy S. The model is termed ”two-phase thermodynamics” because it employs a dual-phase approach to describe the VDoS of liquids, which can be challenging to compute quantitatively. This approach consists of a gas-like phase with diffusional dynamics and a solid-like phase with harmonic vibrations. Additionally, the computation of the total VDoS is segmented into translational and rotational degrees of freedom, as previously discussed, providing a more nuanced understanding of the system’s thermodynamic behavior. The translational motion of water is characterized by the center of mass velocity of the molecule, denoted as $\mathbf{v}_{\text{com}}$, while the rotational motion is described by angular velocities ω^k_{rot} around the three distinct rotational axes of a water molecule, each with their corresponding moments of inertia, I^k [83]. This

leads to the following equations:

$$VDoS_{\text{trans}}(\nu) = \frac{2}{k_B T}\int_{-\infty}^{\infty} e^{i\,2\pi\,\nu\,\tau} m_{\text{w}}\langle \mathbf{v}_{\text{com}}(t)\mathbf{v}_{\text{com}}(t+\tau)\rangle_t \,\mathrm{d}\tau \tag{3.6}$$

$$VDoS_{\text{rot}}(\nu) = \frac{2}{k_B T}\int_{-\infty}^{\infty} e^{i\,2\pi\,\nu\,\tau} \sum_{k=1}^{3} I^k\langle \omega^k_{\text{rot}}(t)\omega^k_{\text{rot}}(t+\tau)\rangle_t \,\mathrm{d}\tau \tag{3.7}$$

If the water vibrations are allowed by making the bonds flexible, then a third vibrational term appears in the VDoS calculations (Chapter 6). The vibrational motion of water is characterized by time-derivative of (intramolecular) vibrational coordinate q_v with reduced mass μ_v [178]. The additional $VDoS$ equation I_{vib} or $VDoS_{\text{vib}}(\nu)$ will be described as:

$$I_{\text{vib}}(\nu) = VDoS_{\text{vib}}(\nu) = \frac{2}{k_{\text{B}} T}\int_{-\infty}^{+\infty} \mathrm{e}^{i2\pi\nu\tau} \sum_{v=1}^{3} \mu_v\langle \dot{q}_v(t)\,\dot{q}_v(t+\tau)\rangle\, d\tau \tag{3.8}$$

In the Two-Phase Thermodynamics Model (2PT), the gas-like component represents the slow diffusive degrees of freedom. For translational motion, this is modeled using a hard sphere (HS) fluid model, while for rotational motion, a combination of a hard sphere fluid and rigid rotor (RR) model is used. Conversely, the solid-like components are applied to the fast vibrational degrees of freedom, allowing the vibrational density of states (VDoS) contributions from both translational and rotational degrees of freedom to be depicted by continuous spectra of harmonic oscillators (HO). This approach provides a detailed and nuanced understanding of the dynamic properties of liquids.

The entropy and partition functions for the slow diffusive anharmonic degrees of freedom in a liquid can be characterized using analytical expressions derived from the hard sphere (HS) fluid model, with the rigid rotor model applied for rotational degrees of freedom. Conversely, the fast vibrational degrees of freedom are handled within the context of classical or quantum harmonic oscillators (HOs), with the quantum HO being the default choice in the 3D-2PT model. The total entropy S of the liquid is then determined by summing up the contributions from these different components, as expressed in the following

equation:

$$\begin{aligned}\frac{S}{k_B} &= \frac{1}{k_B}(S^{\mathrm{HS}}_{\mathrm{trans}} + S^{\mathrm{HO}}_{\mathrm{trans}} + S^{\mathrm{HS}}_{\mathrm{rot}} + S^{\mathrm{HO}}_{\mathrm{rot}}) \\ &= \int W^{\mathrm{HS}} VDoS^{\mathrm{HS}}_{\mathrm{trans}} \omega d\omega + \int W^{\mathrm{HO}}(\omega) VDoS^{\mathrm{HO}}_{\mathrm{trans}} \omega d\omega \\ &+ \int W^{\mathrm{RR}} VDoS^{\mathrm{HS}}_{\mathrm{rot}} \omega d\omega + \int W^{\mathrm{HO}}(\omega) VDoS^{\mathrm{HO}}_{\mathrm{rot}} \omega d\omega \end{aligned} \tag{3.9}$$

Equation 3.9 illustrates that each component of the entropy is represented by a weighted integral of the corresponding vibrational density of states (VDoS). For the slow diffusive rotational degrees of freedom, the hard sphere (HS) model is substituted with the rigid rotor expression to better capture the rotational dynamics. It's important to note that this expression does not provide the exact entropy of the liquids. Instead, it approximates the entropy through a linear superposition of two limiting models that describe the slow and fast degrees of freedom, respectively.

As depicted in Fig. 3.2, the solid-like VDoS accounts for approximately 71% of the total entropy S, whereas the gas-like phase contributes about 29%, indicating the predominant role of the solid-like component in the overall entropy of the system. A fluidicity parameter, f, has been introduced to categorize the distribution of the liquid vibrational density of states (VDoS) into gas-like and solid-like components for any given system. This parameter must adhere to two limiting conditions: i). At high temperatures and/or low densities, the system behaves like a hard-sphere fluid, and the fluidicity factor $f = 1$, indicating the absence of a solid-like component. ii). At high-density limits, the system transitions into a solid phase where $f = 0$, signifying the absence of a gas-like component. The solution that satisfies both of these conditions yields a fluidicity parameter that is directly proportional to the diffusivity of the system.

$$f = \frac{D(T,\rho)}{D_0^{\mathrm{HS}}(T,\rho,\sigma^{\mathrm{HS}})} \tag{3.10}$$

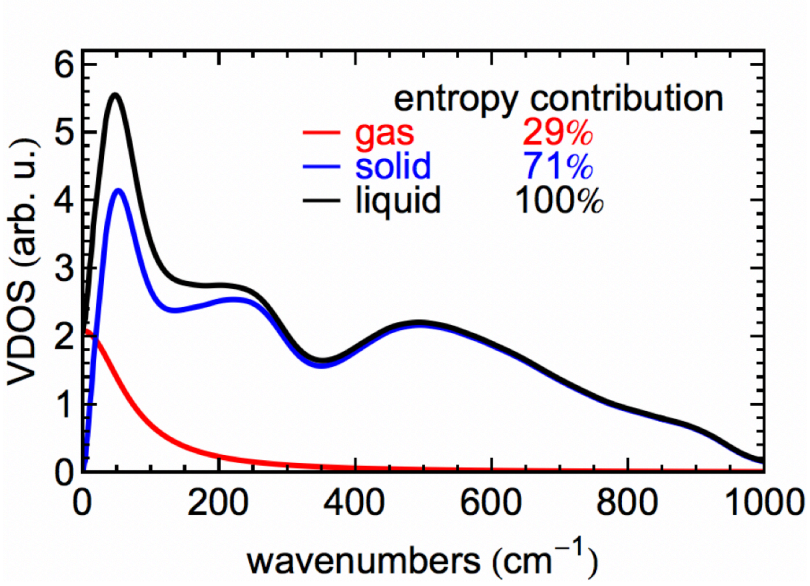

Figure 3.2: The molecular entropy S can be computed by combining the gas-like (Red) and solid-like (Blue) vibrational density of states (VDoS) to form the liquid-like (Black) VDoS through linear superposition. Image courtesy *M. Heyden*.

In this expression, D represents the self-diffusivity coefficient of the simulated molecules as defined in Equation 3.5, while D_0^{HS} refers to the diffusivity of hard spheres, calculated using the Chapman-Enskog theory [180] at the zero-pressure limit.

$$D_0^{\mathrm{HS}}(T,\rho,\sigma^{\mathrm{HS}}) = \frac{3}{8}\frac{1}{\rho\sigma^{\mathrm{HS}2}}\left(\frac{k_B T}{\pi m}\right)^{\frac{1}{2}} \tag{3.11}$$

The fluidicity factor f for translational and rotational motions is given by the following equation, which is derived from Eq. 3.10 by *Lin et al* [81, 82]:

$$2\Delta^{-\frac{9}{2}}f^{-\frac{15}{2}} - 6\Delta^{-3}f^{5} - \Delta^{-\frac{3}{2}}f^{\frac{7}{2}} + 6\Delta^{-\frac{3}{2}}f^{\frac{5}{2}} + 2f - 2 = 0 \tag{3.12}$$

In this equation, Δ represents a dimensionless normalized diffusivity constant that depends on material properties such as temperature T, bulk number density ρ, mass m, and the diffusivity of the $VDoS$ at zero frequency, $VDoS(0)$, of the system:

$$\Delta(T,\rho,m,VDoS(0)) = \frac{2VDoS(0)}{9N}\left(\frac{\pi k_B T}{m}\right)^{\frac{1}{2}}\rho^{\frac{1}{3}}\left(\frac{6}{\pi}\right)^{\frac{2}{3}} \tag{3.13}$$

The fluidicity factors for translational (f_{trans}) and rotational (f_{rot}) motions can be used to calculate the respective hard sphere vibrational density of states ($VDoS^{\text{HS}}_{\text{trans}}$ and $VDoS^{\text{HS}}_{\text{rot}}$) using the following expression:

$$VDoS^{\text{HS}}_{\text{trans}}(\nu) = \frac{VDoS_{\text{trans}}(0)}{1+\left[\frac{\pi VDoS_{\text{trans}}(0)\nu}{6f_{\text{trans}}N}\right]^2} \tag{3.14}$$

$$VDoS^{\text{HS}}_{\text{rot}}(\nu) = \frac{VDoS_{\text{rot}}(0)}{1+\left[\frac{\pi VDoS_{\text{rot}}(0)\nu}{6f_{\text{rot}}N}\right]^2} \tag{3.15}$$

To calculate the harmonic oscillator (HO) vibrational density of states ($VDoS^{\text{HO}}_{\text{trans/rot}}$) for translational and rotational motions, we subtract the translational and rotational $VDoS$ of a hard-sphere fluid from the corresponding $VDoS$ of a water molecule. In Equation 3.9, weighting functions are used to account for different contributions to the entropy. The solid-like component, represented as a set of continuous harmonic oscillators HO, has a weighting function W^{HO} derived from the frequency-dependent partition function of the quantum harmonic oscillator:

$$W^{\text{HO}}\nu = \frac{\beta h\nu}{exp(\beta h\nu)-1} - ln[1-exp(-\beta h\nu)] \tag{3.16}$$

In this equation, $\beta = 1/k_BT$, where k_B is the Boltzmann constant, and T is the temperature. The Planck's constant is denoted by h. The weighting functions W^{HS} for the gas-like diffusive component, described as a hard sphere (HS) fluid for translational motions and by a rigid rotor for rotational motions, are given as follows:

$$W^{\text{HS}}(\nu) = \frac{1}{3}\frac{S^{\text{HS}}}{\text{k}_\text{B}} \tag{3.17}$$

$$W^{\text{RR}}(\nu) = \frac{1}{3}\frac{S^{\text{RR}}}{\text{k}_\text{B}} \tag{3.18}$$

here, S^{HS} represents the entropy of the hard sphere fluid, which is determined by the Carnahan-Starling equation of state [181]:

$$\frac{S^{\text{HB}}}{k_b} = \frac{5}{2}ln\left[\left(\frac{2\pi mk_BT}{h^2}\right)^{\frac{3}{2}}\frac{z(y)}{\rho f_{\text{trans}}}\right] + \frac{y(3y-4)}{(1-y)^2} \tag{3.19}$$

In this equation, the hard sphere packing fraction is represented by $y = f_{\text{trans}}^{5/2}/\Delta^{3/2}$, while the compressibility factor, $z(y)$, can be calculated using the Carnahan-Starling equation of state for hard sphere gases [181].

$$z(y) = \frac{1 + y + y^2 - y^3}{(1 - y)^3} \tag{3.20}$$

The entropy of rotation, S^{RR}, of a rigid-rotor that can be calculated as follows:

$$\frac{S^{\text{RR}}}{k_B} = ln\left[\left(\frac{\sqrt{\pi e^3}}{\sigma}\right)^2 \frac{T^3}{\sqrt{\Theta_a \Theta_b \Theta_c}}\right] \tag{3.21}$$

In this expression, σ represents the symmetry number, which indicates the number of ways a molecule can be arranged to produce indistinguishable configurations. For a rigid water molecule, $\sigma = 2$. The rotational temperatures Θ_a, Θ_b, Θ_c are related to the corresponding moments of inertia I_i according to the following relationships:

$$\Theta_i = \frac{h^2}{8\pi^2 I_i k_B} \tag{3.22}$$

3.4 Spatial Resolution with 3D-2PT

The 3D-2PT method, developed by *Persson et al*. [83], is an extension of the two-phase thermodynamics (2PT) theory originally formulated by *Lin et al*. [81, 82] for bulk liquids. This approach calculates entropic contributions to the solvation free energy by applying the 2PT formalism to each volume element within a 3D analysis grid. The 3D-2PT method enhances existing techniques for analyzing solvation free energies from explicit solvent simulations by providing a more detailed spatial resolution of entropic effects. Thermodynamic integration (TI) [182] is a well-established method for calculating total solvation free energies. While the 3D-2PT approach has been demonstrated to yield results consistent with TI simulations, it offers additional insights by providing spatially resolved local contributions of solvation enthalpies and entropies. Furthermore, 3D-2PT can differentiate between the effects of solute-solvent interactions and the compensating changes in solvent-solvent interactions, known as solvent reorganization.

To utilize the 3D-2PT method, a three-dimensional cubic grid is established, with the solute molecule positioned at the center, as illustrated in Fig. 3.3. For the analysis grid in 3D-2PT, smaller systems typically employ a grid size of 64x64x64 volume elements (voxels) with a voxel size of 0.5 Å (used in Ch. 6), while larger systems may use a grid size of 128x128x128 with a voxel size increased to 1.0 Å (used in Ch. 5).

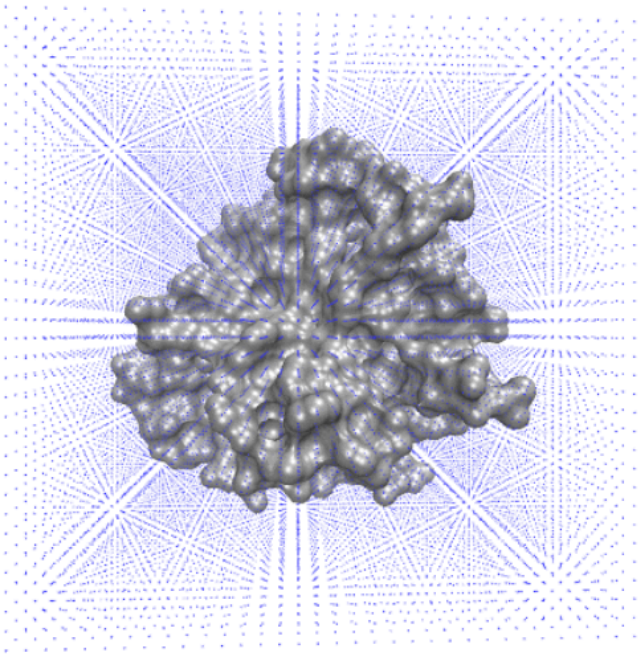

Figure 3.3: Three-dimensional (3D) grid centered on the solute molecule for 3D-2PT analysis. Image courtesy *V. Pattni*.

In the 3D-2PT technique, the entropy per water molecule is calculated for each voxel in the analysis grid mentioned earlier. While rotational and translational degrees of freedom are analyzed separately (unless vibrations are included as discussed in section 3.3), the total contribution to the solvation entropy per water molecule is reported exclusively. To determine the total solvation entropy, the local contribution to the solvation entropy $\Delta S_{\text{solv}}(\mathbf{r})$ per

water molecule in each voxel is assessed. Then, all local contributions are spatially integrated, using the local number density as a weighting factor, to account for the non-uniform distribution of water around the solute.

$$\Delta S_{\text{solv}} = \int \Delta S_{\text{solv}}(\mathbf{r}) n_w(\mathbf{r}) d\mathbf{r} = \int [S(\mathbf{r}) - S^{\text{bulk}})] n_w(\mathbf{r}) d\mathbf{r} \tag{3.23}$$

The local solvation entropy contribution $\Delta S_{\text{solv}}(\mathbf{r})$ can be defined as the deviation of the entropy $S(\mathbf{r})$ in a given voxel from the average bulk water entropy S^{bulk}. In this thesis, S^{bulk} is determined from voxels that are at least 10 Å away from the solute atoms, as illustrated in Fig. 3.4. The symbol $n_{\text{w}}(\mathbf{r})$ represents the spatially resolved water number density.

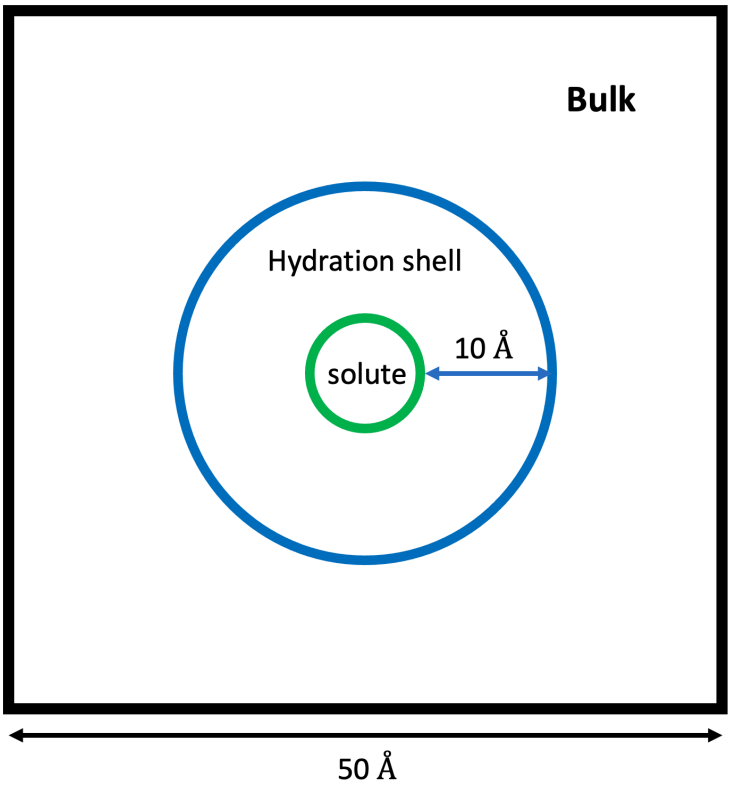

Figure 3.4: Depiction of a simulation box measuring 50 Å on each side, with a solute molecule positioned at the center. Surrounding the solute, there is a hydration layer defined by a radius of 10 Å from the solute's surface. This layer represents the region where water molecules interact directly with the solute. Beyond this hydration layer, the remaining space in the simulation box is considered the bulk region, where water molecules exhibit bulk-like behavior without direct influence from the solute.

In addition to computing spatially resolved local contributions of solvation entropies, the 3D-2PT approach can also calculate solvation enthalpies. However, this calculation is more straightforward and does not require the 2PT approach, as it involves analyzing pairwise intermolecular potentials. The enthalpy is computed using the same analysis grid, where the local solute-solvent potential energy $U_{\mathrm{uv}}(\mathbf{r})$ and solvent-solvent potential energy $U_{\mathrm{vv}}(\mathbf{r})$ are evaluated with spatial resolution as an average over all NVE production simulations. The local contribution to the solvation enthalpy $H_{\mathrm{solv}}(\mathbf{r})$ per water molecule can be determined as follows:

$$\begin{aligned} \Delta H_{\mathrm{solv}}(\mathbf{r}) &= U_{\mathrm{uv}}(\mathbf{r}) + \frac{1}{2}\left[U_{\mathrm{vv}}(\mathbf{r}) - U_{\mathrm{vv}}^{\mathrm{bulk}}\right] \\ &= U_{\mathrm{uv}}(\mathbf{r}) + U_{\mathrm{vv}}(\mathbf{r}) \end{aligned} \tag{3.24}$$

The average bulk solvent-solvent potential energy $U_{\mathrm{vv}}^{\mathrm{bulk}}$ is determined from voxels that are at a minimum defined distance from the closest atom of the solute molecule, which in this study was set to 10 Å. The factor of $1/2$ in the second term accounts for the double counting of water-water interactions when spatially integrating to obtain the total solvation enthalpy.

$$\Delta H_{\mathrm{solv}} = \int \Delta H_{\mathrm{solv}}(\mathbf{r})\, n_{\mathrm{w}}(\mathbf{r})\, \mathrm{d}\mathbf{r} \tag{3.25}$$

The local water number density $n_{\mathrm{w}}(r)$ is incorporated into the calculation to account for the varying distribution of water molecules in the vicinity of the solute. This adjustment is necessary to accurately represent the spatial variation of water concentration around the solute in the analysis.

The solvation free energy can be calculated as the sum of the solvation enthalpy and the solvation entropy, representing the total energetic and entropic contributions to the solvation process.

$$\Delta G_{\mathrm{solv}} = \Delta H_{\mathrm{solv}} - T\Delta S_{\mathrm{solv}} \tag{3.26}$$

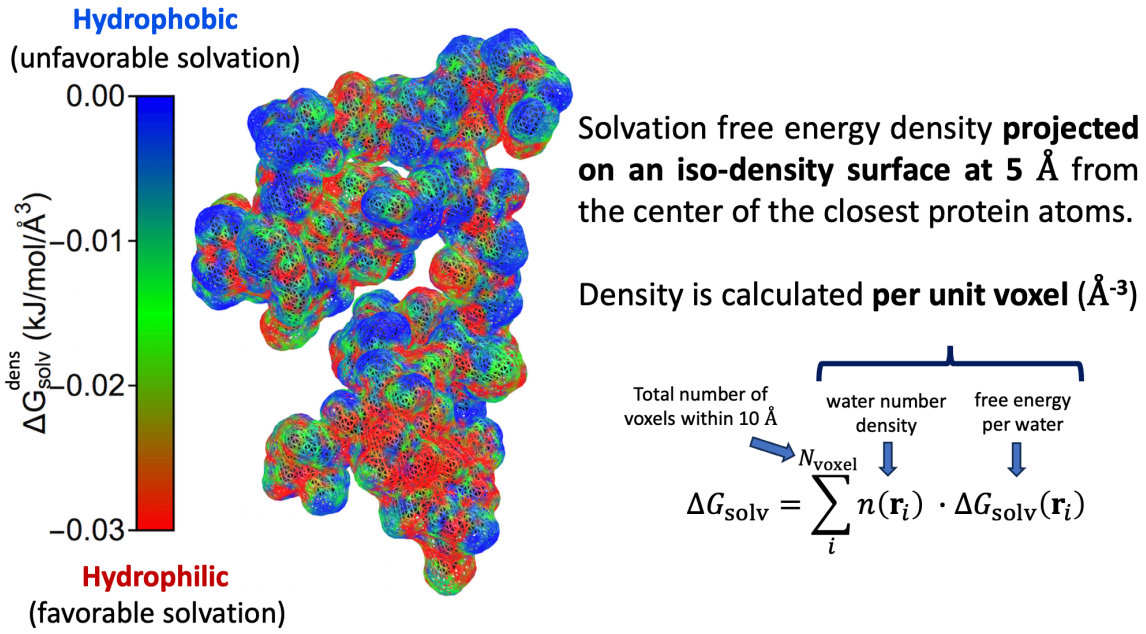

Figure 3.5: Solvation free energy density map of a fixed conformation of the K-18 domain of the Tau protein obtained from 3D-2PT analysis using the force field-water model combination of Amber99SB*-ILDN-TIP4P-D.

A three-dimensional map depicting local solvation free energy contributions due to solute-solvent interactions can be characterized using the spatially resolved solvation free energy density, $\Delta G^{\text{dens}}_{\text{solv}}(\mathbf{r})$, as shown in Figure 3.5. Additionally, the solvation free energy, ΔG_{solv}, can be represented as a weighted integral of local solvation free energy contributions, $\Delta G_{\text{solv}}(\mathbf{r})$ (per water molecule), or solvation free energy densities, $\Delta G^{\text{dens}}_{\text{solv}}(\mathbf{r})$ (per volume). This is achieved by integrating equations Eq. (3.23) and Eq. (3.24) with Eq. (3.25).

$$\begin{aligned} \Delta G_{\text{solv}} &= \int [\Delta H_{\text{solv}}(\mathbf{r}) - T\Delta S_{\text{solv}}(\mathbf{r})]\; n_{\text{w}}(\mathbf{r})\, \mathrm{d}\mathbf{r} \\ &= \int \Delta G_{\text{solv}}(\mathbf{r})\, n_{\text{w}}(\mathbf{r})\, \mathrm{d}\mathbf{r} \\ &= \int \Delta G^{\text{dens}}_{\text{solv}}(\mathbf{r})\, \mathrm{d}\mathbf{r} \end{aligned} \tag{3.27}$$

Chapter 4

RESEARCH OVERVIEW

This thesis primarily investigates the dynamics and solvation thermodynamics of intrinsically disordered proteins (IDPs). The lack of a well-defined structure in IDPs and the challenges associated with experimental studies make direct comparisons between simulations and experiments difficult. The absence of a single standard and accurate force field complicates in-silico studies of IDPs. Various force fields with enhanced protein-water interactions can replicate some properties of IDPs that align closely with experimental data. However, each force field introduces unique modifications that result in an extended conformational ensemble. Therefore, a quantitative analysis of solvent-mediated interactions and the associated thermodynamics is essential to understand the impact of each force field and how it leads to extended conformations. This analysis would facilitate the selection of the most suitable force field based on the specific characteristics of the IDP system being simulated. Next, it has been observed that biomolecules, particularly charged intrinsically disordered proteins (IDPs), exhibit a positive response to external electric fields. Given the large size of IDPs, employing quantum calculations such as *ab initio* molecular dynamics (AIMD) is not practical. To indirectly study the effects of electric fields on IDPs, both AIMD and classical molecular dynamics (MD) are utilized to investigate the response of bulk water to varying strengths of electric fields. Finally, protein folding, as well as protein-protein and protein-ligand interactions, are crucial for the sustenance of all life forms. Intrinsically disordered proteins (IDPs), which constitute about 40% of the human proteome, play a significant role in these essential cellular processes. I conducted a quantitative analysis of the folding free energies of several IDPs to differentiate between fully disordered regions and regions with a low energy penalty for folding into an alpha helix, known as alpha-MoRFs.

I then focused on three specific alpha-MoRFs that act as pro-apoptotic proteins binding to the anti-apoptotic protein MCL-1. Studying such systems provides deeper insights into the properties favorable for an alpha-MoRF and can aid in the design of future peptide-based drugs targeting disease-causing proteins.

In the initial project, we conducted a comparative analysis of the local and total solvation thermodynamics surrounding the K-18 domain of the intrinsically disordered protein (IDP) Tau. This analysis was performed using simulations with four pairs of modified and standard force fields. IDPs are proteins lacking a fixed structure, playing crucial roles in cellular signaling, regulation, and recognition in humans (extensively described in Chapter 1). Empirical force fields employed in molecular dynamics (MD) simulations often exhibit an imbalance between intramolecular interactions within the protein and intermolecular interactions between the protein and water. This imbalance can result in the artificial collapse of IDP structures in simulations, leading to discrepancies with experimental observations. To mitigate this issue, various approaches have been developed to refine the models of protein-water interactions, thereby enhancing the accuracy of simulations. In this research, we utilized both standard and modified force fields in our simulations to closely examine the impact of each modification on the solvation free energy. This investigation aimed to shed light on the effectiveness of different force field adjustments in accurately representing the solvation thermodynamics of IDPs.

The traditional structure-function paradigm in molecular and structural biology is confronted by the existence of intrinsically disordered proteins (IDPs) [12, 13, 17, 18]. These proteins, characterized by their unstructured domains, make up approximately 33% of the total eukaryotic proteome, with the proportion in humans exceeding 40% [28, 183]. IDPs play diverse roles in biological processes, including cellular signaling [23], regulation, and molecular recognition [32, 184]. Intrinsically disordered proteins are linked to a range of diseases, such as cancer, amyloidosis, and neurodegenerative disorders [38, 40–43]. Un-

derstanding the pathological behavior of IDPs and their relationship with sequence, function, and features like liquid-liquid phase separation requires insight into how their conformational ensembles in solution are influenced by interactions with the molecular environment. Techniques such as small-angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR), and Förster resonance energy transfer (FRET) [48–50] can characterize the structural and dynamical properties of IDPs. Additionally, atomistic molecular dynamics (MD) simulations offer a microscopic view of these properties [51–53]. However, standard empirical force fields used in MD simulations often fail to replicate experimental observations, leading to overly compact conformations of solvated IDPs. This has prompted the development of IDP-specific force field parameters that adjust dihedral angles to modify the conformational ensembles in simulations. The limitations of standard force fields in describing IDPs are primarily attributed to an imbalance between non-covalent interactions within the protein, which favor compact states, and protein-solvent interactions, which favor extended conformations with a higher solvent-accessible surface area [51, 52, 54].

To improve the accuracy of standard force fields in simulating experimental results, various modifications have been implemented [53]. One technique involves combining AMBER protein force fields [185–187] with the TIP4P/2005 water model [188] and increasing the Lennard-Jones interaction energies between water oxygens and protein atoms by 10% [52, 52, 189, 190]. Another approach involves scaling LJ interactions between water hydrogens and protein atoms for the CHARMM36m force field [191] in conjunction with the TIPS3P water model [66](a variation of the TIP3P water model [192] employed in CHARMM, which incorporates Lennard-Jones sites for water hydrogen atoms). Another strategy employed by Shaw and colleagues was reparameterizing the TIP4P water model and adjusting parameters in AMBER-type protein force fields to enhance protein-solvent interactions [51, 193–195]. Additionally, innovative water models that prioritize electrostatics and charge distributions have shown promising outcomes [196, 197].

In each of the aforementioned scenarios, altering protein-water interactions leads to extended conformations for IDPs in solution that are more in line with experimental data. However, the impact of these modifications on the thermodynamics of protein solvation has been mainly studied for side-chain analogs [52]. There needs to be more in-depth analysis that includes the breakdown of local enthalpic and entropic contributions to the solvation free energies of proteins. Understanding these details is crucial for grasping the microscopic effects of the modified protein-water interactions, which are not apparent when only compared to structural parameters obtained from experiments.

After simulating the intrinsically disordered K-18 domain of the Tau protein [48, 198] using various combinations of standard and modified protein and water force fields, we examined the conformational ensembles produced by each simulation model and assessed the protein solvation free energy, along with its enthalpic and entropic components, for a specific conformational state with spatial resolution using the 3D-two-phase thermodynamics (3D-2PT) approach [167, 199] (described in Chapter 3). Preferences in enhancing the solvation of polar and nonpolar groups significantly influence the conformational ensembles generated by standard and modified force field parameters. A03ws and A99ws simulations tend to favor extended and solvent-exposed states for nonpolar residues, whereas A99D and C36m simulations prefer similar states for polar residues. In IDPs containing both polar and nonpolar residues, such differences might not be evident in overall conformational measures like the radius of gyration or end-to-end distance. However, these distinctions become more prominent in proteins predominantly composed of either polar or nonpolar amino acids. Despite better agreement between simulations with modified force fields and experimental findings, IDPs and proteins with substantial intrinsically disordered regions (IDRs) continue to pose challenges for atomistic simulations.

The second research topic explores the impact of varying strengths of static electric fields on the entropy of bulk water through classical molecular dynamics (MD) and *ab initio* molecular dynamics (AIMD) simulations [200, 201]. The application of external electric fields to liquid water can lead to a variety of phenomena that have important implications in the realms of electrochemistry and hydrogen-based technologies. Furthermore, a recent study[202] has investigated the impact of electric fields on intrinsically disordered proteins (IDPs) associated with neurodegenerative diseases. The application of oscillating oriented external electric fields (OEEFs) resulted in the irreversible disintegration of β-amyloid aggregates, providing valuable insights into potential treatment strategies for Alzheimer's disease and other neurodegenerative disorders marked by senile plaques. This highlights the wide-ranging applications of electric field effects, spanning from the study of water properties to the development of therapies for neurodegenerative diseases. The effects of external fields can include changes in the molecular structure of water, alterations in the dynamics of hydrogen bonding, and shifts in the solvation properties of ions and molecules in water. While some research efforts have been directed towards elucidating the thermodynamic properties associated with the application of electric fields in aqueous systems, a comprehensive understanding of these effects remains incomplete. Specifically, the impact of electric fields on the total and local entropy of bulk water, which are critical for understanding the disorder and energy distribution within the system, has not been reported in the scientific literature to date. This gap in knowledge represents a significant opportunity for further investigation, as understanding these entropy changes could provide valuable insights into the behavior of water under the influence of electric fields and its implications for various technological applications.

While earlier investigations have examined the impact of electric fields on the equilibrium between solid and liquid phases [203], more recent studies have focused on the changes in entropy caused by the reorganization of hydrogen bonds in water at electrified

interfaces, as revealed through multiscale simulations [204]. In particular, Estejab and colleagues utilized both classical molecular dynamics (MD) and *ab initio* molecular dynamics (AIMD) simulations to explore the solvation thermodynamics of different surface species, postulating that local electric fields could reorient the molecules of hydration water [204]. A key finding from their work was that the entropy of solvation is dependent on the electric field's strength and orientation. Nonetheless, the complexity of the interaction between the electric field and surface phenomena at water/electrode interfaces posed a challenge to their study. Additionally, a significant portion of computational research has been devoted to understanding the hydrophobic effect [205], whose underlying molecular mechanisms are influenced by local charges and, therefore, by electric fields [206, 207]. So far, there were no reported studies that explored the effects of electric fields on the local entropy of bulk water in either a qualitative or quantitative manner.

The primary objective of this study was to investigate the effects of static and homogeneous electric fields on the structure and entropy of bulk liquid water utilizing molecular dynamics (MD) simulations, both classical (force-field) and *ab initio* (AIMD). Additionally, we aimed to analyze the consequences of local entropy descriptors that provided a spatially-resolved quantification of entropy. Specifically, we employed an MD-based analysis method known as 3D-two-phase thermodynamics (3D-2PT) [199, 208] to spatially resolve the entropy per water molecule and the local number density of bulk water under an electric field, as well as their changes compared to the zero-field regime. My contribution to this project was performing the simulation and analysis part involving classical MD with TIP4P/2005 [188], and the 3D-2PT analysis part. In our investigation of local entropy indicators, we examined the bulk water environment in two different scenarios at various field intensities: (i) the ideal system, where a water molecule was perfectly aligned along the electric field (referred to as the "fixed" system), and (ii) the more realistic system, where a water molecule was allowed to freely rotate around a fixed point in the bulk liquid (referred

to as the ”pinned” system). The ideal system provided insights into the complex water environment along specific directions of the perfectly aligned molecule, such as bond vectors and dipole moments. In contrast, the more realistic system was essential for understanding the response of the central water molecule to the externally imposed electric field, including the combined effects of field alignment and local structural preferences.

In the general findings, without an external field, a comparison between the rigid force-field TIP4P/2005 and *ab initio* functionals revealed that the TIP4P/2005 water model, which is computationally efficient, adequately describes key structural properties of water. This model's structure closely aligns with the results from AIMD simulations using the more accurate SCAN+rVV10 functional. Additionally, the molecular entropy of bulk water estimated for the TIP4P/2005 model using the two-phase thermodynamics (2PT) formalism was consistent with the corresponding result from AIMD simulations with the SCAN+rVV10 functional. However, both were approximately 15 J $mol^{-1}K^{-1}$ lower than the experimental reference value of 69.95 J $mol^{-1}K^{-1}$. This discrepancy could be due to a systematic underestimation of water entropy by the 2PT formalism, which relies on the vibrational density of states. Furthermore, the molar entropy of TIP4P/2005 water showed significantly less sensitivity to polarization-induced changes compared to BLYP+D3 water. This is likely because the latter includes electronic polarization, which affects its response to changes in the environment.

In further results, the MD-based analysis (3D-2PT) [199, 208] provided detailed spatial information on the local entropy and number density of water surrounding a reference molecule. This approach enabled the creation of three-dimensional maps depicting the alterations in water structure caused by electric fields. This technique enabled the creation of three-dimensional maps showcasing the alterations in water structure caused by electric fields. Through this method, we observed a field-induced enhancement of hydrogen bonds between the reference molecule and its first neighbors along specific directions. This obser-

vation aligns with the vibrational Stark effect, suggesting that liquid water tends to develop more ”ice-like” structures when exposed to electric fields at intensities of approximately 1.0 V nm^{-1} [200]. The formalism of 3D-two-phase thermodynamics (3D-2PT) was instrumental in revealing that changes in the local entropy density due to an electric field are intricately linked to alterations in the water number density. This allowed us to construct a detailed molecular-level view of how the field influences entropy. The primary effect of a static electric field on water's number density is to promote a more structured layering, particularly an increased coordination in the second shell, along with anisotropic structural changes. In the presence of strong electric fields, water molecules increasingly align their dipoles and arrange themselves into preferred orientations parallel to the field, forming hydrogen-bonded structures mainly along a diagonal direction relative to the field axis.

In the final research topic, we delved into the energetics of specific peptide sequences known as alpha-Molecular Recognition Features (alpha-MoRFs) [209–214], which undergo transitions from intrinsically disordered states to structured secondary motifs like the alpha-helix. The relatively low folding free energy penalty associated with these sequences presents an opportunity for the development of peptide-based or small molecule drugs that can stabilize the helical structure by inducing folding upon binding to alpha-MoRFs. We conducted a comparative study between alpha-MoRFs and completely disordered sequences from the same well-characterized proteins. Our comparison included an analysis of minimum energy structures obtained from umbrella sampling simulations and those predicted by AlphaFold, a leading structure prediction model, to benchmark our results. Furthermore, we explored the binding free energies of various alpha-MoRFs in their folded states to assess whether experimental binding free energies accurately capture the separate contributions of the folding and binding processes, as determined from our umbrella sampling simulations. This comprehensive analysis aimed to deepen our understanding of the energetics underlying the interactions between alpha-MoRFs and their binding partners.

Intrinsically disordered proteins (IDPs) [18, 20] lack a stable three-dimensional structure, often due to their composition of charged residues, glycines, and prolines. Despite this structural flexibility, IDPs are integral to vital cellular functions, including transcription regulation, signal transduction, and the formation of multiprotein complexes. About 35-40% of the human proteome consists of IDPs or proteins with significant disordered regions. A critical aspect of IDPs is their ability to interact with globular proteins, either by adopting a specific conformation upon binding or by remaining mostly disordered. Key functional regions within IDPs, known as Molecular Recognition Features (MoRFs) or Molecular Recognition Elements (MoREs), facilitate these interactions. MoRFs can transition from disorder to order during binding and are essential for regulatory processes and signal transduction [31], despite their flexible structure.

Among the various types of Molecular Recognition Features (MoRFs), alpha-helix forming molecular recognition elements/features (α-MoREs/MoRFs), also known as helical binding motifs (HBMs), are the most common and well-studied. These sequences typically consist of short amphipathic helices that transition from a disordered state to a helical conformation upon binding to their target partners. They can be predicted based on their sequence and are particularly prevalent in eukaryotic transactivation domains [215] and $\alpha\alpha$-hub proteins [216]. In systems where folding occurs upon binding, experimental research has shown that the presence of a helical structure can affect both the binding affinity and the kinetics. However, more investigation is required to comprehend how the sequence of an intrinsically disordered protein (IDP) influences its structural ensemble, both when it is free and when it is bound, and how this influences binding characteristics, encompassing both kinetic and equilibrium aspects. The use of enhanced sampling techniques, such as umbrella sampling, for measuring folding free energy has not been thoroughly explored, primarily due to the difficulty in selecting a suitable collective variable to depict the unfolding process. Furthermore, the detailed effects of point mutations, specifically those

involving helix stabilizers like alanine and helix disruptors such as proline and glycine, on the overall folding free energy remain to be elucidated.

In tackling these uncharted areas, we applied atomistic umbrella sampling to examine the folding free energies of disordered peptides that transition into an alpha-helical structure upon binding to their targets, including peptides with alpha-Molecular Recognition Features (α-MoRFs) and entirely disordered intrinsically disordered regions (IDRs). By comparing the structures with minimum free energy obtained from our simulations with those predicted by AlphaFold, we observed that all the α-MoRF examples we analyzed had folding free energies within a range of 10 kJ/mol (2.4 kcal/mol). This finding is in agreement with the predictions made by Hadzi *et al*. [211, 217]. in their research, which utilized a modified LR model. They accounted for varying lengths of IDP sequences, which could explain the differences between our results and theirs. The relatively low folding free energy penalty of these sequences offers potential for the development of peptide-based or small molecule drugs that can stabilize the helical structure by promoting folding upon binding to α-MoRFs.

Utilizing molecular dynamics simulations and the umbrella sampling method, we first established the credibility of our protocol by examining sequences from the c-Myc protein, observing that the folding free energies were consistent with expected values. Subsequently, we mutated helix-destabilizing elements with alanine in selected c-Myc sequences, further reinforcing the reliability of our approach. These preliminary validations set the stage for a comprehensive exploration of naturally occurring intrinsically disordered regions (IDRs) and alpha-Molecular Recognition Features (α-MoRFs). In this expanded analysis, we leveraged our protocol alongside the predictive capabilities of AlphaFold. Our findings revealed that AlphaFold was proficient in predicting the alpha-helical structures of α-MoRFs, notably in the case of the two naturally occurring segments of α-MoRFs (alpha-c-Myb-D, and alpha-p53-CT-D), and could identify sequences as fully disordered up to a certain threshold, which varied depending on sequence length.

In the successive section, we estimated the binding free energies of two selected α-MoRFs from BH3 domain-binding proteins (PUMA and NOXA-A) to Mcl-1 using coarse-grained (CG) simulations with the SIRAH force field. Our results provide partial support for our hypothesis that the experimental binding free energies reported by Dahal *et al*. [218] are the sum of the calculated binding and folding free energies. This finding sheds light on the potential binding-induced folding pathway exhibited by our chosen set of α-MoRFs. The calculation of binding free energies revealed deviations from experimental results, indicating the possible involvement of a positive folding free energy component in the binding-induced folding process. This hypothesis gained support from our analysis of protein complexes involving PUMA and NOXA-A, where the determined folding free energies were notably lower than those needed to align with experimental observations. These differences may arise from several sources, such as variations in sequence length between our simulations and experimental research, the employment of mutated versions in experiments that are distinct from the structures utilized in our simulations, and differences in overall sequence composition. These elements could significantly influence the folding behavior of the proteins, resulting in the observed discrepancies between our computational findings and experimental data. Addressing these factors in future research could aid in bringing our computational predictions into closer agreement with experimental results.

Chapter 5

MODEL-DEPENDENT SOLVATION OF THE K-18 DOMAIN OF THE INTRINSICALLY DISORDERED PROTEIN TAU

5.1 Summary

A known imbalance between intra-protein and protein-water interactions in many empirical force fields results in collapsed conformational ensembles of intrinsically disordered proteins in explicit solvent simulations, which disagrees with experiments. Multiple strategies have been introduced in the literature to modify protein-water interactions, which improves agreement between experiments and simulations. In this work, we combine simulations with standard and modified force fields with a spatially resolved analysis of solvation free energy contributions and compare the consequences of each strategy.

We find that enhanced Lennard-Jones interactions between protein atoms and water oxygens primarily improve the solvation of nonpolar functional groups of the protein. In contrast, modified electrostatics in the water model or strengthened Lennard-Jones interactions between the protein and water hydrogens mainly affect the hydration of polar functional groups. Modified electrostatics further impact the average orientation of water molecules in the hydration shell. As a result, protein-water interactions with the first hydration layers are strengthened, while interactions with water molecules in higher hydration shells are weakened.

Hence, distinct strategies to balance intra-protein and protein-water interactions in sim-

ulations have qualitatively different effects on protein solvation. These differences are not necessarily captured by comparisons to experiments that report on global parameters describing protein conformational ensembles, e.g., the radius of gyration, but will influence the tendency of a protein to form aggregates or phase-separated droplets.

5.2 Introduction

While the structural information for folded proteins continues to grow[220–223] and *de-novo* predictions are feasible[224–226], the discovery of intrinsically disordered proteins (IDPs) that lack folded domains[12, 13, 17, 18] challenges the structure-function paradigm in molecular and structural biology. In addition to IDPs many proteins with folded domains contain extensive intrinsically disordered regions (IDRs), which are required for function[14, 15]. IDPs and IDRs constitute about 33% of the total eukaryotic proteome, while in humans this fraction is found to be closer to 50%[27, 28]. The expression of IDPs is often tightly regulated in the cell[227] and they exhibit a wide range of biological functions such as cellular signaling[23] and molecular recognition[32, 184]. The biological function of some IDPs and IDRs can be tied to folding or partial folding, e.g. upon binding to a specific target[35, 36], but the formation of high-affinity complexes has been reported even for fully unfolded IDPs[37].

IDPs are involved in multiple diseases such as cancer, amyloidosis, and neurodegenerative diseases[38, 40–43]. To shed light on the pathological behavior of IDPs and the relationship between their sequence, function, and characteristic features such as the tendency for liquid-liquid phase separation[44–47], we need to understand how conformational ensembles accessible to intrinsically disordered proteins in solution depend on interactions with their molecular environment. In principle, structural and dynamical properties of conformational ensembles explored by IDPs can be characterized in small angle X-ray scattering (SAXS), nuclear magnetic resonance (NMR), and Förster resonance

energy transfer (FRET) experiments[48–50], while atomistic molecular dynamics (MD) simulations provide a valuable tool to obtain the underlying microscopic picture[51–53]. However, MD simulations with standard empirical force field models typically fail to reproduce experimental observations and tend to predict overly compact conformations of solvated IDPs[51, 52, 54]. This led to the development of IDP-specific protein force field parameters, which re-weight dihedral angles to modify the conformational ensembles of IDPs in simulations[55]. However, the shortcomings of simulations with standard force fields to describe IDPs are more commonly ascribed to an imbalance between non-covalent interactions within the protein, which stabilize compact states, and protein-solvent interactions, which stabilize extended conformations with a higher solvent-accessible surface area[51, 52, 54].

This motivated multiple strategies to modify standard force fields to improve their ability to reproduce experimental results[53]. In one approach that combines protein force fields in the AMBER family[185–187] with the TIP4P/2005 water model[188], Lennard-Jones (LJ) interaction energies between water oxygens and protein atoms have been increased by 10% [52, 189, 190]. Using a similar strategy, scaled LJ interactions between water hydrogens and protein atoms were proposed for the CHARMM36m force field[191] in combination with the TIPS3P water model[66] (a modified version of the TIP3P[61] water model used in CHARMM that includes water hydrogen LJ sites). In a different approach, Shaw and co-workers re-parametrized the TIP4P water model to improve dispersion interactions between the protein and the solvent, while also modifying electrostatics and parameters in AMBER-type protein force fields[51, 193–195]. Encouraging results have also been achieved with new types of water models that emphasize the role of electrostatics and charge distributions over molecular geometry[196, 197].

In each of the cases cited above, modified protein-water interactions result in more favorable extended conformations for IDPs in solution and improve the agreement between

atomistic simulations and experimental data. However, the thermodynamic consequences of modified protein-water interactions for the solvation of proteins have so far only been analyzed for side-chain analogs[52, 53]. A more detailed analysis of protein solvation free energies including a decomposition of local enthalpic and entropic contributions has so far been missing. The latter is key to understand the microscopic consequences of modified protein-water interactions that remain hidden by comparisons to structural parameters obtained in experiments.

In this study, we simulated the intrinsically disordered K-18 domain of the Tau protein[48, 198] with varying combinations of standard and modified protein and water force fields. In addition to an analysis of the conformational ensembles obtained with each simulation model, we analyzed the protein solvation free energy and its enthalpic and entropic contributions for a selected conformational state with spatial resolution using the previously developed 3D-2PT (3D-two-phase thermodynamics) approach[167, 199]. Our analysis reveals that distinct strategies for the modification of protein-water interactions have distinguishable thermodynamic consequences. Further, protein solvation is not just affected globally but each modification of force field parameters results in unique changes in the solvation thermodynamics of polar and non-polar groups of the protein that affect the conformational ensembles sampled in simulations.

5.3 Methods

5.3.1 Conformational Sampling

All simulations were performed using the GROMACS 2018.1[228] software package. We chose the K-18 domain of the Tau protein as an example to study the impact of varying force field parameterizations on the solvation of IDPs and their thermally accessible conformational space[48, 198]. We used four pairs of force field parameters in our simulations,

each containing one set of standard parameters and one set with modified protein-water interactions intended to stabilize extended conformations and improve agreement with experiments. We provide a list of the pairs of force field parameter sets and modifications applied to protein-water interactions in Table 5.1, where we also introduce shorthand notations used throughout this work. The A03w and A99w protein force fields are modified versions of AMBER03 and AMBER99SB for use in combination with the highly optimized TIP4P/2005 water model[188, 189]. In A03ws and A99ws, the ϵ parameter that describes the strength of Lennard-Jones interactions has been scaled by a factor of 1.10 for interactions between protein atoms and water oxygens.[52, 189] In A99*D, the four-site water model has been re-parameterized with increased dispersion interactions (affecting protein-water and water-water interactions simultaneously) and a larger water dipole moment that modifies electrostatic interactions and screening.[51] The resulting TIP4P-D water model shares its geometry with the TIP4P/2005 water model that we use here as a reference in the A99* system without modifying the protein force field. To ensure that our results do not depend on this combination of protein and water force fields, we performed an alternative version of our analysis with the TIP4P-Ew water model [229] as a reference for comparisons to TIP4P-D (A99*Ew). Lastly, in the C36m* model, the ϵ parameter for Lennard-Jones interactions between protein atoms and water hydrogens has been increased by a factor of ~2.17.[191] The existence of such interactions is a specific feature of the TIPS3P water model, which has been developed in conjunction with the CHARMM force field family for proteins.[66]

To generate an initial atomistic structure model for MD simulations of the protein, we selected the first backbone conformation for the K-18 domain of Tau in the PED00017 entry of the Protein Ensemble Database (PED)[231] (6AAC in a previous version of the PED[48]). We then used the SCWRL4 package[232] to add the missing side chains, which resulted in 1986 protein atoms for the 130 amino acid residues. We set the protonation state of each

Table 5.1: Force field parameter sets used in MD simulations of the K-18 domain of the Tau protein

System Name	Protein FF	Water FF	Modification
A03w	AMBER03w[a]	TIP4P/2005[b]	none
A03ws	AMBER03ws[c]	TIP4P/2005[b]	1.10 scaling of LJ-ϵ_{O_w,X_p}
A99w	AMBER99SBw[c]	TIP4P/2005[b]	none
A99ws	AMBER99SBws[c]	TIP4P/2005[b]	1.10 scaling of LJ-ϵ_{O_w,X_p}
A99*	AMBER99SB*-ILDN[d]	TIP4P/2005[b]	none
A99*Ew	AMBER99SB*-ILDN[d]	TIP4P-Ew[e]	none
A99*D	AMBER99SB*-ILDN[d]	TIP4P-D[f]	re-parameterized water model
C36m	CHARMM36m[g]	TIPS3P[h]	none
C36m*	CHARMM36m[g]	TIPS3P[h]	2.17 scaling of LJ-ϵ_{H_w,X_p}

[a]Ref. 189; [b]Ref. 188; [c]Ref. 52; [d]Ref. 230; [e]Ref. 229; [f]Ref. 51; [g]Ref. 191; [h]Ref. 66.

amino acid residue according to a pH of 7 based on standard side chain pKa's in solution. We then solvated the protein in a cubic box with a 164 Å side length by adding approximately 145,000 water molecules and 828 ions (409 of Na^+ and 419 of Cl^- ions), neutralizing the protein charge and resulting in a physiological salt concentration of 150 mM. Salt can influence the conformational ensemble and phase separation behavior of IDPs via electrostatic screening and salting-out effects as discussed in recent work by Zheng *et al*. [233]

We applied periodic boundary conditions (PBC) in all dimensions and used the particle mesh Ewald algorithm[126] to describe long-ranged electrostatics using a 1.2 Å real space grid. We used, a 10 Å real space cut-off for short-ranged electrostatic and Lennard-Jones interactions, constrained all bond lengths in the protein with the LINCS[234] algorithm, and constrained the geometry of all water molecules in the system with SETTLE[63]. Simulations with all force fields were preceded by 1500 steps of energy minimization with the steepest descent algorithm.

We then equilibrated each system in the isothermal-isobaric ensemble (NPT) at 300 K and 1 bar for 1 ns using a time step of 1 fs, a velocity rescaling thermostat[235] with a 0.1 ps time constant and a Berendsen weak-coupling barostat[129] with a time constant of 2.0 ps. For each system, we then performed 1-2 microsecond simulations in the NPT ensemble at 300 K and 1 bar using a time step of 2 fs and a Nose-Hoover thermostat[236] and a Parrinello-Rahman barostat[133] with time constants of 2.0 ps to sample the conformational space. We monitored the conformational dynamics of the simulated proteins via the root mean squared displacements (RMSDs) from the initial structure, the radius of gyration (R_G), and the solvent accessible surface area (SASA) and considered the first 200 ns of each trajectory as additional equilibration time.

5.3.2 Solvation Thermodynamics

For an in-depth analysis of the solvation thermodynamics with 3D-2PT[167, 199, 237], we selected a moderately compact structure (R_G=24.6 Å) with an extended SASA of 14,100 $Å^2$ from the A03ws (see Table 5.1) simulation. This specific structure was chosen due to the simultaneous presence of fully extended and partially collapsed portions of the protein (see Results).

For the 3D-2PT analysis, we followed procedures in previous work[237–241] and performed an additional set of simulations during which the protein conformation was kept fixed while hydration water configurations were sampled. For this purpose, position restraints with a force constant of 5000 kJ mol^{-1} $Å^{-2}$ were applied to all protein atoms during all the following steps. First, we re-solvated the protein in its selected conformation with 30,000 water molecules in a 100 Å × 100 Å × 100 Å cubic simulation box and generated topologies for all force field models described in Table 5.1. After performing a steepest descent energy minimization, we equilibrated each system for 1 ns in the NPT ensemble using a Berendsen weak-coupling thermostat and barostat[129] with time constants of 1 ps,

a reference temperature of 300 K, and a reference pressure of 1 bar. The applied position restraints maintained the RMSD for all protein atoms below 0.1 Å. Following the equilibration, we performed for each system 10 ns simulations in the NVT ensemble at 300 K using a Nose-Hoover thermostat with a time constant of 1 ps[130]. From this simulation, we extracted 100 independent solvent micro-states (coordinates and velocities) at regular 100 ps time intervals. These microstates were then used to initialize NVE simulations of 100 ps length, which were used for the 3D-2PT analysis. In these simulations, the integration time step was set to 1 fs and coordinates and velocities were stored every 8 fs.

The 3D-2PT analysis, which is described in detail in Ref. 167 and in the SI, was carried out on a 98x98x98 analysis grid with a grid constant of 1 Å. The grid was used to compute local contributions to the solvation enthalpy, ΔH_{solv}, and solvation entropy, ΔS_{solv}, which are combined to describe contributions to the solvation free energy, ΔG_{solv}.

$$\Delta G_{\mathrm{solv}} = \Delta H_{\mathrm{solv}} - T\Delta S_{\mathrm{solv}} \tag{5.1}$$

Ignoring negligible contributions from volume work, ΔH_{solv} is computed as a sum of pairwise solute-solvent (uv) interactions, ΔU_{uv}, and changes in solvent-solvent (vv) interactions, ΔH_{vv}, caused by the presence of the solute. A similar separation is performed for ΔS_{solv} based on Ben-Naim's proof that solute-induced changes of the enthalpy and entropy cancel out exactly.[1, 99]

$$\Delta H_{\mathrm{vv}} - T\Delta S_{\mathrm{vv}} = 0 \tag{5.2}$$

Hence, ΔG_{solv} can alternatively be described as the sum of enthalpy and entropy changes caused solely by solute-water interactions.

$$\Delta G_{\mathrm{solv}} = \Delta H_{\mathrm{uv}} - T\Delta S_{\mathrm{uv}} \tag{5.3}$$

As described in the SI, contributions to each term are computed per voxel of the analysis grid, either as an average local contribution per water molecule or as a spatial density. The

latter can be summed over all voxels of the analysis grid or a fraction of it, e.g., to obtain the total solvation free energy or contributions from distinct components of the hydration shell, respectively.

5.4 Results and Discussion

5.4.1 Comparison of Conformational Ensembles

We begin our analysis of the impact of force field modifications on simulations of IDPs in solution with a comparison of conformational ensembles. For this purpose, we performed microsecond timescale simulations of the K-18 domain of the Tau protein.

As a reference, we used simulations with force field parameters from the AMBER and CHARMM family that have been optimized for simulations of folded proteins (A03w, A99w, A99*/A99*-Ew and C36m in Table 5.1)[189, 191, 230]. We then compared results from these reference simulations to simulations with modified force fields optimized to improve the description of IDPs in solution either via explicit modification of protein-water interactions or a re-parameterization of the water model (A03ws, A99ws, A99*D, and C36m*, see Table 5.1)[51, 52, 189, 191].

In Figure 5.1, we plot the time evolution of the radius of gyration R_G of the protein for all sets of simulations. The results show that the K-18 domain quickly collapses into a compact conformation in simulations with standard force field parameters (A03w, A99w, A99* and C36m) resulting in an average R_G between 20 and 22 Å with minor fluctuations. The sampling of these systems was stopped after 1 microsecond due to the lack of conformational changes after the collapse. These observations are in agreement with previous simulation studies of other protein systems using standard force fields but are incompatible with experimental observations. [52, 54, 191]

Experimentally, the average radius of gyration of the K-18 domain was determined as

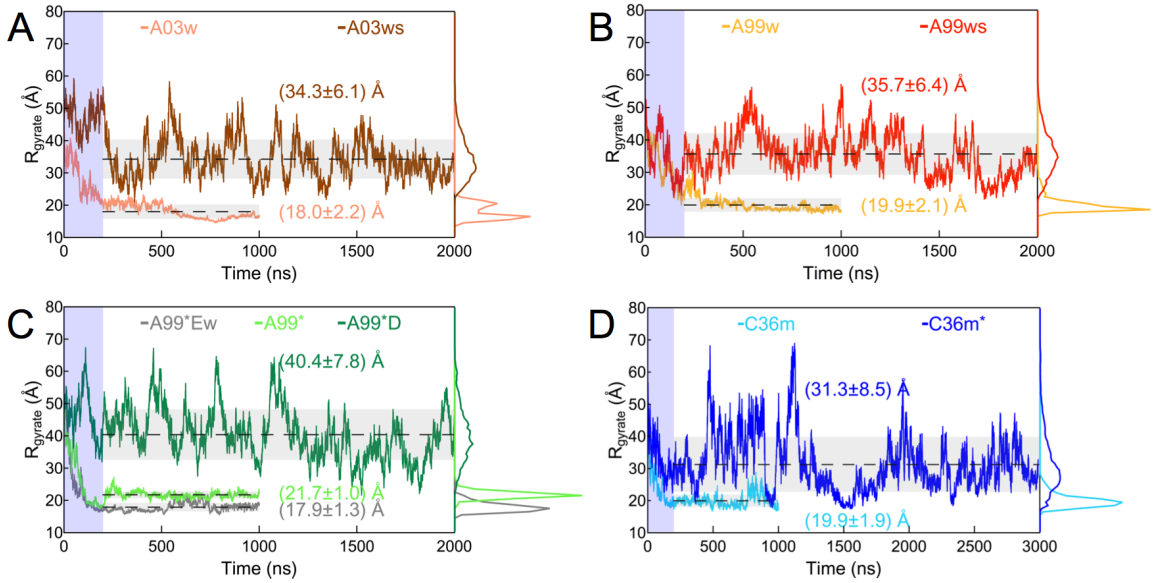

Figure 5.1: Time traces of the protein R_G in MD simulations of the K-18 domain of the Tau protein with distinct sets of force field parameters (see Table 5.1). An equilibration time of 200 ns is highlighted and excluded from further analysis. Dashed horizontal lines indicate averages and shaded gray backgrounds represent standard deviations due to conformational fluctuations. The numerical values are given as insets and histograms of R_G are indicated on the alternative y-axis.

(33.7±7.6) Å by Blackridge and co-workers using chemical shifts and residual dipolar couplings [48]. All modified force fields optimized for simulations of IDPs tested in this study generate conformational ensembles that are in much better agreement with this experimental result. We extended simulations with these force fields up to 2-3 microseconds to sample conformational fluctuations. Especially simulations with scaled Lennard-Jones interactions between protein and water atoms (A03ws, A99ws, and C36m*) accurately reproduce the experimental average, while simulations with the TIP4P-D water model (A99*D) slightly overshoot and predict an average radius of gyration that is too large.

A complementary comparison of the conformational ensembles generated in our simulations to single-molecule FRET data [242] is provided in Figure S1 of the SI. The experiments report an average C_α-C_α distance between the two cysteine residues of the K-18 domain of 38 Å. This distance is significantly smaller in simulations with standard force

fields (20-24 Å), which matches our observations for R_G. For simulations with modified force fields, A03ws shows close agreement with the experiment, while we observed larger average distances of 45-50 Å for A99ws, A99*D and C36m*, albeit with large fluctuations.

Both, the NMR and single-molecule FRET experiments were performed in 50 mM sodium phosphate buffer at pH 6.8, [48, 242, 243] which corresponds to an ionic strength of 78 mM (Debye length 10.9 Å). Compared to the salt concentration in our simulations (150 mM, Debye length 7.8 Å), we expect decreased electrostatic screening of repulsive interactions between the predominantly positive charges of the K-18 domain (more extended conformations) in these experiments. Compared to chloride, salting-out effects for anionic components in the experimental buffer are stronger,[244] but their concentration is lower. [233] In summary, one may expect that conformations of the K-18 domain in experiments are slightly more extended than in our simulation. This is notable, since our results in Figures 5.1 and S1 indicate that the average R_G and C_α-C_α distances observed in our simulations with modified force fields mostly exceed the experimental reference values. A03ws predicts average R_G and C_α-C_α distances that remain closest to the experiments. C36m* is an exception as it predicts an average R_G that is lower than observed experimentally but significantly overestimates the experimentally measured C_α-C_α distance.

In Figure S2 of the SI, we show a complementary analysis of the solvent-accessible surface area (SASA) versus simulation time, which reproduces the general trend observed in Figures 5.1 and S1: Simulations with standard force field parameters result in collapsed conformations, while extended conformations are observed in simulations with the modified force fields.

Joint probability distributions of R_G and SASA shown in Figures S3 and S4 of the SI further illustrate differences between the conformational ensembles observed in simulations with standard and modified force field parameters. More importantly, the joint probability distributions highlight quantitative differences between the simulations with distinct sets

of modified force fields. While all simulations with modified force field parameters result in improved agreement with experimental observables, the conformational ensembles generated in each of these simulations remain noticeably distinct from each other.

5.4.2 Solvation Thermodynamics

To provide additional insights on how the distinct force field modifications promote the formation of more extended and solvent-exposed protein conformations, we analyzed in the following the solvation thermodynamics of the K-18 domain using 3D-2PT [199]. To ensure comparability between the simulations, we perform these studies for a fixed conformation of the protein and sample only the protein-water and water-water interactions. The protein conformation was selected from the A03ws simulation and includes a mixture of compact and extended features. The spatially resolved analysis of the solvation thermodynamics for a single structure allows us to separate contributions to the solvation free energy, ΔG_{solv}, from the hydration shells of polar and nonpolar functional groups of the K-18 domain as well as longer-ranged contributions from water molecules with a minimum distance of 5 Å from the closest protein atom. While ΔG_{solv} will depend strongly on the protein conformation, our analysis aims to highlight overall differences in the protein solvation thermodynamics for distinct force fields that are not expected to depend strongly on the specific protein conformation.

5.4.2.1 IDP Solvation with Standard Force Field Parameters

The 3D-2PT approach spatially resolves local contributions to the solvation free energy, enthalpy, and entropy on a three-dimensional analysis grid. Each property can be expressed either per water molecule at a given site or as an energy density (per volume). The results are shown together with a 3D representation of the selected conformation of the K-18 domain in Figure 5.2. We visualize the selected conformation in panel A and highlight the

location of charged, polar, and nonpolar side chains. In panel B, we illustrate a 3D map of local contributions to the solvation free energy. For this purpose, we used a color code to project spatially resolved solvation free energy contributions per water molecule on an iso-density surface of the water number density at 1.25 g/cm^3. This iso-density surface encloses regions of increased water density characteristic for the first hydration shell of the protein where contributions to the solvation free energy are primarily located. In panel C, we show the closely related solvation free energy density projected on two surfaces with a constant distance to the closest protein atom, i.e., 3.0 Å and 3.5 Å, respectively (see caption). Both representations of the solvation free energy maps provide complementary information. Blue colors indicate sites with small or even positive (unfavorable) contributions to the protein solvation free energy, while green and red colors indicate local sites with significant favorable contributions to ΔG_{solv}. Differences between results obtained from simulations with distinct force fields are not easily discernible at this scale, which is why we only show the result for A03w here.

Quantitative differences between distinct force fields become apparent upon integration over the entire analysis grid, the hydration shell (voxels within 5 Å of the closest non-hydrogen protein atom), and its polar and nonpolar components. Here, we distinguish polar and nonpolar components of the hydration shell based on the partial charge (in the AMBER03 force field) of the closest non-hydrogen atom of the protein. If a voxel of the analysis grid lies within the protein hydration shell and the closest non-hydrogen atom of the protein carries a partial charge q with $|q| > 0.2$, we define it as part of the polar hydration shell. All other voxels of the protein hydration shell are defined as the nonpolar hydration shell.

We first analyzed distinct contributions to the solvation free energy of the K-18 domain in simulations with standard force fields described in Table 5.1. In panel A of Figure 5.3, we compare the integrated solvation free energies of the K-18 domain in simulations with the

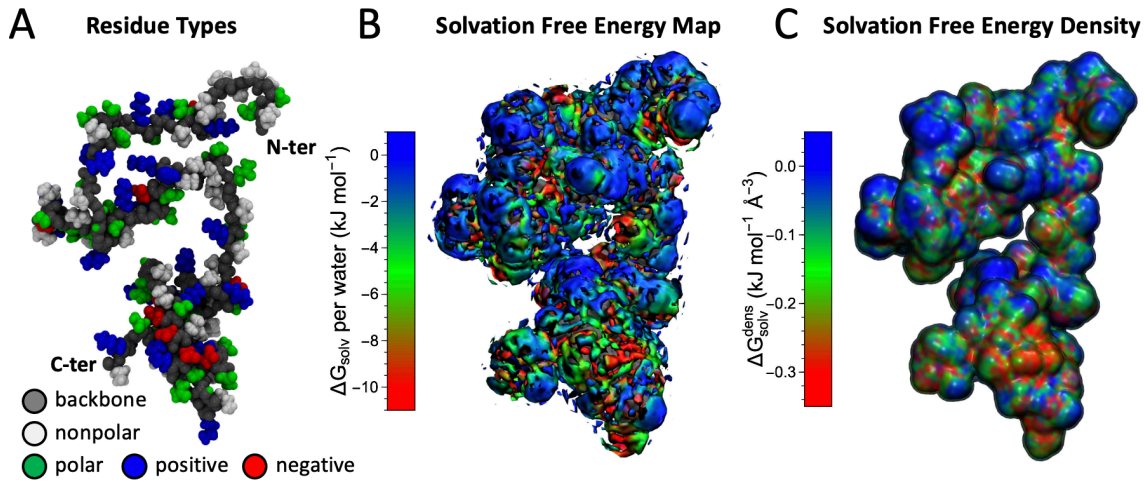

Figure 5.2: Solvation of the K-18 domain in a fixed conformation. (A) The selected conformation with R_G =24.6 Å and SASA=13,260 Å^2 was sampled as a comparably compact state in the A03w simulation, which features extended and collapsed domains. All atoms in the protein are shown as van-der-Waals spheres colored by the corresponding residue type as indicated. (B) Exemplary solvation free energy map obtained with the 3D-2PT formalism using the A03w force field. Local contributions to ΔG_{solv} per water molecule (Eq. S10 in the SI) are projected on an iso-surface of the water density at 1.25 g/cm^3, which effectively highlights the first hydration shell of the protein. (C) Solvation free energy density obtained from the data set in (B) (Eq. S11 in the SI) color-coded on two surfaces with a constant distance of 3.0 Å and 3.5 Å from the closest atom of the protein. The outer surface is semi-transparent to allow an average view on both surfaces.

distinct standard force fields (prior to the introduction of any modifications). The total solvation free energy obtained upon integrating over the entire analysis grid is most favorable for the C36m system and least favorable for A03w. The solvation free energy is a critical component of the free energy surface sampled by the protein in unconstrained simulations. However, the total free energy surface of the protein also includes the intramolecular protein potential energy and its conformational entropy [240], which are excluded from our present analysis.

Figure 5.3A shows that a dominant fraction of the total solvation free energy stems from contributions in the 5 Å hydration shell (~90%). Within the hydration shell, solvation free energy contributions are dominated by contributions from the polar hydration shell in all systems. The latter is partially caused by the respective numbers of water molecules in the

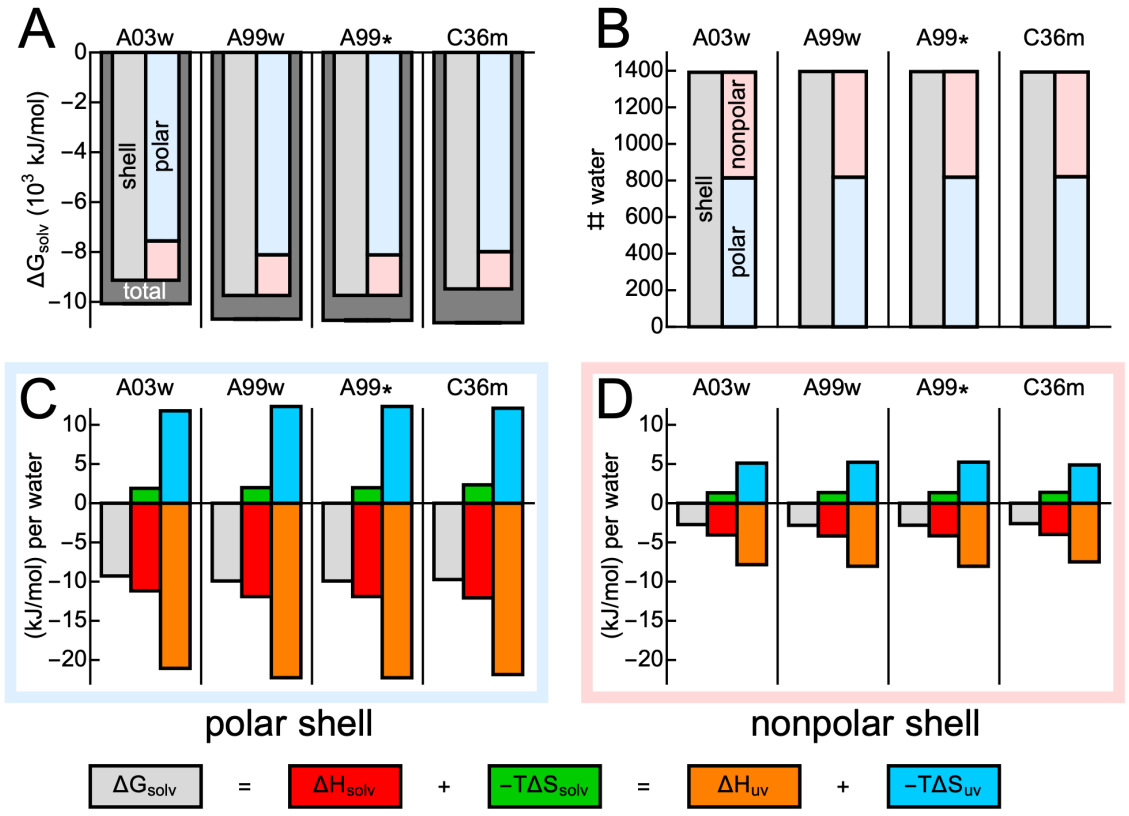

Figure 5.3: Solvation thermodynamics obtained with 3D-2PT for the fixed conformation of the K-18 domain in simulations with standard force field models (see Table 5.1). (A) Total solvation free energy (gray), solvation free energy contributions of a 5 Å shell around the protein defined by the distance to closest non-hydrogen protein atom (light gray), and decomposition into contributions of hydration shells of polar (light blue) and nonpolar (light red) protein atoms. (B) Number of water molecules in the total 5 Å hydration shell and its polar and nonpolar components (colors as in A). (C, D) Solvation free energy contributions per water molecule in the polar (C) and nonpolar (D) components of the hydration shell. The latter (light gray) is decomposed into enthalpic and entropic contributions including (ΔH_{solv} in red and $-T\Delta S_{\mathrm{solv}}$ in green) and excluding (ΔH_{uv} in orange and $-T\Delta S_{\mathrm{uv}}$ in cyan) canceling contributions from solute-induced changes of water-water interactions (see Methods and SI). Error bars describing the error of the mean are included in all panels, but not visible on the shown scale.

polar and nonpolar hydration shells shown in panel B of Figure 5.3. The 5 Å hydration shell consists of just below 1,400 water molecules in all systems. Approximately 60% of these water molecules are located in the polar and ~40% in the nonpolar hydration shell. However, more important than these different populations are the distinct thermodynamic properties of water in polar and nonpolar hydration shells, which are quantified and decomposed into enthalpic and entropic components (per water molecule) in panels C and D of Figure 5.3.

Independent of the force field model, water molecules in the polar hydration shell contribute ~-9 kJ/mol to the solvation free energy, while water molecules in the nonpolar hydration shell only contribute ~-2 kJ/mol. This difference is not surprising given the hydrophilic and hydrophobic properties of polar and nonpolar functional groups of the protein. Decomposing local solvation free energy contributions into enthalpic and entropic contributions allows us to analyze the distinct protein-water interactions in more detail.

If protein-induced changes in the water-water interactions are included, the solvation entropy ΔS_{solv} appears only as a minor factor. The positive $-T\Delta S_{\mathrm{solv}}$ term compensates 18-21% of the corresponding negative ΔH_{solv} term in the polar hydration shell, which seemingly dominates the favorable solvation free energy contribution of ~-9 kJ/mol per water molecule. In the nonpolar hydration shell, both terms are smaller and the positive $-T\Delta S_{\mathrm{solv}}$ term compensates 35-36% of the negative ΔH_{solv} term, which leads to the weakly favorable ~-2 kJ/mol contribution to ΔG_{solv} per water molecule.

As discussed in the context of Eqs. 5.2 and 5.3, ΔH_{solv} and ΔS_{solv} include the exactly canceling protein-induced changes of water-water interactions, ΔH_{vv} and ΔS_{vv}, which have no impact on the free energy. To focus only on non-canceling contributions caused directly by solute-water interactions, we thus also computed ΔH_{uv} and ΔS_{uv}. This reveals the actual thermodynamic driving forces that govern solvation and allows for direct comparisons to analytical theory. Specifically, linear response theory for the solvation of ions, dipoles, etc.

in a polar solvent, i.e., solute-solvent interactions dominated by electrostatics [245, 246], predicts the following ratio [245, 246]:

$$\gamma = \Delta G_{\text{solv}}/\Delta H_{\text{uv}} = 1/2 \tag{5.4}$$

For both the polar and nonpolar hydration shells ΔH_{uv} and $-T\Delta S_{\text{uv}}$ are substantially larger than ΔH_{solv} and $-T\Delta S_{\text{solv}}$ terms. Consequently, the degree of cancellation between ΔH_{uv} and $-T\Delta S_{\text{uv}}$ is also increased. In the polar hydration shell, 43% of the negative ΔH_{uv} term is compensated by positive $-T\Delta S_{\text{uv}}$ contributions. In other words, the ratio γ evaluates to 0.43, which is just below the linear response prediction for electrostatically dominated solvation. In the nonpolar hydration shell, this ratio drops to just ~0.34 with a weak dependence on the force field used.

We previously introduced deviations of γ from the linear response prediction as a measure of the relative importance of electrostatic solute-solvent interactions with $\gamma = 0.5$ as an upper limit [240]. We then used an empirical expression for γ to predict solvation free energies of a small peptide as a function of its conformation based on ensemble averages of ΔH_{uv} [240]. In the present study, the observed decrease of γ from 0.43 in the polar to 0.34 in the nonpolar hydration shell follows the expected trend, i.e., the solvation of polar functional groups is dominated by electrostatics while the solvation of nonpolar functional groups is not.

5.4.2.2 Changes in IDP Solvation with Modified Force Fields

In Figure 5.4, we analyzed changes in the solvation of the K-18 domain induced by modified force field parameters, i.e., we analyzed the difference between solvation free energy contributions obtained in simulations with modified and standard force field parameters. An alternative version of this analysis is shown in Figure S5 of the SI, where the A99*Ew system was used for comparison with A99*D. As shown in panel A, the solvation

free energy changes, $\Delta\Delta G_{solv}$, are favorable for all four force field pairs, i.e., the modified force fields result in an overall more negative ΔG_{solv} of the protein. The latter is expected given the more expanded protein conformations observed in simulations with the modified force fields in Figures 5.1 and S1-S4. However, the magnitude, localization, and decomposition of $\Delta\Delta G_{solv}$ differs between distinct pairs of standard and modified force fields. In Figure 5.4A, the magnitude of the total $\Delta\Delta G_{solv}$ is significantly larger for the comparisons of A03ws and A99ws with their reference systems than for C36m*. For A99*D, the magnitude of the total $\Delta\Delta G_{solv}$ depends on the choice of the reference system. If A99* (with the TIP4P/2005 water model) is used as a reference, the total $\Delta\Delta G_{solv}$ is comparable to C36m*. However, if we use A99*Ew (with the TIP4P-Ew water model) as a reference, the total $\Delta\Delta G_{solv}$ is comparable to A03ws and A99ws. The latter is interesting as it does not directly correlate with changes of the sampled conformational ensembles sampled with the modified force fields analyzed in Figures 5.1 and S1-S4. The lack of correlation can be attributed to different intra-protein and protein-water interactions in the corresponding reference systems.

It is further important to note that different effects of force field modifications on the solvation of polar and nonpolar groups discussed below imply that the total ΔG_{solv} and $\Delta\Delta G_{solv}$ for the compared pairs of modified and standard force fields is expected to depend on the protein conformation selected for the analysis, i.e., the relative exposure of polar and nonpolar functional groups. Hence, we focus our following analysis on changes of solvation parameters that we expect to be less dependent on the protein conformation, i.e., qualitative contributions to $\Delta\Delta G_{solv}$ from the first and higher hydration shells and per water molecule contributions in hydration shells of polar and nonpolar functional groups.

In the systems in which the protein-water interactions were modified directly, i.e., A03ws, A99ws and C36m*, favorable contributions to $\Delta\Delta G_{solv}$ are found in the protein hydration shell as well as to a lesser extent in higher hydration shells for distances larger than 5 Å

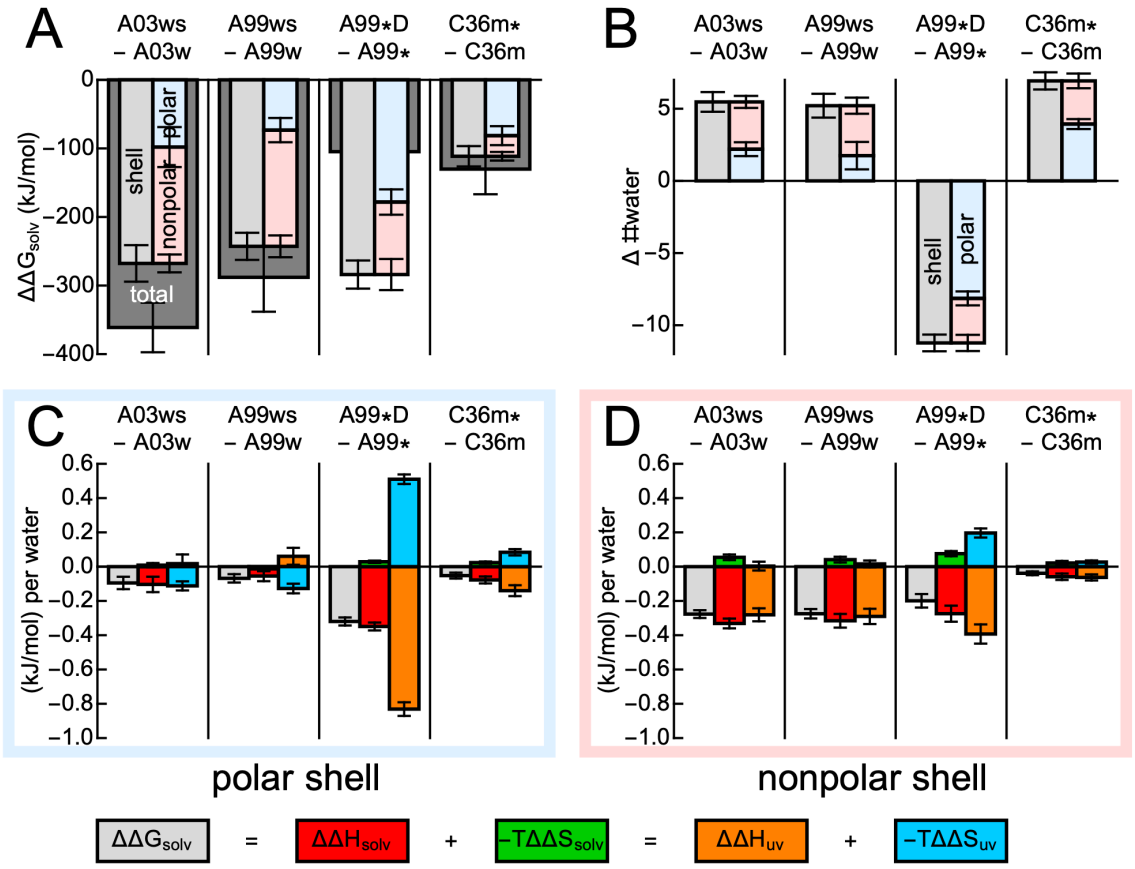

Figure 5.4: Change in solvation due to modified force field parameters. (A) Change in the total solvation free energy (gray), solvation free energy contributions of a 5 Å shell around the protein defined by the distance to closest non-hydrogen protein atom (light gray), and decomposition into contributions of hydration shells of polar (light blue) and nonpolar (light red) protein atoms. (B) Change in the number of water molecules in the total 5 Å hydration shell and its polar and nonpolar components. (C, D) Change of solvation free energy contributions per water molecule in the polar (C) and nonpolar (D) components of the hydration shell. The latter is decomposed into enthalpic and entropic contributions including (ΔH_{solv} in red and $-T\Delta S_{\mathrm{solv}}$ in green) and excluding (ΔH_{uv} in orange and $-T\Delta S_{\mathrm{uv}}$ in cyan) canceling contributions from solute-induced changes of water-water interactions (see Methods and SI). Error bars indicate the statistical error of the mean.

from the protein. The latter results in the increasing magnitude of $\Delta\Delta G_{solv}$, when the integration is performed over the full analysis grid instead of just the hydration shell. Notably, this situation is significantly different for A99*-D with the re-parameterized TIP4P-D water model (independent of the choice of the reference system, see Figure S5 in the SI). Here, favorable $\Delta\Delta G_{solv}$ contributions in the hydration shell are partially compensated by unfavorable contributions at larger distances from the protein. We can rationalize the latter by the nature of the force field modifications.

In the A99*D system, the re-parameterized water model features a larger dipole moment, which affects the preferential orientation of water molecules within the electric field of the protein. These changes are largest at short distances and in the polar hydration shell, where local contributions to $\Delta\Delta G_{solv}$ are most pronounced. Water molecules at distances > 5 Å from the protein interact more strongly with water molecules in the first hydration shell (with which they can form direct hydrogen bonds) than with the protein. Consequently, water molecules in higher hydration shells prioritize their interactions with water in the first hydration shell over weaker direct interactions with the protein. Thus altered water orientations in the first hydration shell of the A99*D system propagate into succinct hydration shells, where they result in unfavorable changes of protein-water interactions and thus unfavorable contributions to $\Delta\Delta G_{solv}$.

In addition to $\Delta\Delta G_{solv}$, we analyzed changes in the number of water molecules within the 5 Å hydration shell due to the modified force field parameters in Figure 5.4B. As one might expect, the number of water molecules increased in all cases, where protein-water interactions were modified to be more attractive, i.e., due to changes in protein-water Lennard-Jones interactions in A03ws, A99ws, and C36m* systems. However, for the A99*D system where the re-parameterized water model affects both protein-water and water-water interactions, the number of water molecules in the hydration shell decreases despite the favorable $\Delta\Delta G_{solv}$ reported in panel A. While small compared to the overall number of hydration wa-

ter molecules (see Figure 5.3B), these changes highlight that the impact of simultaneously modified protein-water and water-water interactions are non-trivial to predict. The same applies to changes in the solvation free energy: only changes in water-water interactions induced by the presence of the protein solute cancel out as described in Eq. 5.2, while modified water-water interactions in the force field model do not.

In Figure 5.4C and D, we decomposed $\Delta\Delta G_{\mathrm{solv}}$ contributions from the polar and nonpolar hydration shells into contributions per water molecule. This decomposition reveals additional details that inform how the distinct modifications of the force field parameters impact protein solvation in each case.

For A03ws and A99ws, $\Delta\Delta G_{\mathrm{solv}}$ contributions per water molecule are largest in the nonpolar hydration shell, while the opposite is the case for A99*D and C36m*. We begin the discussion of these differences with the nonpolar hydration shell in the A03ws and A99ws systems, where Lennard-Jones interactions between protein atoms and water oxygens have been scaled.

In the nonpolar hydration shell, protein-water Lennard-Jones interactions do not compete with electrostatic protein-water interactions such as hydrogen bonds. Existing attractive Lennard-Jones interactions simply become stronger upon scaling, while changes in the water structure are minimal. The latter can be deduced by the minor difference between $\Delta\Delta H_{\mathrm{solv}}$ and $\Delta\Delta H_{\mathrm{uv}}$, which describes the change of water-water interactions $\Delta\Delta H_{\mathrm{vv}}$. Further, the decomposition of $\Delta\Delta G_{\mathrm{solv}}$ shows that the latter is essentially identical to $\Delta\Delta H_{\mathrm{uv}}$ and no significant compensation by the $-T\Delta\Delta S_{\mathrm{uv}}$ term is observed. For hydrophobic surfaces in water, ΔS_{uv} is dominated by repulsive protein-water interactions [247], which are insensitive to scaling of the corresponding Lennard-Jones potentials. The steepness of the repulsive term prevents the population of strongly repulsive configurations even prior to scaling.

This situation is different in the polar hydration shell of the A03ws and A99ws sys-

tems. Here, Lennard-Jones interactions between protein atoms and water oxygens, which are isotropic and do not prefer any specific water orientation, compete with directional protein-water hydrogen bonds. Scaling the Lennard-Jones interactions thus increases the weight of Lennard-Jones interactions within this balance and results in overall less directional protein-water interactions. The consequences can be compared to previous studies of bulk water with modified water models, which showed how an increased weight of isotropic Lennard-Jones interactions weakens the structure of the water hydrogen bond network.[248] In the polar hydration shell of the A03ws and A99ws systems, we thus observe a (small) favorable $-T\Delta\Delta S_{\mathrm{uv}}$ term, i.e., an increase of the solute-solvent entropy. The resulting change in the hydration water structure is particularly pronounced in the A99ws system, where it even leads to a net unfavorable change of the protein-water interaction energy, i.e., $\Delta\Delta H_{\mathrm{uv}}$.

Compared to the observation for A03ws and A99ws, changes in solvation are qualitatively different for the A99*D and C36m* systems, where the introduced changes in the force field affect the preferential orientation of hydration water molecules.

In the A99*D system, the increased dipole moment of the water model results in an enhanced strength of electrostatic protein-water interactions such as protein-water hydrogen bonds, which is particularly evident in the $\Delta\Delta H_{\mathrm{uv}}$ term for the polar hydration shell. This effect on electrostatic interactions seems to be more important than the enhanced dispersion interactions in the TIP4P-D water model. In both, the polar and non-polar hydration shell, ~50% of the favorable $\Delta\Delta H_{\mathrm{uv}}$ is compensated by an unfavorable $-T\Delta\Delta S_{\mathrm{uv}}$ term, which is expected for primarily modified electrostatic interactions (following the same linear response approach discussed in the context of Eq. 5.4). Since electrostatic protein-water interactions are significantly more relevant in the polar hydration shell, the resulting contributions to $\Delta\Delta G_{\mathrm{solv}}$ are larger than in the nonpolar hydration shell. The large difference between $\Delta\Delta H_{\mathrm{solv}}$ and $\Delta\Delta H_{\mathrm{uv}}$, especially in the polar hydration shell, indicates a significant change in the hydration shell structure and water-water interactions. The latter is a con-

sequence of the previously discussed modified preferential orientation of hydration water molecules and the simultaneous change of water-water interaction potentials in the water model.

In the C36m* system, the scaling of short-ranged Lennard-Jones interactions between protein atoms and water hydrogens, which are specific to the TIPS3P water model, primarily enhance the strength of protein-water hydrogen bonds. Consequently, the impact on $\Delta\Delta G_{\mathrm{solv}}$ (while small in comparison to the other modified force fields) is primarily localized in the polar hydration shell. Interestingly, the partial cancellation of ~50% of the favorable $\Delta\Delta H_{\mathrm{solv}}$ term by an unfavorable $-T\Delta\Delta S_{\mathrm{uv}}$ term closely resembles the expectations for modified electrostatic interactions despite the actual nature of the force field modification.

5.5 Conclusion

We used all-atom MD simulations of the K-18 domain of the Tau protein to compare the impact of previously proposed force field modifications to improve agreement between conformational ensembles of IDPs in simulations and experiments. We utilized four distinct pairs of standard and modified force fields that implement distinct strategies to increase favorable protein-water interactions and stabilize extended conformational states. We initially analyzed conformational ensembles sampled in simulations on the microsecond timescale, before we performed a detailed analysis of spatially resolved solvation free energy maps.

As expected, all force field modifications that stabilize extended protein conformations result in favorable changes in the protein solvation free energy compared to standard parameters. However, our analysis based on spatially resolved solvation free energy maps of a single protein conformation identifies significant differences that arise from each specific strategy to increase attractive protein-water interactions. These differences are expected to affect the properties of conformational ensembles sampled in MD simulations as well as water-mediated interactions between proteins that influence the ability to form complexes

in solution, to aggregate, or even phase-separate.[249, 250]

For example, simulations with the TIP4P-D water model were found to enhance protein-water interactions in the first hydration layers relative to reference simulations with either TIP4P/2005 or TIP4P-Ew, while changes of protein-water interactions over longer distance were found to be unfavorable. We can associate this behavior with a change in the preferential orientation of water molecules in the first hydration shell in response to the modified protein-water interactions. Such restructuring of inner hydration shells propagates into the solution, where water molecules tend to optimize interactions with their immediate solvent environment and interactions with more distant protein atoms only play a secondary role.

Quantitative definitions of solvent-mediated interactions, e.g., between two proteins in solution, include the cost of depopulating each other's hydration shells upon complex formation. As a consequence, simulations with the TIP4P-D water model facilitate the approach of other proteins up to distances of up to 5 Å compared to simulations with the reference water models. Only the final desolvation of the protein-water interface, required to form direct protein-protein contacts, is expected to involve more work due to the increased thermodynamic cost of removing water molecules in the hydration shell compared to simulations with a standard force field.

In addition, the distinct force field modifications exhibit distinct effects on the solvation of polar and nonpolar functional groups of the protein. A03ws and A99ws primarily enhance the solvation of nonpolar functional groups, while A99*D and C36m* primarily enhance the solvation of polar functional groups.

Distinct preferences to enhance the solvation of polar and nonpolar groups will have direct consequences on changes of conformational ensembles sampled with standard and modified force field parameters. Simulations with A03ws and A99ws will tend to stabilize extended and solvent-exposed states of nonpolar residues, while simulations with A99*D and C36m* will tend to stabilize extended states and solvent-exposed states of polar residues

(relative to the corresponding standard force field).

For IDPs whose sequence includes both polar and nonpolar residues, such differences are not likely to be evident in comparisons of global parameters to characterize protein conformational states such as the radius of gyration or end-to-end distance. However, such differences will become more dominant if a protein primarily consists of polar or nonpolar amino acids.

Chapter 6

ELECTRIC-FIELD INDUCED ENTROPIC EFFECTS IN LIQUID WATER

6.1 My Contribution

This project was aimed at investigating the electric-field induced effects on the total and local entropy of bulk water from both quantum and classical points of view. Externally applied electric fields in liquid water can result in a myriad of effects that are important to the fields of electrochemistry and hydrogen-based technologies. While there has been some research conducted to comprehend the thermodynamics of electric fields in aqueous systems, as per literature history no previous reports existed regarding the impact of electric fields on the overall and local entropy of bulk water. We used MD-based 3D-2PT analysis to examine the protein-water interactions by spatially resolving the local entropy and number density of water under different strengths of electric-fields. A relation between entropic and structural modifications with atomistic resolution was established through the spatial maps of local order. In this project, my contributions were focused on carrying out the simulation and analysis tasks, which involved classical molecular dynamics (MD) simulations utilizing the TIP4P/2005 water model in Gromacs 2018.1, as well as conducting the 3D-two-phase thermodynamics (3D-2PT) analysis. Our study of local entropy indicators involved investigating the behavior of bulk water under different conditions at varying electric field intensities. Specifically, we analyzed two scenarios: (i) the ideal system, in

which a water molecule was precisely aligned with the electric field (known as the ”fixed” system), and (ii) the more realistic system, where a water molecule was permitted to rotate freely around a fixed point in the bulk liquid (referred to as the ”pinned” system). The ideal system allowed us to gain a deeper understanding of the water environment along specific orientations of the perfectly aligned molecule, including aspects such as bond vectors and dipole moments. Conversely, the more realistic system was crucial for comprehending how the central water molecule reacted to the external electric field, considering both the effects of field alignment and local structural nuances. Through this detailed investigation, we were able to uncover valuable insights into the intricate behavior of water molecules in the presence of electric fields.

In our overall observations, when no external field was applied, a comparative analysis between the TIP4P/2005 force field and various ab initio functionals demonstrated that the TIP4P/2005 water model, known for its computational efficiency, effectively captured the essential structural characteristics of water. The structural representation provided by this model was found to be in close agreement with that obtained from ab initio molecular dynamics (AIMD) simulations utilizing the more sophisticated SCAN+rVV10 functional.

Furthermore, I calculated the molecular entropy of bulk water using the TIP4P/2005 model with the two-phase thermodynamics (2PT) formalism, and the results were in line with those from AIMD simulations employing the SCAN+rVV10 functional. Nonetheless, both these values were about 15 J mol^{-1} K^{-1} lower than the experimental benchmark value of 69.95 J mol^{-1} K^{-1}. This variation might stem from an inherent tendency of the 2PT formalism to underestimate water entropy, as it is based on the vibrational density of states. Additionally, the molar entropy derived from the TIP4P/2005 water model exhibited considerably less sensitivity to changes induced by polarization compared to that from the BLYP+D3 water model. This difference can be attributed to the inclusion of electronic polarization in the BLYP+D3 model, which influences its reactivity to environmental mod-

ifications.

In additional findings, I utilized molecular dynamics-based analysis through the three-dimensional two-phase thermodynamics (3D-2PT) approach, which provided comprehensive spatial insights into the local entropy and number density of water molecules surrounding a reference molecule. This method enabled me to create three-dimensional visualizations that depicted the changes in the water structure induced by electric fields. Through this approach, I observed an enhancement in the hydrogen bonding between the reference molecule and its nearest neighbors, particularly in specific orientations. This observation is in line with the vibrational Stark effect, indicating that liquid water tends to form more ”ice-like” structures under the influence of electric fields with intensities around 1.0 V nm−1. The application of the 3D-2PT formalism was crucial in demonstrating the intricate relationship between changes in local entropy density and alterations in water number density caused by electric fields. This relationship allowed me to develop a detailed understanding of the impact of electric fields on the entropy of water at the molecular level.

Finally, our main observation was that a static electric field primarily affects the number density of water by inducing a more organized layering effect. This effect is characterized by increased coordination in the second shell of water molecules, along with anisotropic structural changes. Under strong electric fields, water molecules progressively align their dipoles and adopt orientations parallel to the field lines, leading to the formation of hydrogen-bonded structures predominantly oriented along a diagonal orientation relative to the electric field axis.

6.2 Summary

Externally applied electric fields in liquid water can induce a plethora of effects with wide implications in electrochemistry and hydrogen-based technology. Although some effort has been made to elucidate the thermodynamics associated with the application of elec-

tric fields in aqueous systems, to the best of our knowledge, field-induced effects on the total and local entropy of bulk water have never been presented so far. Here, we report on classical TIP4P/2005 and *ab initio* molecular dynamics simulations measuring the entropic contributions carried by diverse field intensities in liquid water at room temperature. We find that strong fields are capable of aligning large fractions of molecular dipoles. Nevertheless, the order-maker action of the field leads to quite modest entropy reductions in classical simulations. Albeit more significant variations are recorded during the first-principles simulations, the associated entropy modifications are small compared to the entropy change involved in the freezing phenomenon, even at intense fields slightly beneath the molecular dissociation threshold. This finding further corroborates the idea that electrofreezing (*i.e.*, the electric-field-induced crystallization) cannot take place in bulk water at room temperature. Besides, here we propose a molecular-dynamics-based analysis (3D-2PT) that spatially resolves the local entropy and the number density of bulk water under an electric field, which enables us to map their field-induced changes in the environment of reference H_2O molecules. By returning detailed spatial maps of the local order, the proposed approach is capable of establishing a link between entropic and structural modifications with atomistic resolution.

6.3 Introduction

In addition to quantum mechanics, matter's behavior is ultimately ruled by the subtle interplay between local electric fields. [251] Hence, it is not surprising that the application of external electric fields can be used to alter the thermodynamics of complex molecular systems in much the same way as temperature and pressure have been traditionally exploited. [252–255] This way, a widespread series of chemical as well as physical phenomena may be triggered by opportunely tuned oriented fields.

Nowadays, many laboratories are able to produce controllable static fields on the or-

der of 1 – 10 V·nm^{-1} at the proximity of high-capability field emitter tips in Scanning Tunneling Microscopy (STM) and Atomic Force Microscopy (AFM) apparatus, [256–258] rendering evident the possibility that tailored static electric fields will routinely be employed as smart catalytic tools in the imminent future. [259] Indeed, it has been proven that electrostatic potential gradients can selectively promote and catalyze Diels-Alder [260] and other industrially-relevant [261] chemical reactions. These findings represent some of the first experimental evidences indicating that electric fields with magnitudes on the order of ~ 1 – 10 V·nm^{-1} can drive and finely control specific chemical transformations. On the other hand, detailed conclusions on the effects produced by the application of intense electric fields were achieved through traditional quantum-based computational approaches on single-molecule systems [262, 263] and – by means of *ab initio* and advanced metadynamics techniques – on condensed phases of matter. [264, 265] Field strengths on the order of ~ 5 V·nm^{-1} have been employed in order to reproduce *in silico* Miller [266] and Miller-like [267] experiments and to highlight some crucial aspects of the role played by electrostatics on general chemical reactivity. [259, 264, 268] Those field intensities couple with the local electric fields naturally present in condensed matter, whose intensity falls exactly in the same range of magnitude. [269] Indeed, electric fields on the order of 10 V·nm^{-1} are generated by the solvent in aqueous systems. [270] Besides, as proven by recent investigations carried out *via* two-dimensional infrared (2D-IR) spectroscopy, field strengths of 3 V·nm^{-1} are ubiquitous in neat liquid water and are responsible for the solvation state of protons. [271] Pioneering Car-Parrinello molecular dynamics simulations [272] and more recent and sophisticated *ab initio* molecular dynamics (AIMD) investigations, [200, 201] also including subtle Nuclear Quantum Effects (NQEs) *via* path-integral AIMD, [273] have partly elucidated the role played by static electric fields when applied on bulk liquid water, shedding some light on its spectroscopic, molecular and proton conductivity response upon exposure to electric fields. In vibrational Stark spectroscopy, frequency shifts induced by

similarly intense electric fields in selected vibrational modes are conveniently exploited as local probes. [274] Furthermore, proton transfer reactions triggered by the strong electric fields spontaneously present at the interface between water and different solid materials have recently been characterized. [275–277] This way, electrostatic potential gradients are capable not only of reorganizing molecular dipoles in such a way to freeze liquids like ammonia, [278] but also of opening pathways in complex reaction networks shaping the chemistry of aqueous systems. [266, 267, 279]

Notwithstanding some thermodynamic investigations reported on the effects generated by electric fields on the solid-liquid equilibrium, [203] electric-field-induced entropy modifications associated with the water hydrogen-bond (H-bond) reorganization have only very recently been investigated at electrified interfaces by means of multiscale simulations. [204] In particular, Estejab *et al.*, by means of classical MD and AIMD simulations, have studied the solvation thermodynamics of a series of surface species on the assumption that local electric fields could be able to reorient the hydration water molecules. [204] Although a key finding – concerning the fact that the entropy of solvation depends on both the strength and direction of the electric field – was reported, [204] this study was affected by the (inextricable) competition between the field and surface effects at the water/electrode interfaces. In addition, many computational efforts were traditionally devoted to the well-known hydrophobic effect, [205] whose molecular-scale origins are modified by local charges (and hence fields). [206, 207] However, to the best of our knowledge, neither qualitative nor quantitative assessments on the role played by electric fields in shaping the local entropy of bulk water have been reported so far.

The aim of the current study is to exploit force-field MD and AIMD simulations of bulk liquid water under the effect of static and homogeneous electric fields to unveil field-induced changes on the structure and entropy of the system. Moreover, here we also track the outcomes stemming from local entropy descriptors capable of returning a spatially-

resolved quantification of the entropy. In particular, we use an MD-based analysis (3D-2PT) [199, 208] that spatially resolves the entropy per water molecule and the local number density of bulk water under an electric field and their changes relative to the zero-field regime. As for these local entropy indicators, we investigate the bulk water environment of two different scenarios at diverse field intensities: (*i*) the ideal system of a water molecule perfectly aligned along the field (named hereinafter as *fixed*) and (*ii*) the more realistic system of a water molecule freely rotating around a locked point in the bulk liquid (named hereinafter as *pinned*). Whilst the ideal system allows to characterize the complex water environment along specific directions of the perfectly aligned molecule (*e.g.*, bond vectors, dipole moment), the more realistic system is needed to describe the response of the central water molecule to the externally imposed electric field, *i.e.*, the combined effect of the field alignment and local structure preferences.

6.4 Methods

6.4.1 Simulation Protocol

We used the software package CP2K, [280] based on the Born-Oppenheimer approach, to perform *ab initio* molecular dynamics (AIMD) simulations of a sample of liquid water under the action of static and homogeneous electric fields applied along a given direction (corresponding to the Z-axis). The implementation of an external field in numerical codes based on Density Functional Theory (DFT) can be achieved by exploiting the Modern Theory of Polarization and Berry's phases [281–283] (see, *e.g.*, Ref. [284]). Under periodic boundary conditions, the problem of computing the polarization $\mathbf{P}$ in an electric field provides the solution to the problem of computing any property of an insulator in a finite homogeneous electric field. By exploiting Berry's phase [283] polarization the overall problem is solved. In particular, by considering an oriented electric field, the following variational

energy functional can be built [284]

$$E^{\mathcal{E}}[\{\psi_i\}] = E^0[\{\psi_i\}] - \mathcal{E} \cdot P[\{\psi_i\}]\,, \tag{6.1}$$

where $E^0[\{\psi_i\}]$ is the well-known energy functional in the zero-field system and $P[\{\psi_i\}]$ is the polarization defined by Resta. [285] This way, a finite static and homogeneous electric field can be applied in AIMD simulations.

The liquid water sample contained 128 H_2O molecules (*i.e.*, 384 atoms) arranged in a cubic cell with 15.82 Å edges, so as to reproduce a density of 0.97 g·cm^{-3}. As usual, structures were replicated in space by employing periodic boundary conditions. The intensity of the electric field was gradually increased with a step increment of 0.5 V·nm^{-1} from zero up to a maximum of 2.0 V·nm^{-1}. Albeit the latter value is slightly beneath the computational [272, 273] and experimental [257, 258, 286] water dissociation threshold, from our prolonged AIMD trajectories a non-zero probability of finding field-induced ionizations at 2.0 V·nm^{-1} was recorded, a result in accordance with recent simulations. [273] Thus, to make a one-to-one comparison with the entropy estimates stemming from the classical force-fields simulations, we decided to consider for the analysis of the entropy only the trajectories where proton transfers do not occur. On the other hand, for structural analyses also data stemming from the simulation at the highest field intensity (2.0 V·nm^{-1}) were post-processed. In the zero-field case we performed dynamics of 50 ps whereas, for each other value of the field intensity, we ran dynamics of almost 250 ps with a time-step of 0.5 fs.

Wavefunctions of the atomic species have been expanded in the TZVP basis set with Goedecker-Teter-Hutter (GTH) pseudopotentials using the GPW method. [287] A plane-wave cutoff of 400 Ry has been imposed. Exchange and correlation (XC) effects were treated with the gradient-corrected Becke-Lee-Yang-Parr (BLYP) [288, 289] density functional. Moreover, in order to take into account dispersion interactions, we employed the

dispersion-corrected version of BLYP (i.e., BLYP+D3(BJ)). [290, 291] A nominal temperature slightly higher than the standard one has been simulated in order to better reproduce the water structure (*i.e.*, $T = 350$ K). The dynamics of ions was simulated classically within a constant number, volume, and temperature (NVT) ensemble, using the Verlet algorithm whereas the canonical sampling has been executed by employing a canonical-sampling-through-velocity-rescaling thermostat (CSVR) [292] set with a time constant equal to 10 fs.

In the zero-field regime, comparisons of some structural properties of water were performed by means of the more recent and accurate non-empirical strongly constrained and appropriately normed (SCAN) functional [293] (plus the rVV10 dispersion correction [294]). In this case, a cubic box containing 97 H_2O molecules (*i.e.*, 291 atoms) arranged in a cubic cell with edges equal to 14.30 Å was simulated for $\sim$ 50 ps. Wavefunctions have been expanded in the TZVP basis set with GTH pseudopotentials specifically optimized for SCAN, whilst a large plane-wave cutoff of 1200 Ry has been imposed after several tests. Moreover, optimized parameters for SCAN were used for mimicking dispersion interactions within the rVV10 method (*i.e.*, $b = 11.40$ and $c = 0.93$). A temperature of 300 K was imposed in the SCAN+rVV10 case *via* a CSVR thermostat [292] set with a time constant equal to 50 fs.

Classical molecular dynamics (MD) simulations of liquid water under static and homogeneous electric fields were carried out with the Gromacs 2018.1 software package. [295] The TIP4P/2005 model [296] was used to describe water-water interactions. Although more sophisticated potentials have been developed in recent decades, the reliability of TIP4P/2005 has been demonstrated by recent interesting results, also concerning the application of external electric fields. [255, 297]

To analyze the bulk entropy of TIP4P/2005 water for comparison with AIMD, we performed one set of simulations with 1024 water molecules in the canonical ensemble with a fixed-volume cubic box with 31.62 Å edges. A second set of simulations for the same system was performed in the isobaric-isothermal ensemble at a constant pressure of 1 bar

to study the impact of field-induced density changes on the water entropy, which were not captured in constant-volume AIMD simulations. To spatially resolve local water entropies in the 3D environment of a central water molecule, we used a third and fourth set of simulations with 2147 water molecules in a cubic box with initial edges of 42 Å, which were equilibrated at a pressure of 1 bar for each applied electric field strength prior to sampling at constant volume.

As described in the Introduction section, these last two sets of simulations described two different scenarios, which we refer to as *fixed* and *pinned*. In the *fixed* set of simulations, the dipole moment of a central water molecule was aligned with the z-axis along which the electric fields were applied. The Cartesian coordinates of this water molecule were kept fixed using harmonic position restraints with a large force constant (500 kJ/(mol Å^2). In the *pinned* set of simulations, the harmonic position restraint was only applied to the oxygen atom of the central water molecule, thus allowing the latter to rotate freely. These two scenarios allow for different high-resolution analyses of water properties in the environment of the restrained water molecule, as described in detail in subsection II-B.

Each of the four sets of simulations with the TIP4P/2005 model were simulated in the zero-field regime and in the presence of electric fields E, of varying magnitude between 0.5 and 2.0 V·nm^{-1}. The fields were implemented using an external force along the Z-axis that acts on all atoms proportional to their partial atomic charge. In all simulations, long-range electrostatics were treated with the particle mesh Ewald algorithm on a 1.2 Å spatial grid with a fourth order interpolation and tin-foil boundary conditions. [126] Short-range electrostatics and Lennard-Jones forces were shifted to zero between 9.0 and 10.0 Å. Long-range dispersion corrections were applied for the energy and pressure. Neighbor lists were updated every 10 femtoseconds (fs) (equilibration) or 8 fs (production) with a 13.0 Å distance cutoff. The time step for integration was set to 1 fs (for both equilibration and production). The SETTLE algorithm [63] was used to constrain the intramolecular degrees

of freedom of water molecules.

The bulk water systems with 1024 molecules were simulated with a time step of 0.5 fs and coordinates were stored every 1 fs to allow for direct comparisons with the AIMD simulations using identical analysis protocols. The first set of simulations at constant volume was equilibrated for each applied field strength for 100 ps using the Berendsen weak-coupling thermostat with a reference temperature of 300 K and a time constant of 1.0 ps. [129] This was followed by 250 ps of production simulations with the Nose-Hoover thermostat and a time constant of 2.0 ps. [130] The second set of simulations of these systems at constant pressure used the same protocol with added barostats at a reference pressure of 1 bar. The Berendsen weak-coupling barostat[129] and the Parrinello-Rahman barostat[298] were used during equilibration and production, respectively, with a time constant of 1.0 ps.

Simulations of the *fixed* and *pinned* systems with 2147 water molecules were equilibrated for each electric field strength at constant pressure with the same basic protocol as described above. However, the production simulations in the isothermal-isobaric ensemble were extended to 80 ns in this case to sample 800 independent microstates in the NPT ensemble, *i.e.*, coordinates and velocities every 100 ps. Each of these microstates was then used to initiate microcanonical (NVE) simulations of 100 ps length each with 1 fs time steps, which were used for the subsequent analysis. During these 800 production runs per system, coordinates and velocities were saved every 8 fs. Compared to the previous sets of simulations that were directly compared to the AIMD simulations, we increased the integration time step to 1 fs for the equilibration of the *fixed* and *pinned* systems in consideration of the lack of fast intramolecular vibrations in the rigid TIP4P/2005 model. In the NPT simulations, we increased the time step further to 2 fs to increase the sampling speed. For the following simulations in the microcanonical ensemble used for the analysis, we decreased the time step again to 1 fs to ensure strict energy conservation and accurate dynamics in accord with previously established protocols for entropy estimations using 3D-2PT.[199, 208, 299]

6.4.2 Details on the 3D-2PT Method and System-Specific Analyses

All entropy calculations were performed with the two-phase thermodynamics (2PT) formalism proposed by Goddard and co-workers, [300–302] which we recently expanded into a grid-based version, 3D-2PT, that allows us to analyze solvent properties with spatial resolution.[199, 208, 299] The method is based on the vibrational density of states (VDoS), which corresponds to the Fourier transform of velocity time correlation functions and can be evaluated separately for translational and rotational degrees of freedom as well as intramolecular vibrations, if present.

$$\begin{aligned} I_{\mathrm{tr}}(\nu) &= \frac{2}{k_{\mathrm{B}}T}\int_{-\infty}^{+\infty} \mathrm{e}^{i2\pi\nu\tau} m_{\mathrm{w}} \langle \mathbf{v}_{\mathrm{com}}(t)\, \mathbf{v}_{\mathrm{com}}(t+\tau)\rangle \, d\tau \\ I_{\mathrm{rot}}(\nu) &= \frac{2}{k_{\mathrm{B}}T}\int_{-\infty}^{+\infty} \mathrm{e}^{i2\pi\nu\tau} \sum_{k=1}^{3} I_k \langle \omega_k(t)\, \omega_k(t+\tau)\rangle \, d\tau \\ I_{\mathrm{vib}}(\nu) &= \frac{2}{k_{\mathrm{B}}T}\int_{-\infty}^{+\infty} \mathrm{e}^{i2\pi\nu\tau} \sum_{v=1}^{3} \mu_v \langle \dot{q}_v(t)\, \dot{q}_v(t+\tau)\rangle \, d\tau \end{aligned} \tag{6.2}$$

Here, I_{tr}, I_{rot} and I_{vib} describe the VDoS of the translational, rotational and vibrational degrees of freedom, respectively, computed from time auto-correlation functions of the molecular center of mass velocity $\mathbf{v}_{\mathrm{com}}$ with mass m_{w}, angular velocity ω_k around rotational axis k with moment of inertia I_k, and time-derivative of (intramolecular) vibrational coordinate q_v with reduced mass μ_v; k_{B} is Boltzmann's constant, T the absolute temperature, and angular brackets $\langle ... \rangle$ indicate ensemble averages over the simulation time t and equivalent water molecules. For this study, we modified the open-source version of 3D-2PT [199], intended for simulations with rigid water models, and added a module for the analysis of the intramolecular vibrations (present in the AIMD simulations) as originally described by Lin *et al.*[301, 302].

The main feature of the 2PT framework is that the VDoS of translational and rotational degrees of freedom obtained from MD simulations are decomposed into separate contributions that are treated by simple analytical model systems. Specifically, the translational

degrees of freedom are described as a superposition of a hard sphere fluid and a set of harmonic oscillators that cover the range of all frequencies sampled in the simulation. By construction, the hard sphere fluid describes all diffusive components of the VDoS that are sampled at zero frequency.

A similar approach is used for rotational degrees of freedom of simulated molecules. Here, the hard sphere fluid is replaced by the rigid rotor model to describe zero-frequency components of the rotational VDoS. The intramolecular vibrations of water molecules lack diffusive processes at zero frequency and are thus simply treated as a spectrum of harmonic oscillators.

Standard expressions in statistical mechanics for the thermodynamic properties of hard sphere fluids, rigid rotors and harmonic oscillators are then used to describe the thermodynamic properties of the simulated liquid based on the decomposition of its VDoS described above. While the accuracy of the approach is intrinsically limited by the simplifications introduced through the analytical models, absolute molar entropies obtained with 2PT from molecular simulations have been shown to correlate well with experimental data.[301] In addition, applications within the 3D-2PT framework have shown that changes in entropy, *e.g.*, in response to the presence of a solute, benefit from the cancellation of systematic errors and can be obtained within the accuracy limits of the applied simulation model, *e.g.*, empirical force fields used in the simulation.[208, 299].

Average molar entropies without spatial resolution were obtained for the AIMD simulations and the TIP4P/2005 simulations of the smaller systems with 1024 water molecules. To compute velocity time correlation functions from these simulations, atomic velocities were first reconstructed using finite differences and atomic coordinates saved at 1 fs intervals. For each molecule, the atomic velocities were then converted into velocities of the molecular center of mass (translations), angular velocities around the rotational axes with distinct moments of inertia (rotations), and in the case of AIMD simulations, intramolecular

vibrations. Time correlation functions of the distinct velocity components were then computed for a maximum correlation time of 1.0 ps, averaged for all water molecules, Fourier transformed into the corresponding VDoS and used to estimate the entropy with the 2PT formalism described briefly above and in detail elsewhere. [208, 300, 301]

For the *fixed* and *pinned* systems simulated with the TIP4P/2005 water model, we used atomic velocities evaluated during the simulation and stored in the trajectory together with atomic coordinates at time intervals of 8 fs. To analyze molecular entropies with spatial resolution, we defined a three-dimensional (3D) grid around the restrained central water molecule. The grid contained $80 \times 80 \times 80$ volume elements (voxels) with a grid constant of 0.5 Å. The orientation of the analysis grid was aligned with the axes of the coordinate system and was kept unchanged during the analyses, for both the *fixed* and *pinned* systems. The analysis of the translational and rotational VDoS described above (the TIP4P/2005 water model lacks intramolecular vibrations) was then performed for each voxel of the analysis grid using a standard protocol adopted from previous 3D-2PT studies [208, 299]. Water molecules analyzed for a given time frame of the trajectory were assigned to a voxel based on their center of mass position. Auto-correlation functions of water translational and rotational velocities were computed for correlation times up to 1.6 ps and Fourier transformed into the corresponding VDoS. The resulting local VDoS for the translational and rotational degrees of freedom for each voxel were averaged over all simulation trajectories for each system and used to estimate the local water entropy with the 2PT formalism. [208, 300, 301] In the *pinned* set, at zero field we expect isotropic physical properties in the surrounding of the freely rotating central water molecule while at non-zero field any sign of anisotropy with increasing field intensity is due to the alignment of the molecules of the system (including the central water molecule). This is not valid for the *fixed* set: due to the set orientation of the central water molecule, the results are anisotropic even in the zero-field regime. Notably, since both the analysis grid and the central water molecule in the *fixed* system do not

change their orientation with respect to the simulation box and therefore do not change their reciprocal orientation either, an advantage of the *fixed* system is that we can characterize properties of water in the environment along specific directions (*i.e.*, analysis along the O-H bond vectors or along the water dipole moment). On the other hand, in the *pinned* set of simulations the rotation of the central water molecule in the fixed analysis grid results in a spherically symmetric environment at zero field strength and cylindrical symmetry along the *Z*-axis with the applied electric fields, as schematically shown in Fig. 1.

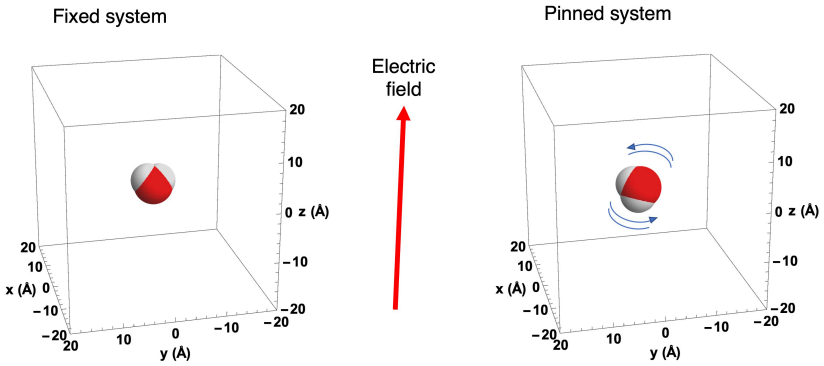

Figure 6.1: Schematic visualization of the simulated reference systems. (Left) *Fixed* system: a water molecule located at the center of the simulation box and analysis grid is fully restrained to maintain near-perfect alignment with respect to the electric field direction. (Right) *Pinned* system: only the position of the oxygen atom of central water molecule is restrained, thus allowing free rotation of the H_2O molecule. Hydration environment is not shown for clarity.

6.5 Results

6.5.1 Classical Force-field vs. ab initio *Water Entropies*

Before determining a spatially-resolved molecular picture of the structural and thermodynamic properties of the bulk water environment, it is worth to preliminarily test the

accuracy – in the absence of the field – of the commonly employed TIP4P/2005 [296] water parametrization with respect to two different dispersion-corrected Density Functional Theory (DFT) exchange-correlation (XC) functionals belonging to diverse classes, *i.e.*, the BLYP+D3 [288–291] and the SCAN+rVV10. [293, 294] Whereas the former represents by far one of the most employed Generalized Gradient Approximation (GGA) density functionals in *ab initio* MD (AIMD) simulations of bulk liquid water, the latter is considered the forefront of meta-GGA DFT functionals for reproducing the behavior of aqueous systems. In particular, the SCAN functional corrected by the rVV10 dispersion parametrization is capable of overcoming some of the serious drawbacks characterizing the previous generation of DFT XC functionals belonging to the GGA class. As shown in Fig. 2-a, SCAN+rVV10 (dotted blue curve) is capable of faithfully reproducing the oxygen-oxygen radial distribution function $g_{OO}(r)$ emerging from neutron scattering experiments [4] (profile on top of the grey shaded area) at ambient conditions. Interestingly, whilst Grimme's D3 [290, 291] dispersion corrections are not capable of adequately softening the intermolecular H_2O interactions, which result to be sizably overestimated by the BLYP functional (dashed red curve) even at a temperature higher than the ambient one, the TIP4P/2005 force-field reproduces spatial correlations between the water molecules in bulk liquid water at room temperature remarkably well (black solid curve). Thus, the hydrogen-bond (H-bond) network organization emerging from the $g_{OO}(r)$ appears to be adequately described by the computationally efficient and rigid TIP4P/2005 water model, similarly to cutting-edge and demanding dispersion-corrected first-principles meta-GGA functionals.

The over-structuring induced by the employment of GGA XC functionals visibly emerges from the evaluation of structural order parameters such as the well-known orientational tetrahedral order parameter q. [303] As shown in Fig. 2-b, indeed, the distribution $P(q)$ in absence of the external electric field shows that the population of tetrahedrally-ordered water molecules is overestimated in the BLYP+D3 framework with respect to the TIP4P/2005

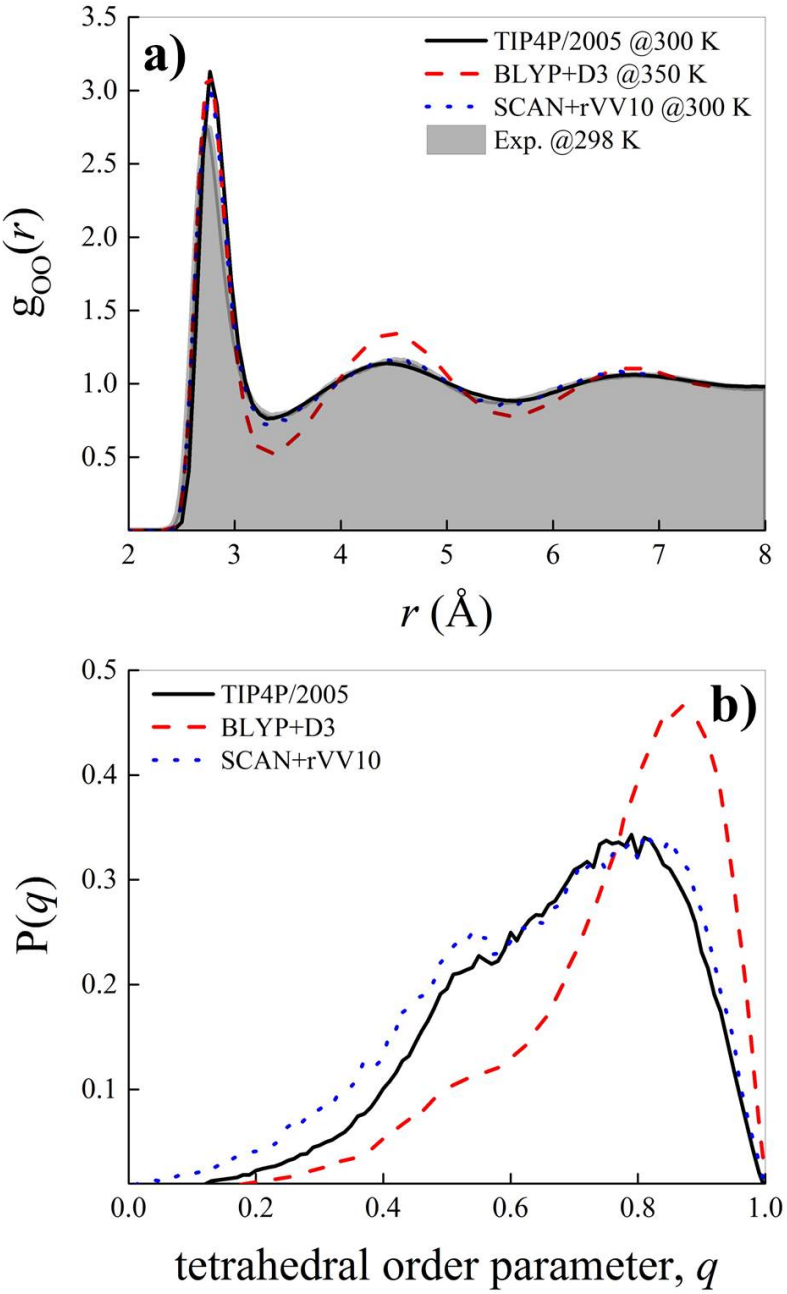

Figure 6.2: (a) Oxygen-oxygen radial distribution functions as determined *via* the classical force-field TIP4P/2005 (black solid line), the dispersion-corrected DFT functionals BLYP+D3 (dashed red line) and SCAN+rVV10 (dotted blue line), and by neutron scattering experiments (grey line above the shaded area). [4] (b) Distribution of the orientational tetrahedral order parameter at zero-field regime in liquid water as simulated by means of the classical force-field TIP4P/2005 (black solid curve), the GGA DFT functional BLYP+D3 (dashed red curve), and the meta-GGA DFT functional SCAN+rVV10 (dotted blue curve).

force-field case. Incidentally, the orientational order observed with this latter classical potential is adherent to results stemming from the predictive meta-GGA SCAN+rVV10 functional (Fig. 2-b). Albeit meta-GGA DFT functionals are currently computationally too demanding to allow for the generation of prolonged trajectories in presence of an electric field, the adherence of the structural data stemming from the rigid TIP4P/2005 water model gives us confidence in regard of the quality of its description of structure-related thermodynamic properties, such as the entropy. On the other hand, before considering the local entropy density estimates for the TIP4P/2005 model only, we compared total entropy changes induced by the application of an external electric field in simulations with the TIP4P/2005 force-field and BLYP+D3, one of the most employed gradient- and dispersion-corrected DFT functionals.

Since the structural ordering emerging from the analysis of the data shown in Fig. 2 is tightly related to the estimate of the entropy, it is not surprising that the total entropy estimated for BLYP+D3 water is lower than that determined for TIP4P/2005 (*i.e.*, 44.50 ± 0.26 J·mol^{-1}K^{-1} *vis-à-vis* 54.17 ± 0.06 J·mol^{-1}K^{-1}, respectively), as shown in Fig. 3 (dashed black lines). Here, statistical uncertainties of the computed molar entropies are reported as the standard error of the mean obtained by averaging over non-overlapping 10 ps segments of the trajectory as shown in Fig. 3 (the first 10 ps segment in each trajectory is treated as additional equilibration time and omitted from the analysis). Due to the already discussed over-structuring of water in GGA XC functionals, [304] the recorded value is significantly lower than the experimentally measured one at room temperature (*i.e.*, 69.95 J·mol^{-1}K^{-1}). [305] In fact, our AIMD results for zero field are comparable to an earlier 2PT calculation reported by Galli's group for AIMD simulations of water with the PBE functional (*i.e.*, 42.74 J·mol^{-1}K^{-1}). [306] However, we note that a study by Kühne and co-workers observed a higher value for the entropy of PBE water (*i.e.*, 51.32 J·mol^{-1}K^{-1}). [302] While our trajectory for the meta-GGA SCAN+rVV10 functional is relatively short (*i.e.*, $\sim$ 50 ps), we

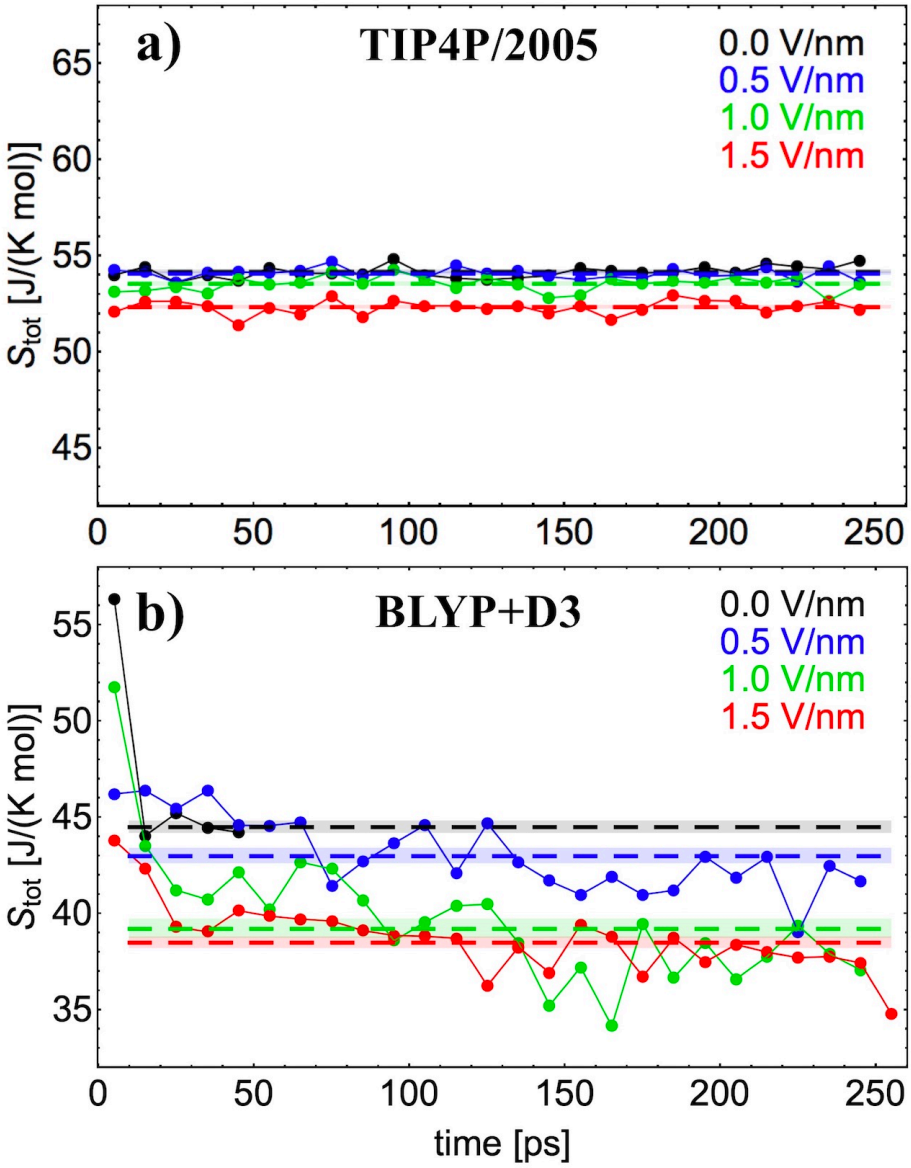

Figure 6.3: Total entropy as determined from classical force-field (a) and *ab initio* molecular dynamics (b) trajectories under the action of progressively increasing field strengths (see legend) in constant volume simulations. For the analysis, all simulation trajectories were split into 10 ps fragments, which were analyzed separately. The first 10 ps fragments of each simulations were considered as equilibration and averages over all remaining fragments are shown as dashed colored lines.

obtained an average water entropy of 52.17 ± 0.97 J·mol^{-1}K^{-1}, which is still significantly lower than the experimental value but only marginally lower than the entropy obtained from classical simulations with the TIP4P/2005 model. 2PT provides quantitative estimates of the molar entropy of liquids, which can contain systematic errors in addition to errors in the simulation model. For most water models, 2PT tends to underestimate the absolute molar entropy of water [302] as observed here for both the TIP4P/2005 model and the SCAN+rVV10 simulations in comparison to the experimental reference value. Further, by construction, the formalism is sensitive to the diffusion coefficient and related finite-size effects. [307] However, such systematic errors cancel out for entropy differences, [299] for example, between simulations of water with the same model but in the presence of distinct external fields.

When the external electric field is switched on at 0.5 V·nm^{-1}, the initial decrement of the total entropy appears to be very modest, especially in the TIP4P/2005 series of simulations (Fig. 3-a). When electrons are treated as quantum entities in BLYP+D3 simulations, the observed field-induced entropy changes are also moderate, accounting to a reduction – in the first-principles case – of less than 2 J·mol^{-1}K^{-1} (Fig. 3-b). However, when we compare Figs. 3-a and 3-b, it is apparent that the entropy of TIP4P/2005 water is significantly less sensitive to polarization than BLYP+3D water. In fact, whereas a maximum entropy decrement of ∼ 3% is observed at the highest field regime in the classical simulations (*i.e.*, from 54.17 ± 0.06 to 52.33 ± 0.08 J·mol^{-1}K^{-1}), a reduction of ∼ 14% (*i.e.*, from 44.50 ± 0.26 to 38.49 ± 0.30 J·mol^{-1}K^{-1}) emerges for the BLYP+D3 AIMD simulations at 1.5 V·nm^{-1}. This is likely due to the fact that the system treated at the BLYP+D3 level is able to respond to the external electric field up to higher field strengths owing to an explicit treatment of the electronic polarization. In fact, a more evident conversion toward more ordered molecular configurations, associated with a smaller total entropy, is recorded in the BLYP+D3 case, as displayed in Fig. 3-b. Beyond this threshold, AIMD simulations executed at the

BLYP+D3 level have shown only a marginal further decrease of the total entropy. Besides, for field intensities beyond 1.5 $V \cdot nm^{-1}$ not negligible field-induced molecular dissociations and consequent proton transfers are recorded, [273] rendering the entropy analysis not comparable to the classical – non-polarizable and non-dissociable – force-field case. In further contrast to the rigid TIP4P/2005 water model, vibrational degrees of freedom are taken into account in the AIMD case, in addition to translational and rotational contributions to the total entropy. In a nutshell, the spectral features emerging upon field exposure in the VDoS reveal an increased formation of H-bonds, in agreement with infrared (IR) and Raman spectroscopic analyses conducted on equivalently treated *ab initio* water. [200]

The AIMD and TIP4P/2005 simulations discussed above were performed at constant volume and thus ignore field-induced effects in the structure of water that affect its density. To estimate the importance of density effects for the reported molecular entropy, we performed a second set of simulations for the TIP4P/2005 water model at constant pressure as described in the Methods section. At zero field, the TIP4P/2005 water simulation at constant pressure increased its density to 1.00 $g \cdot cm^{-3}$, which reduced its entropy to 51.91 ± 0.16 $J \cdot mol^{-1}K^{-1}$ relative to 54.17 ± 0.06 $J \cdot mol^{-1}K^{-1}$ observed at the constant volume with a density of 0.97 $g \cdot cm^{-3}$. With increasing field strength, the density increased gradually to 1.04 $g \cdot cm^{-3}$ for the highest field strength of 1.5 $V \cdot nm^{-1}$. However, the response of the molecular entropy remained very similar to the constant volume simulations. For the largest field strength of 1.5 $V \cdot nm^{-1}$, the molar entropy decreased relative to the zero-field regime by $\sim 2\%$ to 50.69 ± 0.11 $J \cdot mol^{-1}K^{-1}$ compared to a $\sim 3\%$ decrease observed in the constant volume simulations. Hence, compared to the $\sim 14\%$ decrease of BLYP+D3 water, the TIP4P/2005 water simulations remained equally insensitive to the field strength under constant volume and constant pressure conditions.

In light of relatively recent path integral AIMD (PI-AIMD) simulations executed under the effect of externally applied electric fields, [273] it is tempting to consider the impact

that the inclusion of Nuclear Quantum Effects (NQEs) would have on the entropy estimate of liquid water. However, the 2PT formalism here adopted is based on the evaluation of the velocity time correlation functions which is, by definition, a dynamical property that cannot be reliably measured from path integral simulations. [298]

On the other hand, some effort has been done in the literature to determine the impact of NQEs on some structural properties of liquid water. It is well-known, indeed, that zero-point motion and proton tunnelling lead to faster reorientational dynamics and to a slight de-structuring of the radial distribution functions, [308] presumably producing larger entropy estimates. However, it is worth stressing the fact that the inclusion of NQEs in classical force-fields molecular dynamics simulations and in the first-principles ones may generate entropic contributions of completely different physical nature. In fact, PI-AIMD trajectories simulating the quantum nature of nuclei and electrons exhibit a non-negligible amount of autoprotolysis events even in absence of external electric fields. [273, 309] Such a circumstance, which cannot be caught and reproduced within non-dissociable water force-fields, prevents reliable one-to-one comparisons of the genuine impact of NQEs on the entropy of classical and *ab initio* simulations of liquid water.

Although it has been shown that the application of increasingly stronger fields produces a progressive reduction of the total entropy of water, a corresponding order-maker action is not clearly reflected by field-induced changes of the tetrahedral order parameter of water. As shown in Fig. 4-a, the tetrahedral order actually decreases with increasing field strength in water simulations with the TIP4P/2005 force field. For the *ab initio* simulations with BLYP+D3, the tetrahedral order also initially decreases for a field strength of 1.0 $V \cdot nm^{-1}$, but then increases when the strength of the polarizing field is raised to 2.0 $V \cdot nm^{-1}$. Hence, the decrease in water entropy with increasing field strength observed in Fig. 3 is not associated with an increase in tetrahedral order. Instead, the decrease of the water entropy is primarily associated with the increase in the orientational polarization of the water molecules.

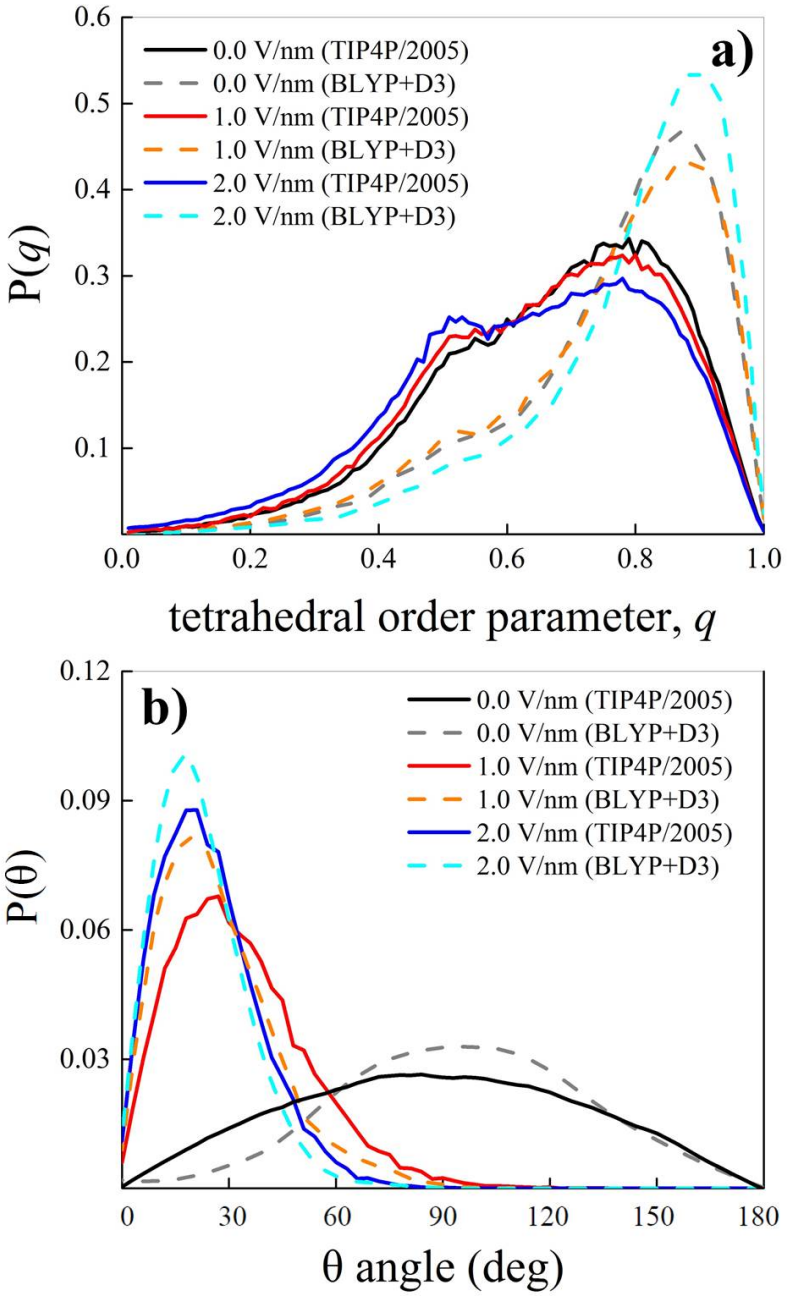

Figure 6.4: (a) Distributions of the orientational tetrahedral order parameter for different field strengths as determined from TIP4P/2005 molecular dynamics (solid curves) and *ab initio* BLYP+D3 molecular dynamics (dashed curves) simulations. (b) Distributions of the angle θ formed between the instantaneous molecular dipole vectors and the electric field direction for diverse field intensities applied along the *Z*-axis in TIP4P/2005 water (solid curves) and BLYP+D3 *ab initio* water simulations (dashed curves).

This can be observed in Fig. 4-b, where we show the distributions of the instantaneous angles θ between molecular dipoles of water with the Z-axis of the coordinate system. Small values of the angle θ correspond to molecular dipole moments that are aligned with the electric fields applied along the Z-axis. Albeit in absence of the field all dipole orientations are equivalent, the imposition of the external electrostatic potential gradient induces a prompt reorganization of the water dipole vectors. As has been shown for other computational models of water, [310] the gradual saturation observed at the strongest fields resembles the behavior of free dipoles described by the Langevin-Debye equation for simulations with the TIP4P/2005 force field and *ab initio* BLYP+D3 simulations. [311] This observation indicates a weak resistance of the molecular dipoles to polarization by very intense fields. Such a field-induced dipole locking leads, among other things, to a progressive reduction of the tetrahedral translational order parameter q_k, [303] as listed in Table I.

Table 6.1: Average tetrahedral translational order parameter $< q_k >$ at different field intensities determined from simulations of water performed with the TIP4P/2005 classical parametrization and with the first-principles BLYP+D3 functional. Since typical variations of this order parameter are manifest at the 4^{th}-5^{th} digit, parentheses are used to highlight the field-induced changes.

Field strength ($V \cdot nm^{-1}$)	TIP4P/2005	BLYP+D3
0.0	(0.998)9901913	(0.998)8946246
1.0	(0.998)9259704	(0.998)8113564
2.0	(0.998)7114147	(0.998)8017632

Interestingly, the water over-structuring induced by the adoption of the BLYP+D3 functional leads to a more complete and concerted reorientation of the dipoles with respect to the TIP4P/2005 case at the same field strengths, in accordance with the results of Fig. 3 and as displayed in Fig. 4-b. Finally, our finding that the orientational polarization of BLYP+D3 water is already nearly complete for 1.0 $V \cdot nm^{-1}$ – as well as the respective translational order parameter – is consistent with the observation that also the entropy reduction reaches a

substantial saturation above this threshold (Fig. 3-b). By contrast, in the TIP4P/2005 case, the orientational polarization of water molecules at 1.0 V·nm^{-1} is much less pronounced (Fig. 4-b) and increases further with increasing field strength – similarly to the reduction of the respective translational order parameter – while the total entropy decreases (Fig. 3-a).

6.5.2 Local Molecular Entropies

To analyze molecular entropies of TIP4P/2005 water as a function of the electric field, we used the 3D-2PT approach introduced by Persson *et al.* [199, 208, 300, 301] to spatially resolve the local water entropy in the environment of a central water molecule. It is noteworthy that, differently from standard calculations of isotropically averaged distribution functions, the spatially-resolved approach adopted here allows us to resolve structural anisotropies with a high resolution owing to the 0.5 Å grid constant of the 3D analysis grid. For every volume element of the analysis grid, we have calculated the entropy per water molecule for the two investigated systems: the *fixed* and the *pinned* system (see the Methods section and Fig. 1).

In Tables II and III we report the bulk water entropy, obtained as an average for voxels separated by 10 Å or more from the central reference water molecule in the investigated field regimes for the *fixed* system (Table II) and the *pinned* one (Table III). The changes of

Table 6.2: Total, translational and rotational entropy – in J·mol^{-1}K^{-1} – determined for bulk water with the 3D-2PT method for different intensities of the electric field in the *fixed* system. Statistical uncertainties are standard errors of the mean obtained from 800 independent microcanonical simulations.

Field strength (V·nm^{-1})	S_{tot}	S_{tr}	S_{rot}
0.0	55.77±0.02	46.87±0.02	8.90±0.02
1.0	55.27±0.02	46.56±0.02	8.71±0.02
2.0	53.69±0.01	45.32±0.01	8.38±0.01

the entropy per molecule relative to the zero-field regime are visualized in the histograms

Table 6.3: Total, translational and rotational entropy – in $J \cdot mol^{-1}K^{-1}$ – determined for bulk water with the 3D-2PT method for different intensities of the electric field in the *pinned* system. Statistical uncertainties are standard errors of the mean obtained from 800 independent microcanonical simulations.

Field strength ($V \cdot nm^{-1}$)	S_{tot}	S_{tr}	S_{rot}
0.0	55.75±0.02	46.85±0.02	8.90±0.02
1.0	55.17±0.02	46.48±0.02	8.68±0.02
2.0	53.37±0.02	45.06±0.02	8.31±0.02

of Fig. 5. For the total molar entropy of bulk water in the zero-field regime, S_{tot}^{E0}, we obtain

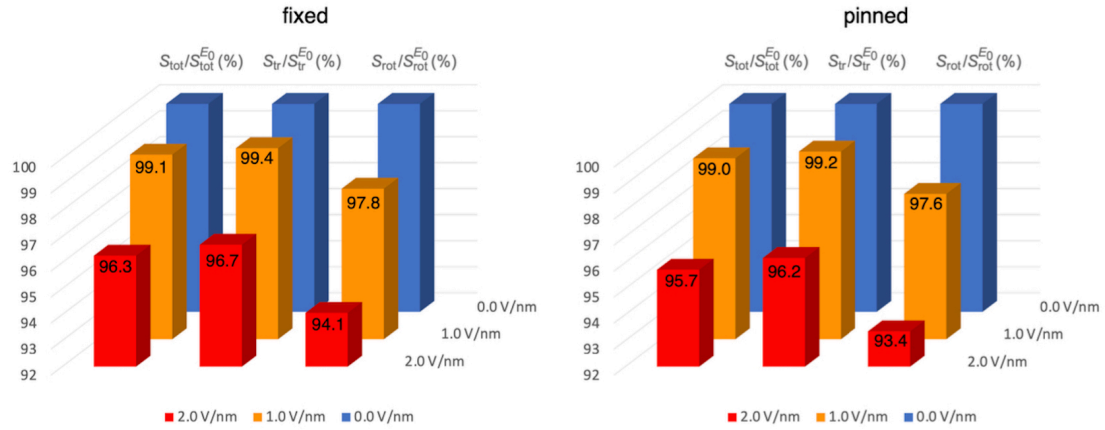

Figure 6.5: Histograms showing the field-induced changes of the bulk water entropy per molecule (total, translational and rotational contributions) relative to the zero-field regime at different field intensities. Respective values are reported in Tables II and III. Left: *fixed* system. Right: *pinned* system.

55.77 ± 0.02 $J \cdot mol^{-1}K^{-1}$ (*fixed* system) and 55.75 ± 0.02 $J \cdot mol^{-1}K^{-1}$ (*pinned* system). These results deviate slightly from the molar entropy of water obtained for the NVT simulation of 1024 water molecules at zero field (54.17 ± 0.02 $J \cdot mol^{-1}K^{-1}$) used for comparison to the AIMD simulations in the previous section, which is the result of finite size effects in the latter case as well as a lower frequency resolution in the analyzed VDoS.

The bulk water entropies in the *fixed* and *pinned* are statistically identical at zero field, because the position restraints applied to the central water molecule in both cases have no effect on an isotropic system at rest (no net translations or rotations). Minor differences be-

tween both systems are only observed for non-zero fields as the systems become polarized, even though only water molecules with a minimum distance of 10 Å to the central water molecule are considered for the analysis of bulk water properties in both systems. The position restraint applied to the oxygen atom of the central water molecule in the *pinned* system still has no impact on thermodynamics. However, under polarized conditions the restrained orientation of the central water molecule in the *fixed* system modifies interactions between water molecules and thus has thermodynamic consequences, which are still detectable at distances of 10 Å and beyond.

In addition to the total entropy per bulk water molecule, we also report separate contributions from translational and rotational degrees of freedom in Table II and III and in Fig. 5. These are computed separately from fluctuations of the center-of-mass velocity and angular velocities for rotations around the three distinct rotational axes of water molecules. The largest contribution ($\sim$ 84%) to the entropy of bulk water molecules in the absence of a polarizing field comes from the translational degrees of freedom, *e.g.*, $S^{E0}_{trans} = 46.87 \pm 0.02$ J·mol^{-1}K^{-1} in the fixed system. In contrast, the rotational degrees of freedom contribute $\sim$ 16% to the entropy per water molecule, e.g., $S^{E0}_{rot} = 8.90$ J·mol^{-1}K^{-1}. With increasing field strength, the total entropy per bulk water molecule decreases by $\sim$ 4% for the highest field strength of 2.0 V·nm^{-1}. This total entropy decrease is primarily associated with the translational degrees of freedom due to their dominant contribution. However, as shown in Fig. 5, the relative change of the entropy contributions is significantly larger for the rotational degrees of freedom, *e.g.*, $\sim$ 6% and $\sim$ 7% in the *fixed* and *pinned* systems, respectively. This corresponds to the expectation of preferentially oriented water molecules in the presence of electric fields. However, given that the rotational entropy in our decomposition is only a minor contributor to the total entropy of water, even a large relative change remains a minor contribution to the field-induced change of the total entropy. Instead, the total entropy change in response to the field remains dominated by the small relative change

of the significantly larger translational entropy.

We note that the polarization-induced changes in the total entropy per water molecule remains small compared to structural changes associated with phase transitions.The ΔS_{tot} associated with the freezing of liquid water at 273 K is $\sim$ 22 J·mol^{-1}K^{-1}. Thus, it is straightforward to conclude that electric fields below the ionization threshold are not capable to cause orientational polarization-induced electrofreezing in bulk water (no confinement) at room temperature. This finding is in agreement with – and extends the range of applicability of – recent experiments that demonstrate that the ice nucleation temperature of water is independent from the application of external electric fields up to 0.1 V·nm^{-1}. [312] This is in contrast with recent AIMD and PI-AIMD simulations that have shown the possibility to electrofreeze another important H-bonded system like ammonia at temperatures and pressures at which it is in the liquid state. [278] Of course, determination of electric-field-induced effects on enthalpy and free energy, additionally to those on the entropy, would further clarify on this intriguing phenomenon of enormous importance. However, the estimate of the free energy and of the enthalpy in presence of an external electric field is not straightforward. In fact, the value of the electric field strength to be included in the thermodynamic equations is, in general, quite different from the applied external field due to the intrinsic additional field generated by the polarized surface of the cavity, which depends on the specific geometry of the system and on the anisotropy of the dielectric constant. [297]

In Fig. 6, the entropy density defined as the product of the water entropy per molecule and the water number density for each voxel, is shown for the *fixed* and *pinned* systems in the *XZ*-plane. The color code describes changes in the local entropy density in the environment of the molecule placed at the origin of the analysis grid relative to bulk water at zero field. Due to the fully restrained position of the central water molecule in the *fixed* system, the results are already anisotropic in the absence of an externally applied electric field. Interactions with the restrained central water molecule force water molecules to adapt

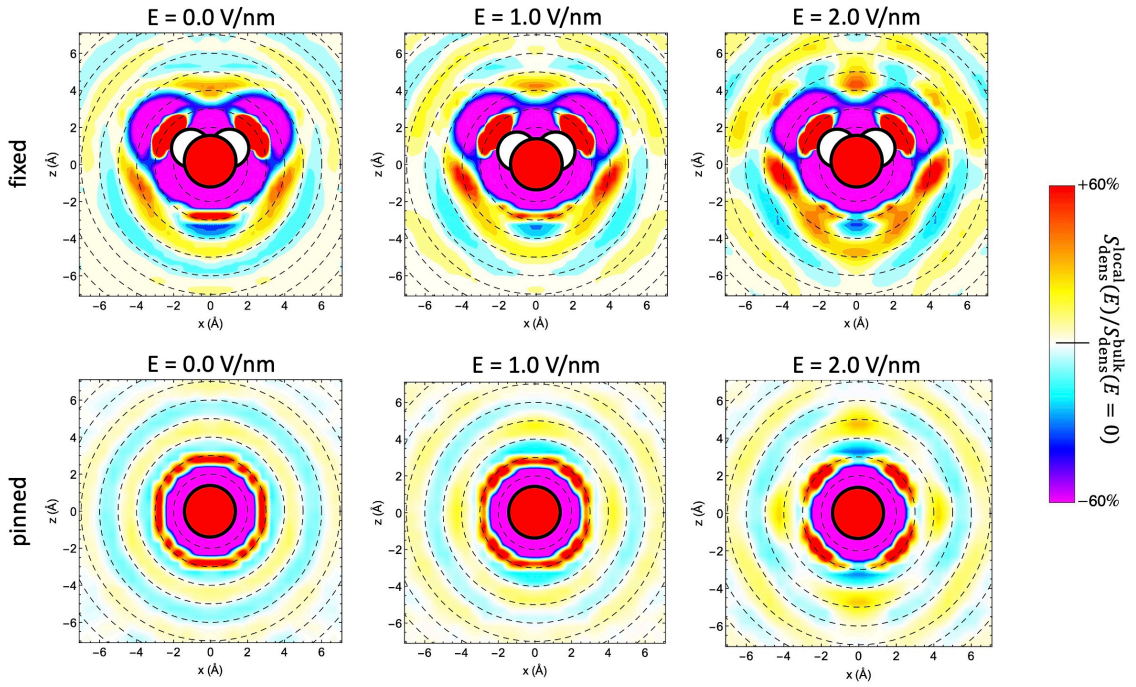

Figure 6.6: *Fixed* (top) and *pinned* (bottom) systems. Spatially-resolved changes of the entropy density at different field intensities relative to the bulk entropy density at zero field in the *XZ*-plane. Left column: no field; middle column: $E = 1.0$ V·nm^{-1}; right column: $E = 2.0$ V·nm^{-1}. In the center of each panel, the restrained atom(s) of the central water molecule are shown and dashed circles indicate 1 Å distance increments from the center starting at 2 Å.

their position and orientation. In the *pinned* system, the environment of the central water molecule is isotropic at zero field, while anisotropy is gradually induced when an electric field is applied. In the *fixed* system, a local increase in the entropy density relative to bulk water with increasing field strength is particularly visible for sites localized in the second coordination layer (4-5 Å) as shown in Fig. 6 (top) despite the overall decrease in the entropy of bulk water. In the *pinned* system (Fig. 6, bottom), whose thermodynamic properties are equivalent to a bulk water system, the applied fields restructure the isotropic entropy density into a cylindrically symmetric density around the *Z*-axis. A decrease of the local entropy density is observed in the first coordination layer (3 Å) in the *XY*-plane and along the *Z*-axis, which is followed by a local increase in the second coordination layer (4-5 Å). The trend is opposite at ±45° angles with the *Z*-axis, where an increase in the entropy

density is observed in the first coordination layer, while a decrease is found for the second coordination layer. Notably, the local water entropy densities in the environment of the central water molecule observed in Fig. 6 are strongly affected by the local water number density in both systems, which effectively describes the 3D water-water pair correlation function. Thus, we analyze in the following the spatially resolved water number density and its response to the applied fields in more detail.

6.5.3 Water Number Density

We analyzed the average local water number density in the environment of the central water molecule on the same analysis grid as used for the analysis of local entropy densities above. In Fig. 7, we show the contour plots of the water number density, $\rho(E)$, normalized to the bulk value at zero field, $\rho_{bulk}(E0)$, in the environment of the *fixed* molecule placed at the origin of the analysis grid. Data are shown for different field intensities (*i.e.*, 0.0 V·nm^{-1}, 1.0 V·nm^{-1}, and 2.0 V·nm^{-1}), in the *XZ*-plane (top), the *YZ*-plane (middle), and the *XY*-plane (bottom) whilst – as stated above – the field is applied along the *Z*-direction. The average number density of bulk water at zero field has been obtained from voxels separated by 10 Å or more from the central water molecule. As seen earlier in Fig. 6, interactions with the restrained central water molecule already result in an anisotropic environment at zero field (left panels in Fig. 7) in accord with other reports of 3D pair correlation functions of water in a molecular reference frame. [313] Increased local densities of water are observed in the first coordination layer, specifically in H-bond acceptor sites in the *XZ*-plane (top panel) and the H-bond donor sites in the *YZ*-plane (middle panel) at 3 Å distance from the center. H-bond acceptor and donor sites are surrounded by a reduced number density nearby due to repulsions with water molecules in the first shell. Additional enhanced densities can be observed at ~4 Å, which correspond to water molecules in the second coordination layer of the central water molecule. The anisotropic structure of water changes notably as a func-

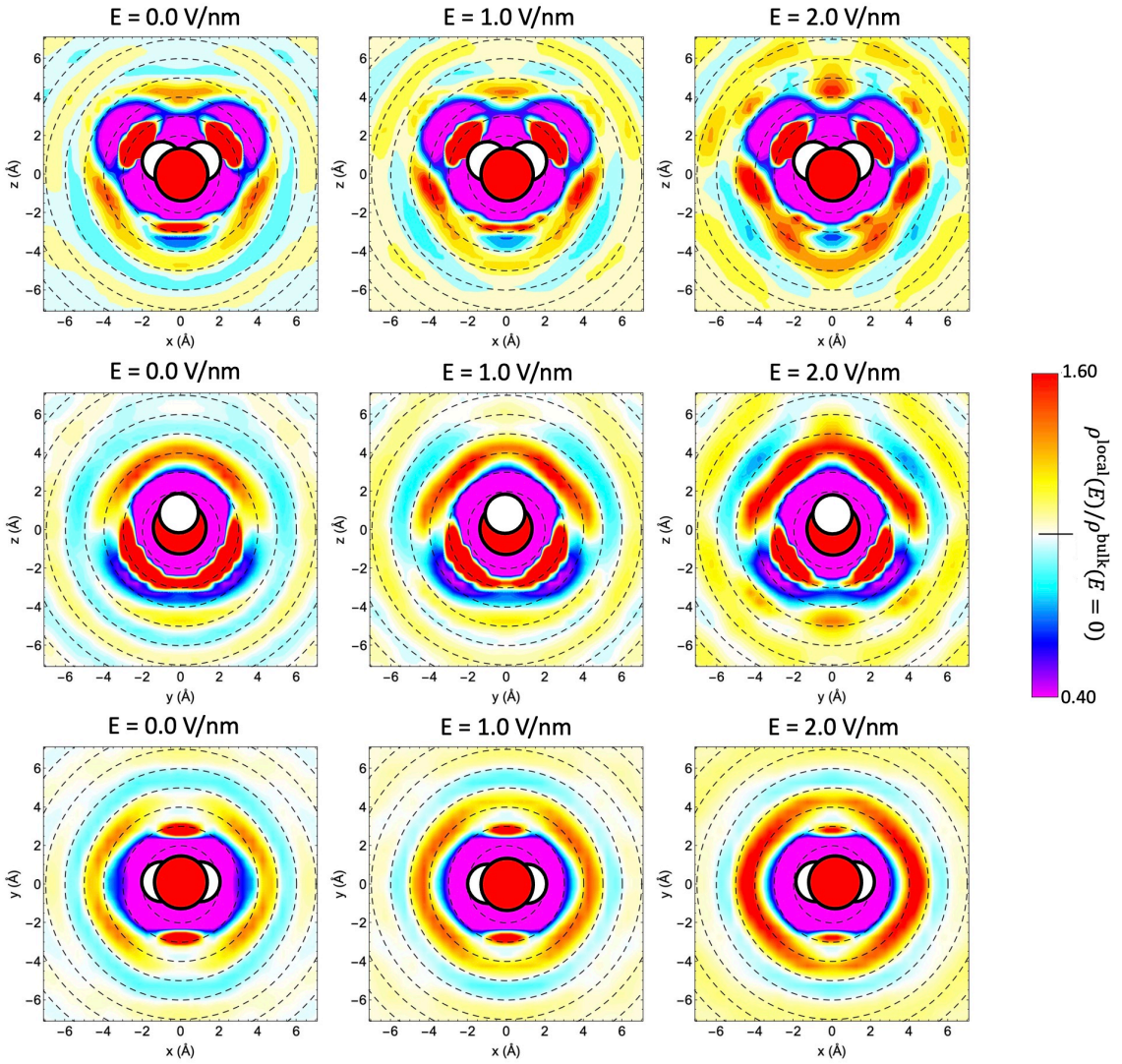

Figure 6.7: *Fixed* system. Contour plots of the average water number density normalized to the bulk value at zero field for different electric field intensities (field in the Z-direction). Left column: no field; middle column: $E = 1.0$ V·nm^{-1}; right column: $E = 2.0$ V·nm^{-1}. Top: XZ-plane. Middle: YZ-plane. Bottom: XY-plane. In the center of each panel, the restrained atoms of the central water molecule are shown and dashed circles indicate 1 Å distance increments from the center starting at 2 Å.

tion of the applied electric field. In particular, Fig. 7 shows that the locally increased water densities in the second hydration shell increase further with electric field strength in both the XZ- and YZ-planes. This is accompanied by a less structured first hydration shell that is best observed in modified peak intensities of oxygen-oxygen radial distribution functions (see discussion of Fig. 8 below). The first peak in the radial distribution function decreases

slightly with increasing field strength, while the second peak increases significantly.

Our analysis of the local environment of the *fixed* water molecule thus shows that the application of a static electric field induces an enhanced layering of the first two hydration shells. The second solvation shell is affected by the field to a larger extent with respect to the first shell, in fairly good agreement with previous force-field molecular dynamics simulations by Shiafiei *et al.* [310] and classical AIMD and path-integral AIMD simulations by some of us. [273] This observation can be explained by the comparatively weaker correlations in the second hydration shell, which are easier to perturb by the applied electric field compared to the strong correlations in the first hydration shell that are the result of direct H-bonds with the central reference water molecule.

In Fig. 8, we analyzed changes in the local water structure around the central water molecule in the *pinned* system using contour plots of the local water number density in the *XZ*- and *XY*-planes. Due to the unhindered rotation of the central water molecule, its environment is isotropic in the absence of an electric field (left panels) and retains cylindrical symmetry around the *Z*-axis once the field is applied (*i.e.*, the *XZ*- and *YZ*-planes are symmetrically equivalent). The coordination layers, in particular the first one identifiable by a red ring at $\sim$ 3 Å, can be easily located in the field-free contour plot maps of Fig. 8 (left column). By applying an increasingly intense electric field, we can observe that the progressive dipolar alignment of the entire system – and in particular of the central water molecule – along the field direction results in manifest structural anisotropies in the *XZ*- and *YZ*-planes (Fig. 8, top). In other words, by increasing the field strength, the coordination shells show an enhanced layering toward specific directions due to the preferential orientation of all water molecules in the system and the resulting changes of intermolecular interactions.

In particular in the *XY*-plane and along the *Z*-axis, we observe a decrease in the local water number density in the first hydration shell ($\sim$3 Å) and an increase in the second hydra-

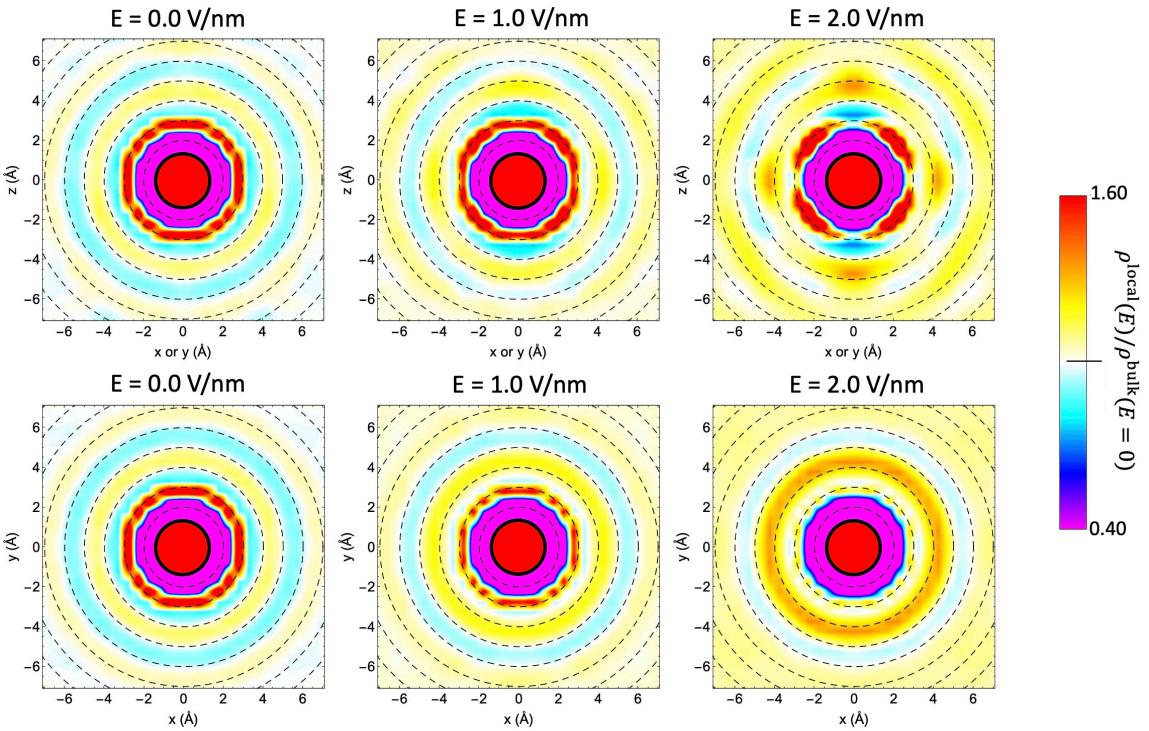

Figure 6.8: *Pinned* system. Contour plots of the average water number density normalized to the bulk value at zero field for different electric field intensities (field in the *Z*-direction). Left column: no field; middle column: $E = 1.0$ V·nm^{-1}; right column: $E = 2.0$ V·nm^{-1}. Top: *XZ*- or *YZ*-planes. Bottom: *XY*-plane. In the center of each panel, the restrained oxygen atom of the central water molecule is shown and dashed circles indicate 1 Å distance increments from the center starting at 2 Å.

tion shell (~4-5 Å). However, the opposite trend is observed at an angle of 45° with respect to the *Z*-axis, especially at the most intense field strength explored here (*i.e.*, 2.0 V·nm^{-1}), as shown in Fig. 8 (top right panel). This latter evidence suggests a field-induced strengthening of the H-bonds of the central reference molecule with the first-neighboring molecules along specific directions, a finding in agreement with the computational spectroscopic indication of the tendency of liquid water toward the formation of more "ice-like" structures upon field exposure at these intensities. [200]

The increased local density of water in the second hydration shell at a distance of ~4-5 Å from the center in the *XY*-plane and along the *Z*-axis are accompanied by a depletion of water density in the first hydration shell. Depleted zones are observed along the *Z*-axis

at about ~3-4 Å (Fig. 8, top right panel) as a consequence of the steric exclusion imposed by the nearby water molecules which – under the field action – come closer to and establish stronger H-bonds with the central water molecule. On the other hand, a different response to the field can be visualized in the *XY*-plane (Fig. 8, bottom panel) where, at the highest field intensity, the peak of the first coordination layer essentially disappears, whilst that of the second coordination layer is increased. Such a finding strongly indicates that whereas first neighbors approach each other in the field direction (*Z*-axis), they are pushed farther in the plane perpendicular to the field (*XY*-plane). This peculiar scenario is also caught and somehow exaggerated within the simpler *fixed* scenario, as shown in Fig. 7 (bottom panel, right).

6.6 Conclusion

In this work, we analyze the electric-field-induced changes in the total and local entropy of bulk liquid water by means of classical and first-principles simulations. In the absence of an external field, a comparison between the rigid force-field TIP4P/2005, the GGA BLYP+D3 and the meta-GGA SCAN+rVV10 *ab initio* functionals, has shown that key structural properties of water appear to be adequately described by the computationally efficient TIP4P/2005 water model, which is in close agreement with the structure of water obtained from AIMD simulations with the cutting-edge SCAN+rVV10 functional.

Related to this, the molecular entropy of bulk water that we estimated for the TIP4P/2005 water model using the 2PT formalism (54.17 $J{\cdot}mol^{-1}K^{-1}$) is in close agreement with the corresponding result for AIMD simulations with the meta-GGA SCAN+rVV10 functional (52.17 $J{\cdot}mol^{-1}K^{-1}$). That both estimates are significantly lower than the experimental reference value of 69.95 $J{\cdot}mol^{-1}K^{-1}$ can be attributed to a systematic underestimation of the water entropy using the 2PT formalism, which is based on the vibrational density of states. However, both the TIP4P/2005 force field and the meta-GGA SCAN+rVV10 functional

result in significantly higher estimates for the molar entropy of water than GGA DFT functionals such as BLYP+D3, which results in significant overstructuring of the liquid even at elevated temperatures and yields a molar entropy of only 44.50 $J \cdot mol^{-1}K^{-1}$. However, the molar entropy of TIP4P/2005 water seems to be significantly less sensitive to polarization-induced changes compared to BLYP+D3 water due to the inclusion of the electronic polarization in the latter.

In addition to these general results, we have presented a molecular-dynamics-based analysis (3D-2PT) that provides spatially resolved information on the local entropy and number density of water in the environment of a reference molecule. This method allowed us to generate 3D maps of changes in the water structure that are induced by electric fields. By means of this approach, a field-induced strengthening of the H-bonds of the reference molecule with the first-neighboring molecules along specific directions has been observed. This resulted to be in agreement with the vibrational Stark effect indicating the tendency of liquid water of evolving toward more "ice-like" structures upon field exposure at intensities of $\sim$ 1.0 $V \cdot nm^{-1}$. [200] However, the largest field-induced entropy changes are too feeble (*i.e.*, about 4% decrease for TIP4P/2005 water and 14% in the BLYP+D3 case) for realizing the so-called electrofreezing phenomenon. The fact that this finding holds up to strengths just beneath the molecular dissociation threshold (2.0 $V \cdot nm^{-1}$) extends the range of applicability of recent experiments showing the independence of the water nucleation temperature from the application of external electric fields up to intensities of 0.1 $V \cdot nm^{-1}$. [312]

Owing to the 3D-2PT formalism, we have shown that the field-induced local entropy density changes are strongly modulated by field-induced water number density changes, allowing to depict a microscopic picture of the molecular entropic effects of the field. It turns out that the main effect produced by a static electric field on the water number density is that of inducing an enhanced layering, *i.e.*, an over-coordination in correspondence of the second shell, though with anisotropic structural modifications. Under intense fields,

water molecules undergo a progressive dipolar alignment and adopt preferential orientations parallel to the field lines, forming H-bond structures preferably along a direction which is diagonal with respect to the field axis.

Chapter 7

PREDICTION OF DISORDERED ALPHA-MOLECULAR RECOGNITION FEATURES (α-MORFS) BY EXAMINING FOLDING AND BINDING FREE ENERGIES

7.1 Summary

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are pivotal in performing critical biological functions such as cell signaling, regulation, and recognition. Unlike conventional proteins, IDPs and IDRs do not possess a well-defined three-dimensional structure, which is typically vital for the function of most proteins. This inherent lack of structure presents a significant challenge in the field of molecular biology, particularly in developing innovative methodologies to elucidate the functional mechanisms of IDPs and IDRs. Furthermore, this structural ambiguity complicates the design of high-affinity ligands, which are essential for modulating the activity of these proteins in therapeutic applications.

Here in this research, we utilize free energy calculations to investigate the potential of using binding-induced folding mechanisms for the development of high-affinity ligands targeting intrinsically disordered proteins (IDPs). Molecular recognition features (MoRFs) are specific functional regions within intrinsically disordered regions (IDRs) that undergo transitions from disorder to order upon binding, such as the transition from a disordered state to an alpha-helix structure in alpha-MoRFs. Theoretically, MoRFs can also serve as targets for the design of drug molecules that can bind to these regions in IDPs associated with diseases, thereby deactivating or modulating their properties. We measure the free energy cost associated with the transition of variable peptide sequences from unstructured forms into

folded secondary structure motifs. Unlike existing structure prediction tools, our atomistic simulation protocol enables us to identify the minimum free energy structures and quantify the free energy difference between folded and disordered conformational states. This quantification is crucial for the development of drugs targeting a specific IDR in a folded conformation, as the binding free energy must offset the free energy cost of folding. Our approach seeks to address the challenge of the ”undruggability” of IDPs and IDRs. Furthermore, we compute the binding free energy of three distinct alpha-MoRFs bound to the Mcl-1 host protein to test a hypothesis. This hypothesis posits that the experimental binding free energy of the host-guest complex is equal to the sum of the folding and binding free energies derived from umbrella sampling simulations. This pathway assists us in predicting the experimental free energy of binding in protein complexes.

7.2 Introduction

Intrinsically disordered proteins (IDPs) are characterized by their absence of a stable three-dimensional structure, primarily due to their unique amino acid composition, often rich in charged residues, glycines, and prolines [18, 20]. Despite their structural variability, IDPs play crucial roles in vital physiological functions. They are involved in several fundamental cellular processes, including the regulation of transcription, signal transduction, the formation of multiprotein complexes, and the development of membraneless organelles [23, 314]. Approximately 40% of the human proteome is believed to be disordered or to have lengthy disordered regions (exceeding 100 residues), with a more significant fraction containing disordered segments of up to 30 residues [24–26]. A key aspect of their functionality is their ability to interact with globular proteins, where IDPs can either adopt a specific conformation when bound or maintain their dynamic and disordered nature [209]. These significant functional regions in IDPs are referred to as Molecular Recognition Features (MoRFs) or Molecular Recognition Elements (MoREs). MoRFs facilitate interactions with structured partner proteins and can experience transitions from disorder to order during interaction [210, 211, 315]. MoRFs typically vary in size and can be up to 70 residues in length, often situated within larger intrinsically disordered regions [316]. Due to their flexible structure, MoRFs can accurately interact with their partners, making them crucial players in regulatory processes and signal transduction [31].

Based on the specific folded structure they adopt after interacting with globular targets, MoRFs can be classified into four subtypes: α-MoRFs, β-MoRFs, ι-MoRFs, and complex-MoRFs [18, 316, 317]. These subtypes are characterized by the formation of α-helices, β-strands, irregular secondary structures, and a combination of multiple secondary structures, respectively.

Among the various types of MoRFs, the most prevalent and extensively studied class in-

volves IDPs and IDRs that fold into α-helices upon binding to their targets [212–214]. This type is known as α-helix forming molecular recognition elements/features α-MoREs/MoRFs [212] or simply as helical binding motifs (HBMs) [318]. Usually, these sequences are composed of short amphipathic helices that transition from a disordered state to a helical conformation upon binding with their target partners [319]. These motifs can be predicted based on their sequence [320, 321] and they are especially prevalent in eukaryotic transactivation domains [215] and $\alpha\alpha$-hub proteins [216].

Sections of proteins containing α-MoRFs were initially thought to be predominantly disordered when unbound and unrestrained. However, subsequent research has revealed that these regions can exhibit a considerable degree of residual helical structure [322]. The partially formed helical configurations observed in α-MoRFs are often termed pre-folded elements, pre-structured motifs, or transient and residual helical structures. [36]. A previous bioinformatics study revealed a significant overlap between the residual structure and the structure of the IDP when bound to its target, suggesting that the residual structure predisposes the IDP to its target-bound conformation [323]. Even with a limited number of systems available at that time, the findings strongly suggested that pre-structuring plays a crucial role in target binding. A contemporary review of experimentally analyzed IDPs in their free state documented more than 50 instances of α-MoRFs exhibiting considerable secondary structure, without exploring the correspondence between residual and target-bound conformations [324].

The transient, residual helical structures are believed to impact the binding affinity (thermodynamics), kinetics, specificity, and structural characteristics of the IDP-target ensemble. There is increasing interest in understanding how these structures influence these properties. An early hypothesis from Dunker *et al*. suggested that the disorder-to-order transition in α-MoRFs leads to a reduction in binding affinity due to the folding penalty associated with α-MoRFs [10]. The significance of the folding penalty in IDPs, and particularly in

α-MoRFs, is now well-established and has been thoroughly reviewed in the literature, as discussed in Skriver *et al*. [325]. The initial analysis of binding affinities from various studies revealed that the average affinity for associations involving folding upon binding is generally lower compared to that of globular complexes, indirectly suggesting an estimate of the folding penalty in such interactions [326, 327]. In the research conducted by Teilum *et al*. [326], which examined a substantial number of protein complexes using experimental data, it was found that the less favorable entropic contribution in complexes involving ordered regions (helical in our case) and intrinsically disordered proteins (ORD-IDP complexes) is primarily due to a reduction in conformational entropy. This finding is consistent with the fundamental difference in the binding mechanisms of ordered and disordered ligands. A disordered polypeptide needs to undergo folding to create the final complex, resulting in a significant decrease in conformational entropy. It should be noted that equilibrium isothermal titration calorimetry (ITC) data cannot definitively determine the pathway followed by the ligand's folding during the binding process. Therefore, the contribution of the amount of folding free energy to the observed binding free energy is uncertain. It is likely that in some protein complexes, the IDP undergoes folding followed by binding in a conformational selection mechanism, whereas in others, the IDP folds upon binding in an induced fit mechanism. However, the variation in average Gibbs free energy ΔG° could still be explained by the necessary folding of the disordered ligand in the ORD-IDP complexes. The assessment of folding penalties has been conducted through a thermodynamic breakdown of entropic factors [328, 329] and the application of helix-coil models [217]. Additionally, the alteration in the helicity of α-MoRFs upon binding to their targets can provide estimates of the folding penalty [217]. A thorough evaluation of 25 distinct α-MoRFs using the helix-coil approach has indicated that the average folding penalty stands at around 2.5 kcal/mol. Nonetheless, numerous instances of ultra-tight picomolar binding affinities in folding-upon-binding interactions have been documented [330],

demonstrating that multiple factors such as post-translational modifications (PTMs), phosphorylation, oxidative stress, or initial binding to certain regions of the binding partners, can effectively counteract the folding penalty to attain ultrahigh binding affinity [331]. A key factor in overcoming the folding penalty is the presence of substantial amounts of residual helical structure, also referred to as pre-structuring or pre-folding. An examination of folding penalties in α-MoRFs revealed that pre-folding can lower the penalty by an average of 25% [217]. Experimentally, a reduction in the α-MoRF folding penalty can be seen when an enhancement in residual helicity (such as that induced by mutations) leads to an elevated binding affinity of the α-MoRF. Relationships between residual helicity and binding affinity have been documented across various systems, and they are in quantitative agreement with predictions from models [211, 217], underscoring the significance of residual structure in influencing the affinities between IDPs and their targets. Nevertheless, some research has identified a less pronounced correlation or even a more intricate relationship between residual helicity and binding affinity [332, 333].

In multiple distinct systems exhibiting folding-upon-binding, experimental studies have shown that the residual helical structure plays a role in modulating both binding affinity and kinetics. However, further insights are needed into how the IDP sequence relates to the structural ensemble (both unbound and bound) and how this affects the associated binding parameters, including both kinetic and equilibrium aspects. Specifically, the measurement of folding free energy using enhanced sampling techniques like umbrella sampling has not been extensively documented, due to challenges in defining an appropriate collective variable to describe the unfolding pathway. Additionally, the quantitative effects of point mutations involving known helix stabilizers like alanine, and helix disruptors such as proline and glycine, on the overall folding free energy have yet to be reported in detail.

To address these unexplored pathways, we employed atomistic umbrella sampling to explore the folding free energies in disordered peptides that transition to an alpha-helix upon

target binding, such as peptides containing α-MoRF, as well as fully disordered intrinsically disordered regions (IDRs). After comparing the free energy minimum structures from simulation, and from AlphaFold [334] predictions, all the α-MoRF examples we studied exhibited folding free energies within a range of 10 kJ/mol (2.4 kcal/mol), which is consistent with the range predicted by Hadzi *et al*. in their study [217] using a modified LR model[211, 335, 336]. They considered different sequence lengths of IDPs, which is one of the factors contributing to the discrepancy between our results and theirs. The relatively low folding free energy penalty of these sequences can be implemented in designing peptide-based or small molecule drugs that stabilize the helix structure by inducing folding upon binding to α-MoRFs.

Furthermore, we calculated the binding free energies of two selected alpha-MoRFs from BH3 domain-binding proteins (PUMA and NOXA-A) to Mcl-1 using coarse-grained (CG) simulations with the SIRAH force field. Our findings partially support our hypothesis that the experimental binding free energies reported by Dahal *et al*. [218] are the sum of both the computed binding and folding free energies. This result offers valuable insights into the potential binding-induced folding pathway followed by our chosen set of alpha-MoRFs.

7.3 Methods

7.3.1 Simulation Protocol for Folding Free Energy Calculation

All simulations concerning the folding free energy were performed using the GROMACS 2018.1 software package [228] and the PLUMED library [337, 338]. For system preparation, we constructed fully alpha-helical peptide sequences. We generated a template PDB file of a helical poly-alanine with the desired sequence length. The template file was then used to create a PDB file of the sequence with alanines replaced by other amino acids according to the sequence. We subsequently employed the SCWRL4 package [232] to incorporate any missing residues and adjust the conformations of the side chains, thereby generating the initial helical state of the sequence. We adjusted the protonation state of each amino acid residue to correspond with a pH of 7, utilizing standard side chain pKa values in solution.

Once the system was prepared, we carried out steered molecular dynamics (SMD) simulations [147] on the helical peptide. This was performed to generate snapshots or starting configurations for further analysis. The peptide was placed in a cubic simulation box with an edge length of 80 Å, and it was surrounded by explicit TIP3P water molecules to mimic an aqueous environment. To ensure the stability of the initial conformation of the peptide, we applied a force constant of 1000 kJ/mol to all atoms to maintain its initial configuration during equilibration. Next, we solvated the system with approximately 16700 water molecules to ensure adequate hydration of the peptide. Additionally, we introduced ionic species to the system to imitate a physiological salt concentration of 150 mM. We added close to 50 Na^+ and Cl^- ions. It's important to note that the exact numbers of these ions were adjusted based on the specific peptide sequence being studied, as different peptides may require different ionic concentrations to maintain charge neutrality and mimic physiological conditions. The presence of salt can affect the structural dynamics and phase separation tendencies of in-

trinsically disordered proteins (IDPs) through mechanisms like electrostatic screening and the salting-out effect, as highlighted in a recent study by Wohl *et al*.[233] This setup aimed to create a realistic simulation environment for the helical peptide, allowing us to study its behavior and interactions under conditions that closely resemble those in a biological system.

We implemented periodic boundary conditions (PBCs) across all dimensions to simulate an infinite system and avoid edge effects. For handling long-range electrostatic interactions, we employed the particle mesh Ewald (PME) algorithm [339], setting up a real space grid with a resolution of 1.2 Å. This method efficiently calculates electrostatic forces by dividing them into short-range and long-range components. For short-range electrostatic and Lennard-Jones (LJ) interactions, we established a cutoff distance of 10 Å. Interactions beyond this distance were not considered, which helps to reduce computational cost without significantly affecting accuracy. LINCS (Linear Constraint Solver)[234] algorithm was used to constrain high-frequency bond lengths such as NH, OH, and CH within the protein allowing us to use a longer time step of 2fs in our case when integrating the equations of motion. Additionally, we employed the SETTLE algorithm [340] to preserve the geometry of water molecules in the system. We performed an energy minimization process using the steepest descent algorithm before starting the molecular dynamics simulations with all the selected force fields. This initial step, involving 2000 steps of minimization, helps in relieving any steric clashes or unfavorable conformations in the system, providing a stable starting point for the dynamic simulations.

We then conducted a two-step equilibration process for each system. Initially, we equilibrated the system in the isothermal-isochoric (NVT) ensemble, applying a position restraint of 1000 kJ $mol^{-1}nm^{-2}$ to each heavy atom to maintain structural stability. Subsequently, we transitioned to the isothermal-isobaric (NPT) ensemble, maintaining the system at a temperature of 300 K and a pressure of 1 bar for 1 ns. During this phase, we utilized a time

step of 1 fs, a velocity rescaling thermostat with a time constant of 0.1 ps for temperature control, and a Berendsen weak-coupling barostat with a time constant of 2.0 ps to regulate pressure.

To represent the folding process of the helix during steered MD (SMD), we selected a single collective variable (CV), ξ that quantifies the number of hydrogen bonds between backbone atoms of residues that are four sites apart along the sequence (i.e., i, i+4 hydrogen bonds).

$$\xi(r_{\mathrm{i,i+4}}) = \sum_{i=1}^{n-4} \frac{1-(r_{\mathrm{i,i+4}}/d_0)^n}{1-(r_{\mathrm{i,i+4}}/d_0)^m} \tag{7.1}$$

This particular choice of CV was previously employed by Camilloni et al. [341] in their study to effectively capture the folding dynamics of helical structures. In enhanced sampling techniques, the differentiability of collective variables is essential for accurate force calculations, efficient sampling, convergence of simulations, and physical interpretability of the outcomes. In order to meet the algorithm's requirement for the CV to be a differentiable function of the microscopic coordinates, we calculated the hydrogen bond count using specific parameters: n = 6, m = 10, and d_0 = 0.32 nm. The variable $r_{i,i+4}$ represents the distance between the backbone oxygen of residue i and the backbone nitrogen of residue i+4. At the limit of $r_{i,i+4} = d_0$ for any of the hydrogen bonds in the sum ξ, L'Hôpital's rule [342, 343] is applied to that single term in the sum to obtain a finite value given by $\frac{n}{m}$, which evaluates to 0.6 when substituted with our chosen values of n and m. This CV does not account for angular contributions. Therefore, conformations where an oxygen atom at site i is near a nitrogen atom at site i+4 will contribute to the CV even if their positions do not meet the angular criterion for a hydrogen bond (i.e., the angle θ_{H-N-O} is less than 30°). Incorporating such a criterion would, on one hand, enhance the accuracy with which the collective variable (CV) depicts the formation of helices. However, it would also increase the region of the phase space where the gradient of the CV with respect to the microscopic atomic co-

ordinates is zero. The chosen values for the exponents m and n ensure that the derivatives across the range of distances $r_{i,i+4}$ never reduce to zero within the limits of numerical precision. This approach comes at the cost of slightly diminished accuracy in identifying the exact number of formed i - (i+4) hydrogen bonds. It's important to note that these variables do not necessitate any specific knowledge of the folded structure. For each system, we conducted 100 ns steered molecular dynamics (MD) simulations in the NPT ensemble at 300 K and 1 bar, using a time step of 2 fs. To maintain temperature and pressure, we employed a Nose-Hoover thermostat [344] and a Parrinello-Rahman barostat [345], both with a time constant of 2.0 ps. These simulations were aimed at generating snapshots of the unfolding process, transitioning from a fully alpha-helical state with a non-zero CV value to an unfolded state with a ξ value close to zero. The PLUMED library, patched on GROMACS, was utilized to specify the CV and execute the biasing.

We generated 61 starting structures using steered molecular dynamics (MD). Subsequently, we conducted umbrella sampling simulations for each of these structures. In the umbrella sampling (US) method [144], a biasing potential is applied at specific positions along the reaction coordinate (in this case, the collective variable (CV) representing the hydrogen bond number, ξ) to enhance the sampling of states with high free energy barriers. The reaction coordinate (RC) was segmented into 61 windows, each corresponding to one of the starting structures. A harmonic potential, as detailed in equation 7.2, was added to the original potential (unbiased potential) in each window. This was done to facilitate the exploration of the collective variable space by the protein.

$$\omega_i(\xi) = \frac{1}{2}k(\xi(r) - \xi_i)^2 \tag{7.2}$$

In this expression, ξ_i serves as the reference coordinate for a single umbrella window. During the simulation, if the system deviates from the reference coordinate, the added bias potential exerts a force to push the system back towards the mean position, ensuring en-

hanced sampling around this target coordinate. Here, k is the force constant and r (H-bond distances) are the microscopic coordinates. We employed the Weighted Histogram Analysis Method (WHAM) [159] to remove the bias from the restrained umbrella sampling (US) simulations and to construct potential of mean force (PMF) profiles using 400 bins along the reaction coordinate (RC).

7.3.2 Simulation Protocol for Binding Free Energy Calculation

All simulations pertaining to the binding free energy calculation were carried out using the GROMACS 2018.1 and 2022.5 software packages [228]. As test systems, we chose two α-MoRF BH3-only domain proteins of PUMA(PDB ID: 2ROC) and NOXAA (PDB ID: 2ROD), where each bound in a helix conformation to an antiapoptotic protein MCL-1. BH3-only domain proteins act as pro-apoptotic agents in many organisms including humans. Each test complex underwent identical preparation steps, starting with: (i) obtaining the experimental structures from the Protein Data Bank (PDB) website (http://www.rcsb.org/); (ii) utilizing PyMOL [346] for editing and removing additional chains from the PDBs to create interacting protein-ligand systems; (iii) filling all the missing residues or atoms in the PDB file using MODELER v9.15 [347].

Coarse-grained simulations for each of the three systems were carried out using the SIRAH force field (http://www.sirahff.com) [348]. The SIRAH CG force field is designed to overcome several limitations commonly found in coarse-grained force fields, such as the use of a uniform dielectric constant, the absence of long-range interactions, reliance on topological information for maintaining secondary structures, and limited consideration of ionic strength effects. Unlike the popular Martini force field [349, 350], which typically maps four heavy atoms to one coarse-grained bead, SIRAH maintains a higher level of detail in the peptide bonds, retaining the positions of backbone nitrogen (N), α-carbon (Cα), and oxygen (O) atoms, while the side chains are represented more coarsely. The WT4 water model

incorporated in SIRAH consists of four interconnected beads with partial charges, enabling the simulation of dielectric permittivity. Furthermore, the coarse-grained electrolytes in SIRAH can accurately replicate the effects of ionic strength and osmotic pressure. This residue-based coarse-grained model encompasses all interactions within a classical Hamiltonian, which is a standard feature in most molecular dynamics simulation packages.

Before using SIRAH for umbrella sampling, we generated 91 windows using VMD [351] software. The collective variable, d_{COM}, here is the distance of separation of the two center of masses (COM) between the MCL-1 and BH3-only domain proteins. With a window spacing of 0.05 nm, each receptor-ligand pair was separated to a maximum d_{COM} of 4.5 nm to allow sufficient separation, complete solvation and avoid residual interaction at maximum separation. Next, we applied SIRAH CG simulation protocol to perform umbrella sampling on each window. We adhered to the protocol described in Darré et al. [348] for conducting coarse-grained (CG) simulations. The mapping of coordinates and analysis were executed using the SIRAH tools [352]. Before transitioning to the SIRAH CG model, protonation states were determined assuming a neutral pH, utilizing the PDB2PQR server [353] and selecting the AMBER [187] naming scheme for the output. After converting from all-atom to coarse-grained (CG) representation, the protein complexes were positioned in a cubic box of 8 nm filled with SIRAH WT4 water. To achieve a neutral charge, Na+ and Cl− ions were added at a concentration of 0.15 mol/L. A water shell of 0.3 nm surrounding the protein was removed, leading to the addition of the following numbers of WT4 water molecules to each protein complex: 4995 for PUMA (2ROC) and 4991 for NOXA-A (2ROD). A large box size was chosen to guarantee that, at their maximum separation distance, the two proteins do not interact with their own periodic images. Each system was initially minimized using the steepest descent method for 10,000 steps. To allow for the equilibration of water molecules around the protein complex, each system was then subjected to a 5 ns simulation with the positions of all coarse-grained (CG) atoms in the complex harmonically restrained.

During these restrained simulations, the temperature of the systems was maintained at 300 K and the pressure at 1 atm using the V-rescale thermostat [235] and the Parrinello-Rahman barostat [345] with isotropic pressure coupling, respectively. Following this, unrestrained simulations were carried out for a duration of 15 ns. All the simulations were conducted

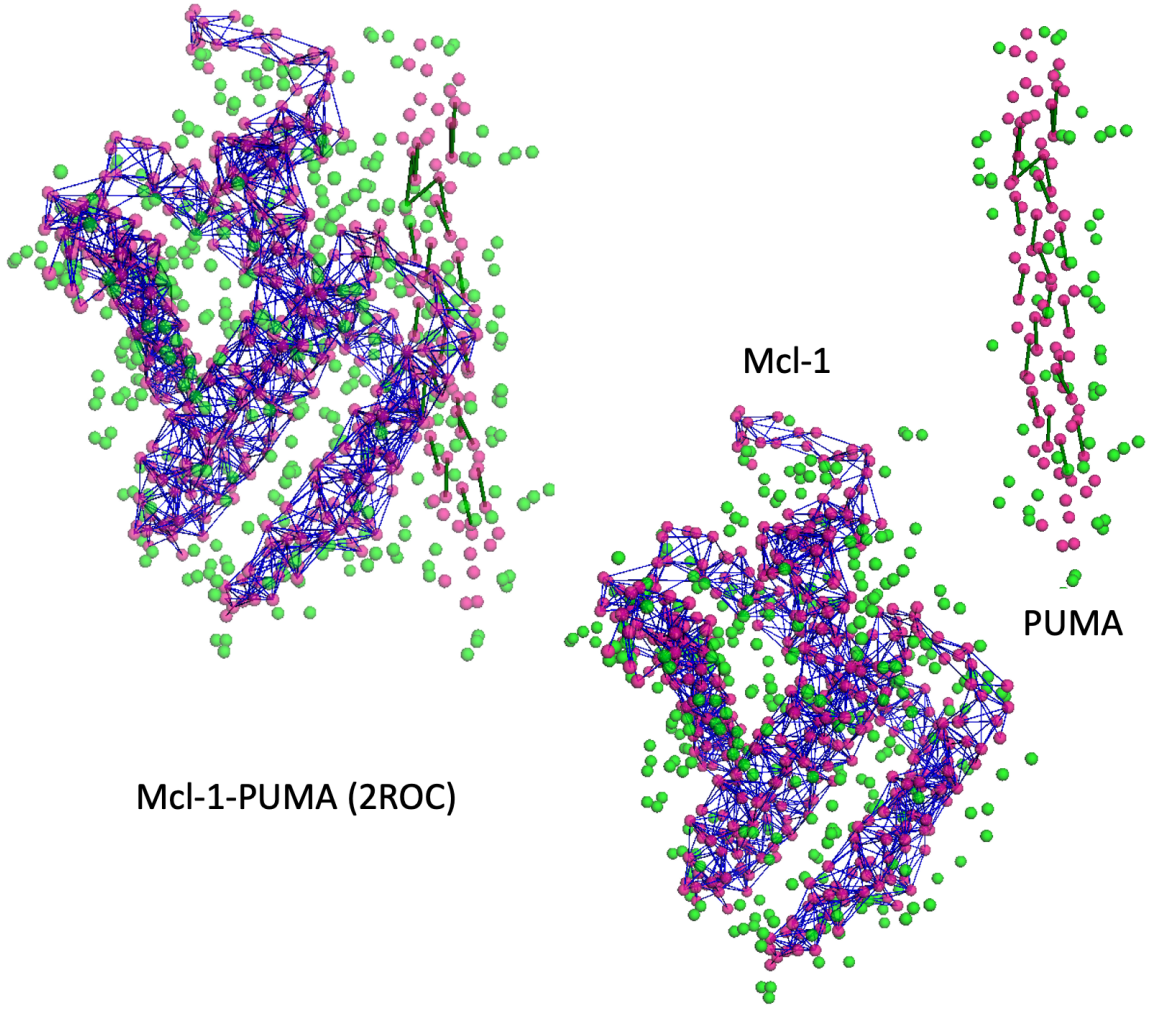

Figure 7.1: Elastic Network Model (ENM) applied to the backbone atoms of Mcl-1 and PUMA individually to maintain a stable coarsegrained structure while applying biasing potential during umbrella sampling. The distance restraints are applied between O-N, C-N, C-O, and N-O atoms within 5.25 Å radius. The pink and green spheres represent backbone and sidechains respectively.

with a time step of 20 fs and updated neighbor lists every 10 steps. Electrostatic interactions were calculated using the particle mesh Ewald method [339] with a direct cutoff of

1.2 nm and a grid spacing of 0.2 nm. For van der Waals (vdW) interactions, a cutoff of 1.2 nm was employed. To each umbrella window, a harmonic restraint with a force constant of 500 kJ $mol^{-1}nm^{-2}$ was applied to the center of mass of each protein undergoing unbinding to ensure sufficient overlap of the adjacent histograms.

Prior to initiating the umbrella sampling simulations we implemented distance restraints on the backbone atoms of Mcl-1 and PUMA separately. This was done to preserve a stable coarse-grained structure while applying the biasing potential during umbrella sampling, as illustrated in Fig. 7.1. The Elastic Network Model restraints [354–359] were applied between O-N, C-N, C-O, and C-C atom pairs within a 5.25 Å radius. As a result, the root-mean-square deviations (RMSDs) of both Mcl-1 and PUMA were maintained within a range of 5-6 Å, ensuring the structural integrity of the proteins during the simulations. To remove the bias from the simulations, we utilized the Weighted Histogram Analysis Method (WHAM) and constructed the PMFs using 400 bins along the reaction coordinate (RC). The final one-dimensional WHAM (1D-WHAM) profile was adjusted to eliminate the translational, configurational entropy term [360], in order to obtain the accurate potential of mean force (PMF) [361–365].

$$PMF_{True}(R) = PMF_{WHAM}(R) + k_B T ln(4\pi r^2) \tag{7.3}$$

This adjustment ensures a flat potential of mean force (PMF) at greater distances by compensating for the entropic decrease in the PMF, which is due to the increase in the number of configurations on a sphere of a given radius [365]. This is also done to keep the $Q_{unbound}$ from Eq. (7.4) concentration independent. Now, from the flattened PMF, the binding free energy (ΔG) for a protein complex was calculated by taking a difference of free energy in the bound and unbound (flattened PMF) states.

$$\Delta G_{bind} = -k_B T \; ln\left(\frac{Q_{bound}}{Q_{unbound}}\right) = -k_B T \; ln\left(\frac{P(\xi_{bound})}{P(\xi_{unbound})}\right) \tag{7.4}$$

Here, Q's are the partition functions, while $P(\xi_{bound})$ and $P(\xi_{unbound})$ represent the probabilities of states on the PMF associated with the bound and unbound states along the RC, ξ. This is discussed in Eq. (2.24) from Ch. 2.6.1 on umbrella sampling.

7.4 Results and Discussion

7.4.1 Folding Free Energy Penalty Analysis

To confirm the accuracy of our method for quantitatively estimating the folding free energies of disordered regions transitioning from coiled to alpha-helical states, we conducted a set of analyses. We focused initially on a 37-amino-acid segment of the c-Myc protein for this evaluation. This particular length was selected because it represents the longest disordered sequence identified as having the potential to fold into an alpha-helical structure. The c-Myc protein [366–371], with a molecular weight of 62 kDa and composed of 439 amino acid residues (https://www.uniprot.org/uniprotkb/P01106), contains multiple functional domains. These domains enable c-Myc to interact with a wide variety of binding partners, each contributing uniquely to the intricate network of interactions known as the c-Myc interactome. The transcription factor c-Myc is overexpressed in more than 70% of human cancers, both constitutively and abnormally. Studies have demonstrated that directly inhibiting c-Myc can lead to rapid tumor shrinkage in mice, accompanied by only mild and completely reversible side effects. This indicates that targeting c-Myc could be a feasible therapeutic approach. While it does contain intrinsically disordered regions (IDRs) that contribute to its flexibility and interaction with various binding partners, c-Myc also has structured domains, such as the basic helix-loop-helix/leucine zipper (bHLH/LZ) domain, which is crucial for DNA binding and dimerization. The presence of both structured and disordered regions allows c-Myc to function as a versatile transcription factor involved in the regulation of numerous genes.

We primarily chose four different sequences from the entire c-Myc protein sequence for our study. The first sequence was identified as a potential target due to presence of several adjacent alanines, and leucines known to stabilize helices. Moreover, AlphaFold's predicted structure for c-Myc (https://www.uniprot.org/uniprotkb/P01106) shows this region to be partially folded into an alpha-helix. The other three sequences were characterized by their high content of glycine, proline, or a combination of both. The residues of glycine and proline are known to destabilize alpha-helical structures in solution [372–375]. For membrane proteins, however, they were found to stabilize the helical structures instead [376]. As a reference sequence, we selected a polyalanine chain of 37 residues in length. This choice was based on its ability to maintain the highest alpha-helicity among all amino acids when simulated using the Amber03 force field [186, 377] and the TIP3P water model [378] in a 100 ns atomistic classical MD simulation. According to the literature, alanine is known to be the most stabilizing amino acid for α-helical structures, followed by leucine [374, 379–381].

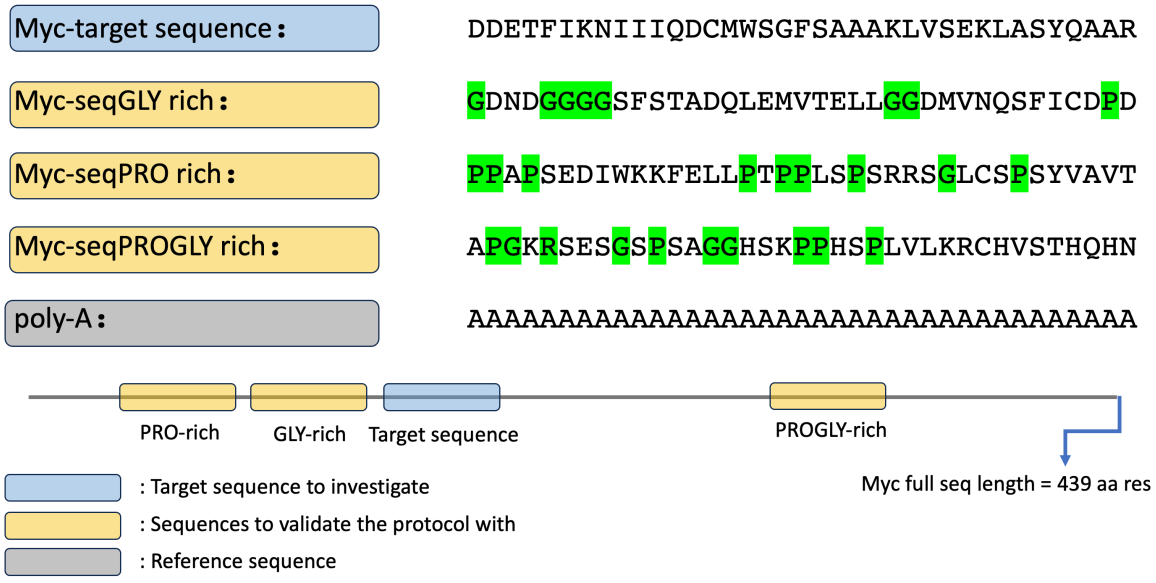

Figure 7.2: Selected sequences to examine for validation of protocol. The 37 sequence polyalanine acts as a reference with a nearly intact alpha-helix structure.

Fig. 7.2 presents the sequences we initially tested our protocol on. These sequences are located within the 439-residue chain of the c-Myc protein. The green highlighted residues are helix destabilizers present in the chosen sequences. These sequences were specifically chosen to deviate from the α-helical conformation. This deviation was intended to assist us in measuring the folding free energies with respect to a fully α-helical polyalanine reference.

Having determined the initial set of sequences for validating our protocol, we proceeded with the calculations of folding free energies. These calculations were carried out using umbrella sampling, as detailed in Sec. 7.3.1.

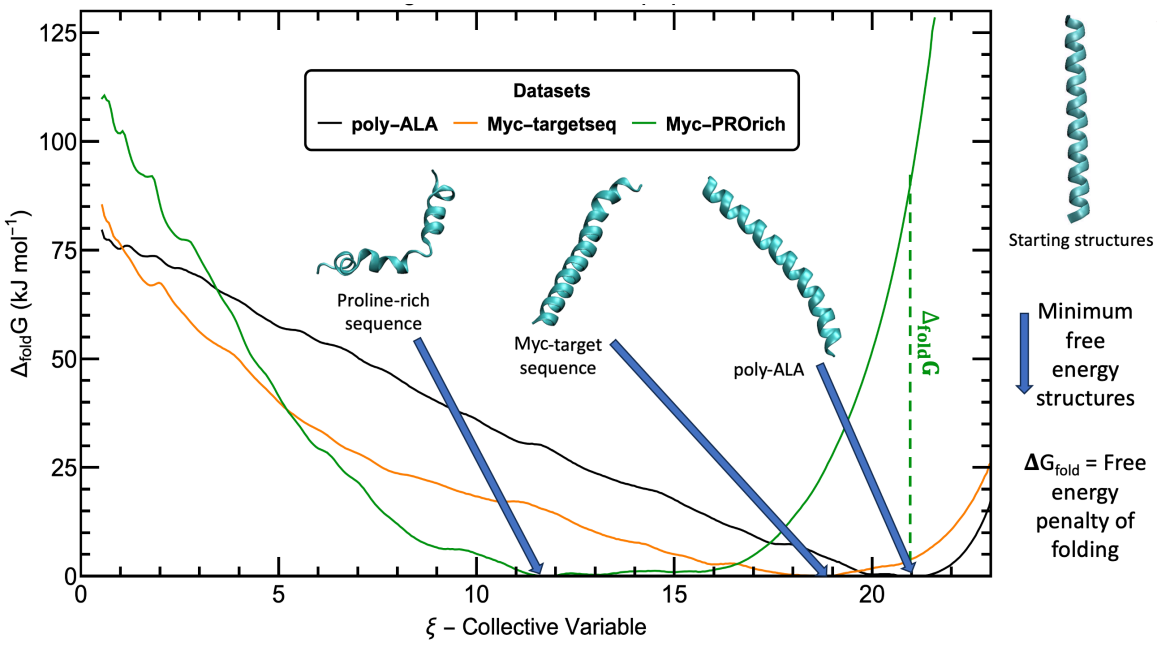

Figure 7.3: Potential Mean Forces (PMFs) for the formation of alpha-helical secondary structures obtained via umbrella sampling simulations. Each structures (shown by the representative arrows) represent the free energy minimum structure in that unfolding pathway from helix to coil. $\Delta_{fold}G$ penalty is measured by the vertical distance from the free energy minimum ξ or CV value of polyalanine(reference) to the corresponding free energy value of the PMFs of other sequences.

The graphs presented in Fig. 7.3 illustrate the unfolding process of a structure that begins in a fully alpha-helical conformation and transitions to an unfolded, coiled state. In this

representation, the reference structure of polyalanine is shown to maintain the closest resemblance to an alpha-helix. This is followed by our target sequence from the Myc protein and then by the sequence that is rich in prolines, which are known to destabilize alpha-helices. The deviation of a sequence from the position directly above the minimum value of the collective variable ξ for polyalanine provides us with the value of $\Delta_{fold}G$, which represents the folding free energy penalty associated with transitioning from a disordered state to an α-helical state. This analysis helps us quantify the stability of alpha-helical structures and the energetic cost of folding for different sequences within the c-Myc protein.

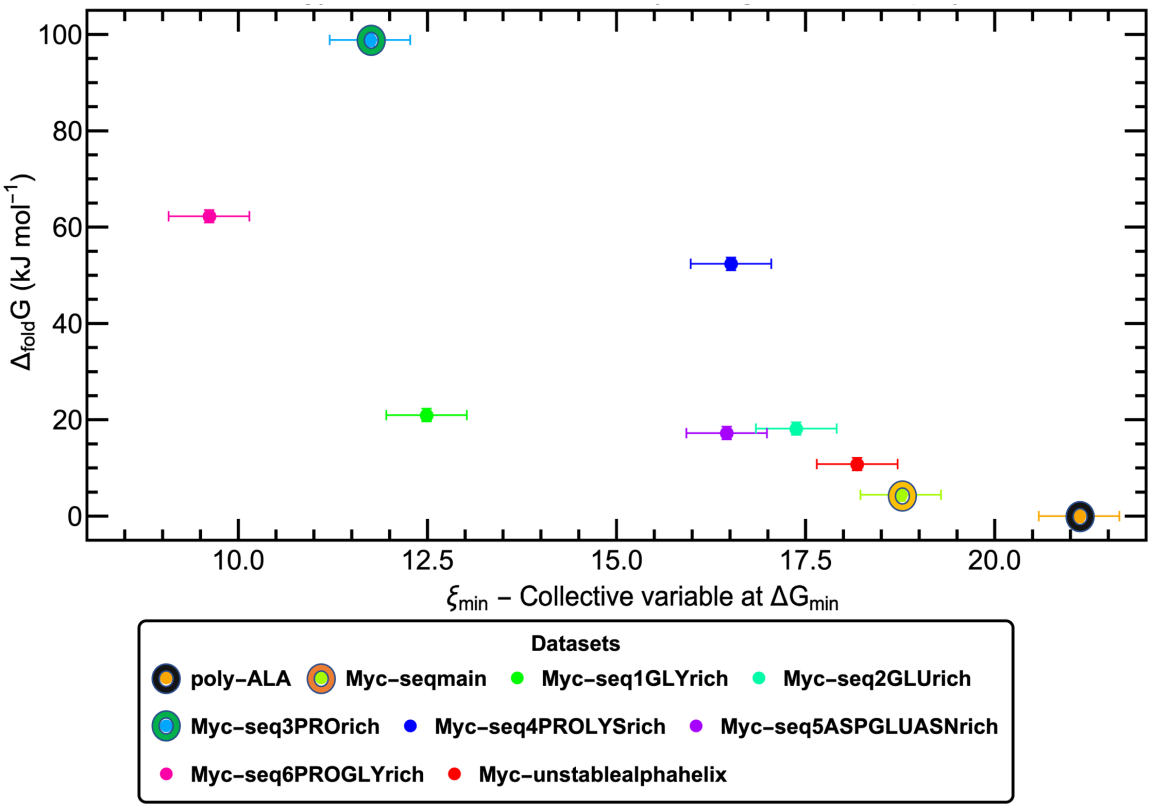

Figure 7.4: Folding free energy, $\Delta_{fold}G$ penalties with respect to polyalanine and position ξ of free energy minima structures for a collection of chosen fragments within the c-Myc sequence.

Here, in Fig. 7.4, the free energy landscape is depicted for different segments of the c-

Myc protein, including our target fragment. The free energy penalty for folding our target c-Myc fragment into an alpha-helical structure is estimated to be around 5 kJ/mol. This value is significantly lower compared to the energy barrier of 98 kJ/mol observed for a segment rich in proline residues, which are known to destabilize alpha-helices. The relatively low folding free energy penalty for our target c-Myc fragment suggests that it is energetically favorable for this segment to adopt an α-helical conformation. This characteristic makes it a promising candidate for drug development, as it implies that the segment could readily fold into an α-helix upon binding to a suitable ligand. This property is crucial for the binding-induced folding mechanism, where the interaction with a drug molecule triggers the folding of the target protein segment into a specific structure, in this case, an alpha-helix. Thus, the favorable folding energetics of our target c-Myc fragment highlights its potential as a drug target sequence for therapeutic interventions aimed at modulating the function of the c-Myc protein in cancer and other diseases.

Following our investigation of the target c-Myc fragment, we aimed to examine the impact of alanine point mutations on the stability of the alpha-helix. Given that alanine is known to stabilize alpha-helices, we hypothesized that substituting residues in a sequence with alanines would reduce the folding free energy penalty. To test this hypothesis, we introduced alanine mutations into sequences rich in glycine, proline, and their intermediates, which are known to destabilize alpha-helices. The results of these mutations are presented in Fig. 7.5. We anticipated that replacing destabilizing residues with alanines would lead to a decrease in the folding free energy penalty, thereby increasing the stability of the alpha-helical structure. The observed trend in our results aligns with our hypothesis. We noticed a significant reduction in the free energy penalty for the proline-rich sequence upon mutation to alanine. This reduction is expected because proline, with its rigid structure, impedes the formation of alpha-helices in solution by enthalpically destabilizing the helical struc-

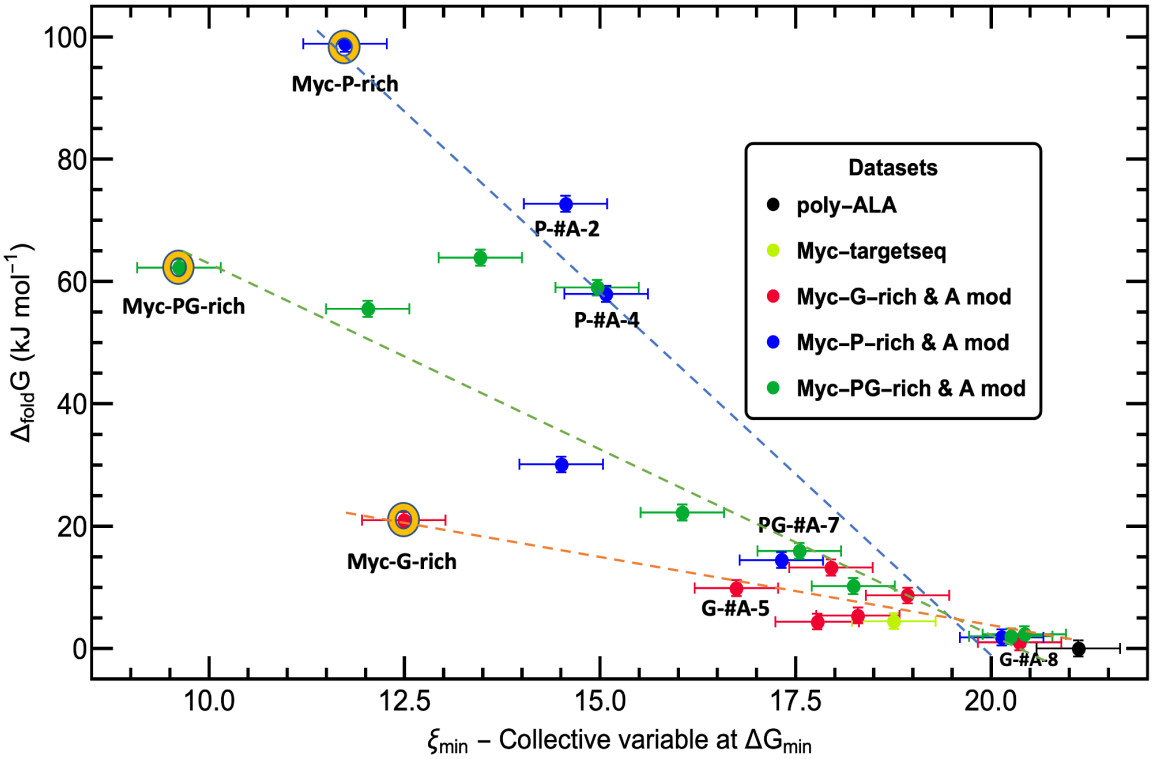

Figure 7.5: The scatter plot comparing the potential of mean force (PMF) features for polyalanine and the selected peptide sequences. A reduction in the folding free energy penalty, denoted as $\Delta_{fold}G$, is observed with the gradual point mutation of destabilizing residues in each sequence to alanines. This trend highlights the stabilizing effect of alanine on the alpha-helical structure, as the replacement of destabilizing residues with alanines leads to a decrease in the energy barrier for folding into an alpha-helix.

ture. On the other hand, the glycine-rich sequence, which is entropically unfavorable for alpha-helix formation, exhibited a smaller decrease in the folding free energy penalty upon mutation to alanine. This is due to the intrinsic flexibility of glycine, which contributes to a higher entropy in the unfolded state. The sequence rich in both proline and glycine demonstrated an intermediate slope in the reduction of the folding free energy penalty upon mutation to alanine. This intermediate behavior reflects the combined effects of proline's enthalpic destabilization and glycine's entropic contribution to the overall stability of the α-helix. To establish a general trend for the folding free energies, we conducted umbrella sampling on sequences of 37 residues in length for every naturally occurring poly-amino acid. Our results indicate that leucine, serine, and threonine have a high propensity for forming

alpha-helical structures. In contrast, isoleucine, valine, tyrosine, and histidine demonstrate significant deviations from the alpha-helical form, primarily due to steric hindrance. The positively charged amino acids, lysine and arginine, exhibit low and comparable free energy penalties for alpha-helix formation. Among the negatively charged amino acids, glutamate shows a lower folding free energy penalty ($\Delta_{fold}G$) compared to aspartate, which can be attributed to the reduced steric effects resulting from glutamate's longer side chain. These trends can be observed in Fig. 7.6.

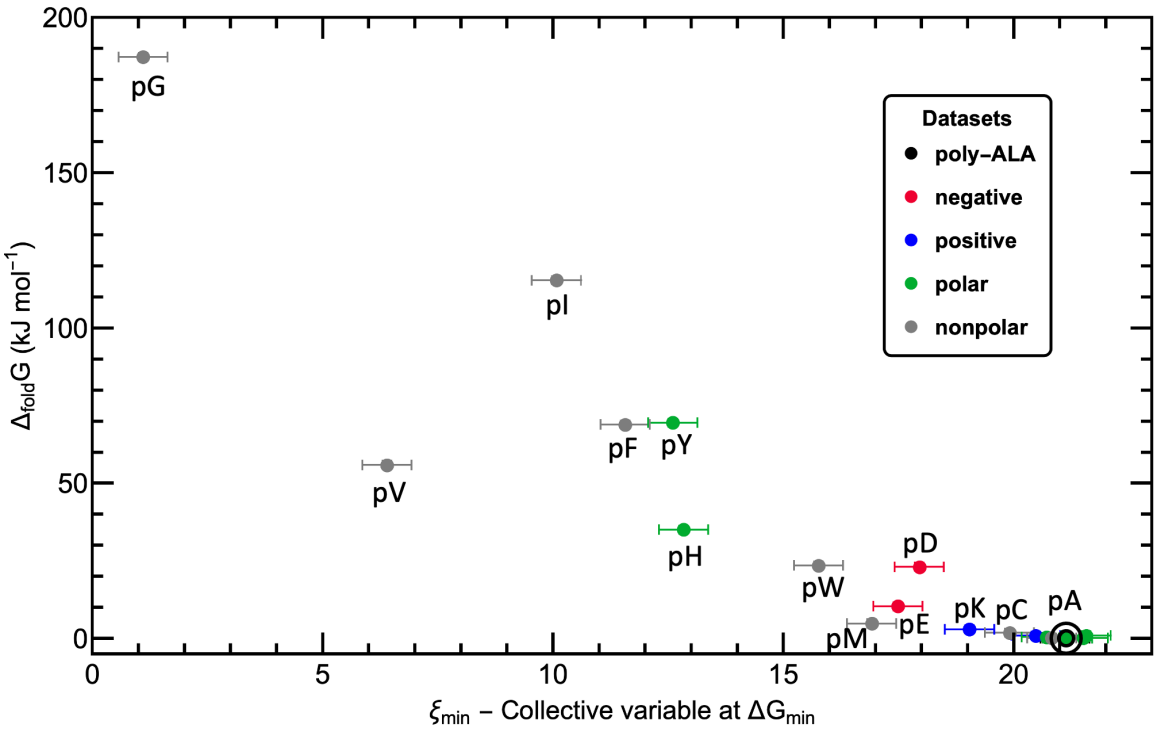

Figure 7.6: The scatter plot comparing the free energy penalties ($\Delta_{fold}G$) and position of energy minima ξ_{min} for polyalanine and all other natural poly-amino acid sequences.

Following the validation of our protocol using the selected sequences from the c-Myc protein, we expanded our research focus to include other naturally occurring intrinsically disordered regions (IDRs) and alpha-Molecular Recognition Features (α-MoRFs). Our objective was to integrate the capability of our protocol to quantify the folding free energy penalty with the structural prediction prowess of the cutting-edge AlphaFold tool. By do-

ing so, we aimed to accurately identify or predict α-MoRFs and differentiate them from fully disordered IDRs. This approach would enhance our understanding of the structural dynamics of IDRs and their potential to transition into functionally relevant α-helical conformations upon interaction with binding partners. We selected two α-MoRFs (p53-CT (C-Terminal domain of p53) and α-c-Myb-D (α-MoRF region of the c-Myb protein. D stands for doubling the length of the single sequence for our simulation purpose)), and three fully disordered sequences (dis-c-Myb (disordered fragment taken from c-Myb protein), dis2-p53-CT(disordered segment from p53 protein), and dis-Vc-HigA2-NT (disordered segment from the N-terminal region of Vc-HigA2 protein)).

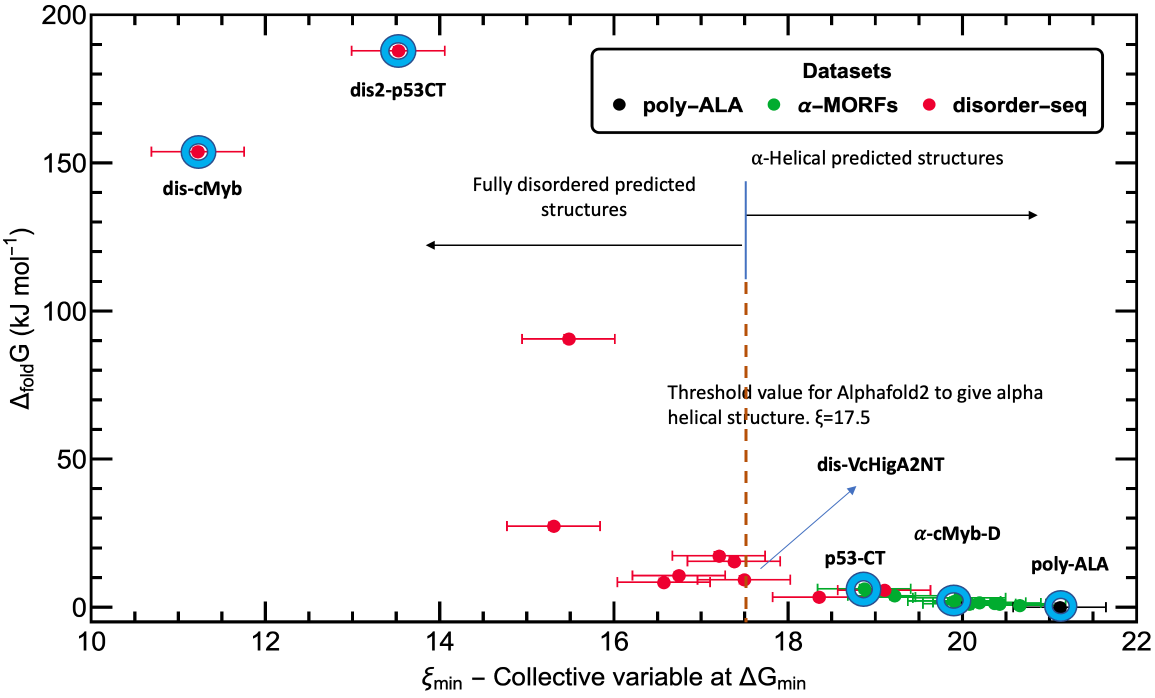

Figure 7.7: Folding free energy penalty ($\Delta_{fold}G$) comparisons between IDRs from the same proteins with opposite α-helix propensities. The threshold value of ξ for AlphaFold prediction of an α-MoRF has been shown with a dotted line.

From the Fig. 7.7, we have the green cluster of α-MoRF sequences, and red cluster of fully disordered ones. In evaluating the minimum free energy structures from our simulations against the predictions made by AlphaFold [334], we observed that the folding free

energies ($\Delta_{fold}G$) of all the α-MoRF examples we investigated fell within the range of 0-10 kJ/mol (2.4 kcal/mol). This observation is in line with the findings of Telium *et al*. [326], who reported that disordered protein complexes are, on average, less stable by approximately 2.5 kcal/mol or about 10.5 kJ/mol. This suggests that a certain amount of folding free energy might be associated with the binding-induced folding mechanism, where the binding event triggers the folding of the disordered protein into a more stable structure. This energy contribution is an important factor to consider when studying the dynamics and stability of protein complexes involving intrinsically disordered regions. Our finding also aligns with the estimates provided by Hadzi and coworkers in their research [217], where they utilized a modified Lifson-Roig (LR) model [211, 335, 336] to assess the folding free energies of intrinsically disordered proteins (IDPs). The discrepancies between our findings and those of Hadzi *et al*. could be attributed to the variations in the sequence lengths of IDPs that were considered in their analysis. The notably low folding free energy penalties associated with some of these sequences indicate their potential for use in the development of peptide-based or small-molecule drugs. Such therapeutic agents have the potential to stabilize the helical structure of α-MoRFs by inducing folding upon binding, thereby offering a strategic approach for influencing protein-protein interactions in medical treatments. This approach could lead to the development of targeted therapies that specifically modulate the function of proteins involved in various diseases. We also identified a threshold value for the collective variable ξ that AlphaFold requires to accurately predict the structure of a protein. Specifically, for a 37-residue peptide, AlphaFold consistently predicts every sequence as fully disordered beyond a ξ value of 17.5. However, for disordered sequences (shown in red) that fall within the range of 17.5 in Fig. 7.7, AlphaFold struggles to accurately predict their structures. Additionally, we observed that AlphaFold has a threshold for the folding free energy penalty, $\Delta_{fold}G$, of 0-10 kJ/mol. Consequently, all sequences with a folding free energy penalty within this range were predicted by AlphaFold to adopt

an alpha-helical conformation. This finding is in alignment with the research conducted by Hadzi and colleagues [217], further validating the accuracy of AlphaFold's predictions within these specific parameters.

In the concluding section of our study, where we focused on calculating folding free energies as depicted in Fig. 7.8, we conducted a detailed comparison. This comparison involved analyzing the final structures obtained from a 100 ns atomistic classical molecular dynamics (MD) simulation for five selected sequences. These sequences comprised two α-MoRFs and three disordered segments. We then juxtaposed these structures with

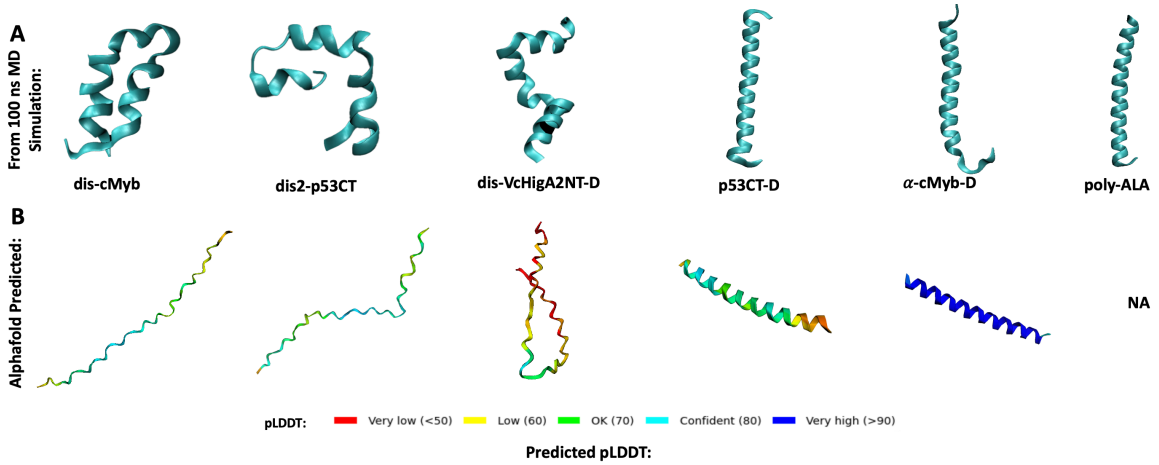

Figure 7.8: Structure comparison between 100 ns vanilla MD and AlphaFold predictions. (A) Conformations after 100 ns MD simulation starting from α-helical form. (B) Structure prediction by AlphaFold. Here, the predicted local distance difference test (pLDDT) estimates the extent to which predictions will match the experimental structure.

the ones predicted by AlphaFold for the same sequences. This comparison allowed us to evaluate the accuracy of AlphaFold's predictions in relation to the dynamic behavior and structural evolution observed in our MD simulations, providing insights into the predictive capabilities of AlphaFold for different types of protein sequences. In our analysis, we noted that AlphaFold is proficient in predicting the helical structure of the two α-MoRFs, particularly in the case of α-c-Myb-D. Additionally, AlphaFold is capable of accurately predicting sequences that will remain fully disordered up to a collective variable (ξ) of 17.5 for a 37

residue sequence. This threshold is exemplified by the structure of dis-VcHigA-2NT-D, a disordered protein that lies at the boundary of AlphaFold's ability to predict alpha-MoRFs. For values below this threshold, AlphaFold tends to predict that sequences will adopt an α-helical conformation. This observation underscores the effectiveness of AlphaFold in distinguishing between disordered regions and alpha-MoRFs, particularly when considering the specific threshold value of the collective variable ξ.

7.4.2 Binding Free Energy Calculations of Mcl-1–BH3 only Domain Proteins and Comparison to their Folding Free Energies

In this section, our aim was to investigate the binding free energies of protein complexes involving intrinsically disordered proteins (IDPs). Experimentally characterizing

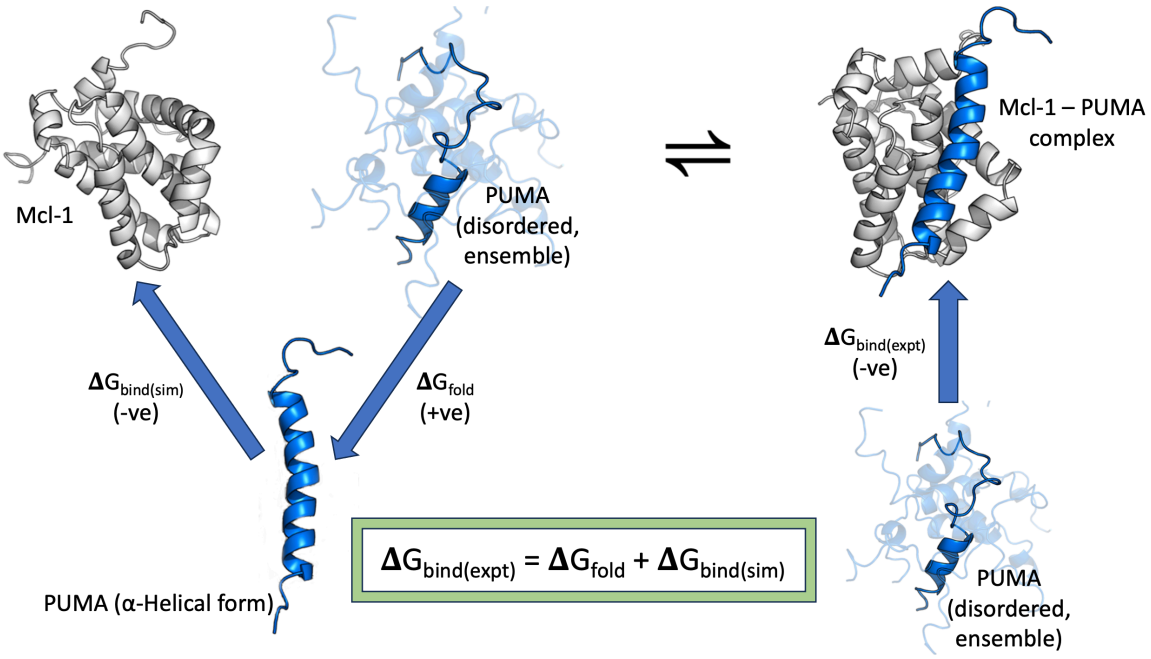

Figure 7.9: Expressing binding free energy as the sum of two processes in the binding induced folding mechanism- folding and binding free energies. Image courtesy: Rogers *et al*. [5]

the affinity of IDP target associations typically yields only the combined contributions of IDP folding and binding. Separating these two contributions is not straightforward using

conventional methods. Our goal was to compare the sum of the folding and binding free energies derived from simulations with the experimentally determined binding free energies. Through this comparison, we sought to validate our hypothesis and provide a more detailed understanding of the energetics involved in the interactions between IDPs and their binding partners.

To test our hypothesis of $\Delta_{bind}G_{(expt)} = \Delta_{fold}G_{(sim)} + \Delta_{bind}G_{(sim)}$, we conducted coarse-grained umbrella sampling simulations of the binding process of two Mcl-1-BH3 domain protein complexes. BH3 domains [382] are characteristic of the pro-apoptotic subset of Bcl-2 proteins [383], which includes proteins like BIM, BAD, BID, PUMA, and NOXA-A. Mcl-1 [218, 383–385] is a key anti-apoptotic protein that plays a critical role in cell survival and has been implicated in the development and progression of various cancers. Its ability to bind and inhibit pro-apoptotic proteins makes it a crucial regulator of the apoptotic pathway. As a result, Mcl-1 has emerged as an attractive target for cancer therapy.

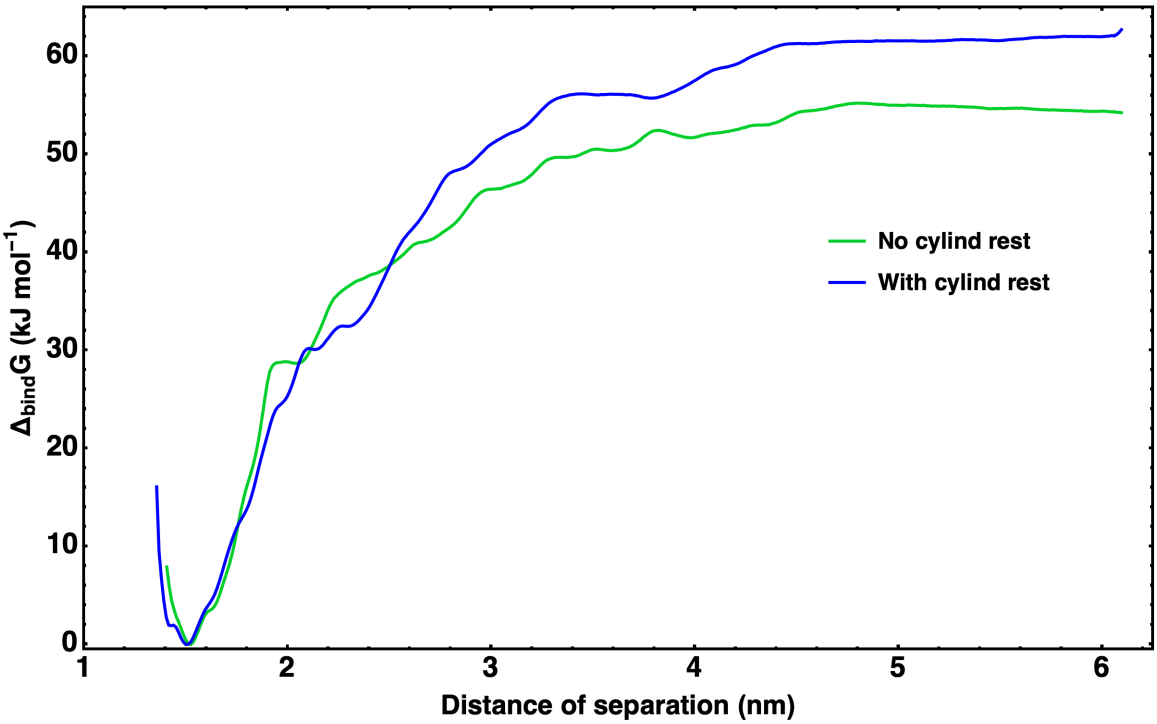

Figure 7.10: Expressing binding free energies $\Delta_{bind}G_{(sim)}$ of PUMA with and without the application of cylindrical restraints while performing umbrella sampling.

Achieving adequate sampling is a challenge in enhanced sampling techniques for processes like binding through host-guest interactions. To address this issue, cylindrical restraints can be applied along the separation path of the two binding partners. This approach restricts the ligand from exploring the entire space around the substrate in a spherical pattern, leading to faster convergence of the simulation with reduced sampling time and computational resources. However, to compare with unrestrained PMF to verify its convergence, it is important to account for the entropic term, as the ligand's access to the space around the binding substrate is restricted. Here, $\Delta_{bind}G_{(sim)} = \Delta_{bind}H_{(sim)} - T\Delta_{bind}S_{rot+vib(sim)}$, where the

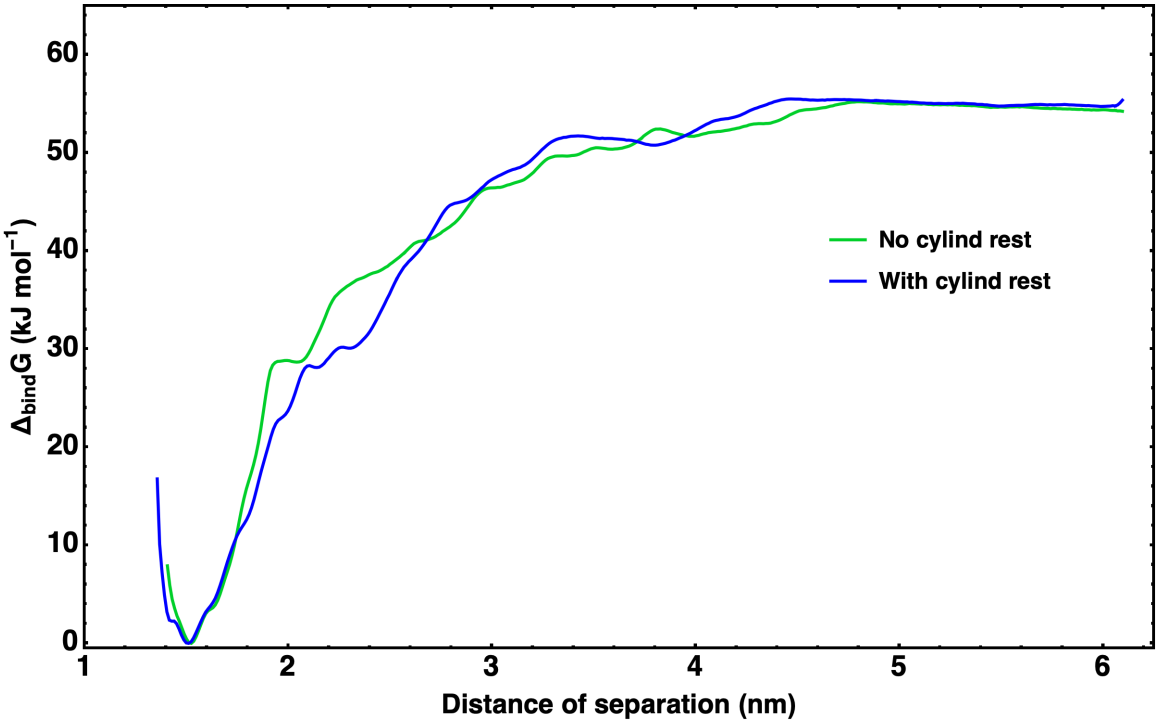

Figure 7.11: Implementation of the entropic adjustment term to the potential of mean force (PMF) that has cylindrical restraints, in order to align it with the PMF that is free of restraints.

translational entropic part ($-T\Delta S_{trans(sim)}$) is missing.

$$-T\Delta S_{(sim)} = -RT\ ln\left(\frac{Q_2}{Q_1}\right) \tag{7.5}$$

where Q_1 and Q_2 are partition functions for state 2 and 1. Here the entropic part is calculated

separately and added to the PMF with cylindrical restraints in blue (Fig. 7.10) to obtain a similar feature to the unrestrained PMF in green (Fig. 7.11)

Based on Fig. 7.11, it is apparent that the unrestrained PMF, represented in green, has undergone sufficient sampling, as it closely aligns with the PMF that has been adjusted for entropy without cylindrical restraint, indicated in blue. This congruence is highlighted by the gradual decline in entropy following a peak, which is due to the configurational entropy term from Eq. (7.5) [360, 361, 365], related to the $4\pi r^2$ increment in area explored by the unbinding ligand with increasing distance of separation, as detailed in Eq. (7.3). The adequate sampling observed in the unrestrained simulation enabled us to employ this method for the subsequent protein, NOXA-A in our analysis.

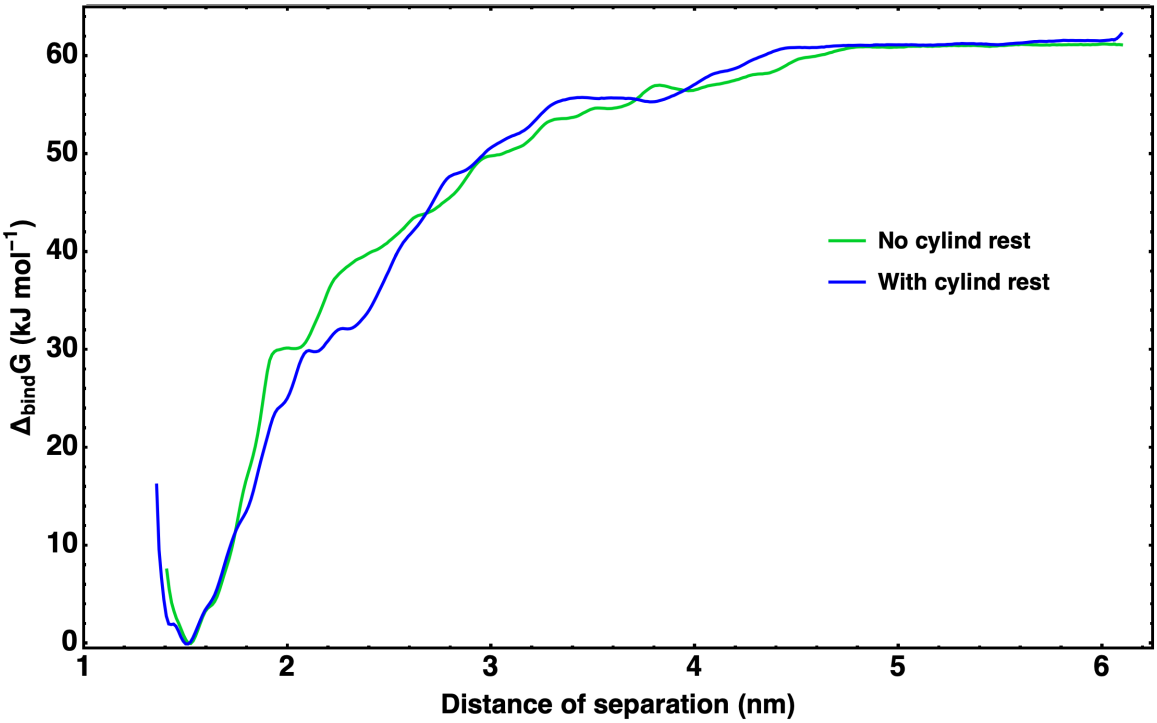

Figure 7.12: Implementation of the entropic adjustment term to the potential of mean force (PMF) that has cylindrical restraints, in order to align it with the PMF that is free of restraints.

This modification by addition of the term $k_B T ln(4\pi r^2)$ ensures that the potential of mean force (PMF) remains flat at larger separation distances by counterbalancing the entropic

decline in the PMF. This decline results from the growth in the number of configurations on a sphere of a specific radius [365]. In this approach, we utilize the flattened potential of mean force (PMF) to determine the binding free energy (ΔG) for a protein complex. This is achieved by calculating the difference in free energy between the bound and unbound states (Eq. 7.4), where the unbound state is represented by the flattened PMF. By comparing these two states, we can accurately quantify the energetics of the binding process and gain insights into the stability and affinity of the protein complex. This method provides a comprehensive understanding of the molecular interactions that drive the formation of protein complexes.

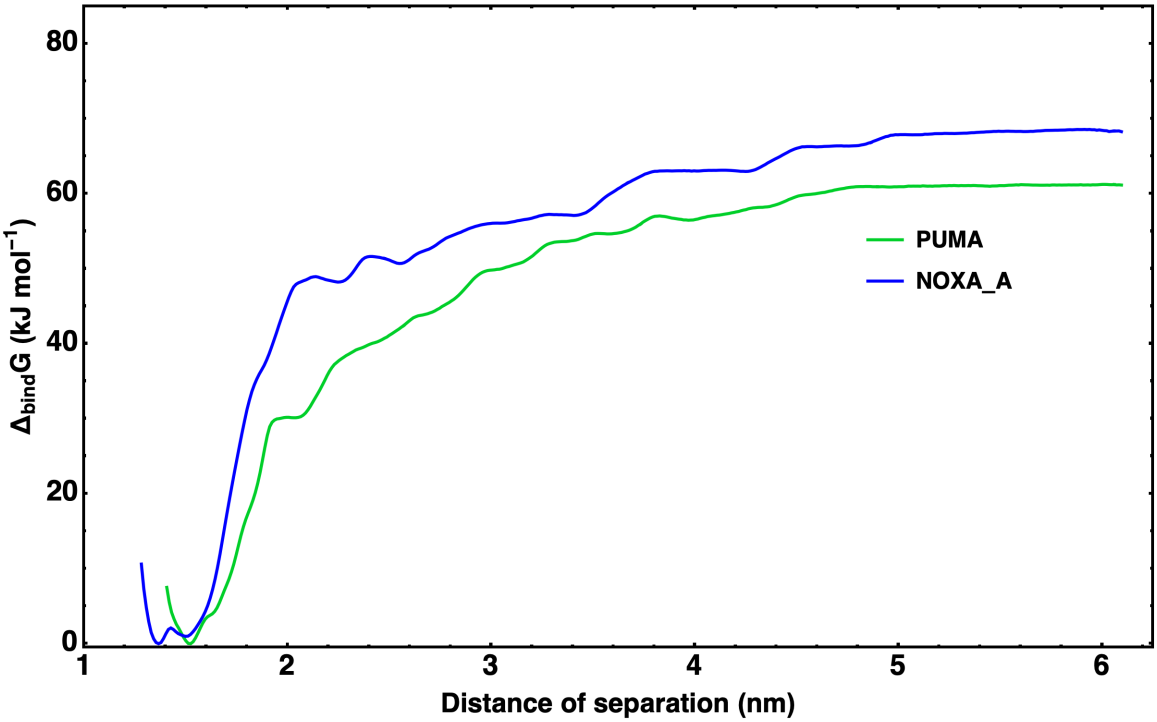

Figure 7.13: Binding free energy curves for Mcl-1 with the BH3-only domains of PUMA (green) and NOXA-A (blue)

In Fig. 7.13, the calculated binding free energies ($\Delta_{bind}G_{(sim)}$) for the protein complexes involving PUMA and NOXA-A were found to be -61.25 kJ/mol (-14.64 kcal/mol) and -68.58 kJ/mol (-16.39 kcal/mol), respectively. These values are noticeably lower than the experimental values reported by Dahal *et al*. [218], which are -55.97 kJ/mol (-13.38 kcal/mol)

for PUMA and -45.9 kJ/mol (-10.9 kcal/mol) for NOXA-A. This discrepancy suggests that there is a positive folding free energy contribution ($\Delta_{fold}G_{(sim)}$) that needs to be accounted for to reconcile the excess binding free energy obtained from simulations with the experimentally observed values. The presence of this positive $\Delta_{fold}G_{(sim)}$ indicates that the folding process upon binding contributes additional energy, which must be considered to accurately model the energetics of the protein-protein interactions in these complexes.

To calculate the folding free energies, we ran two sets of umbrella sampling simulations for PUMA and NOXA-A separately.

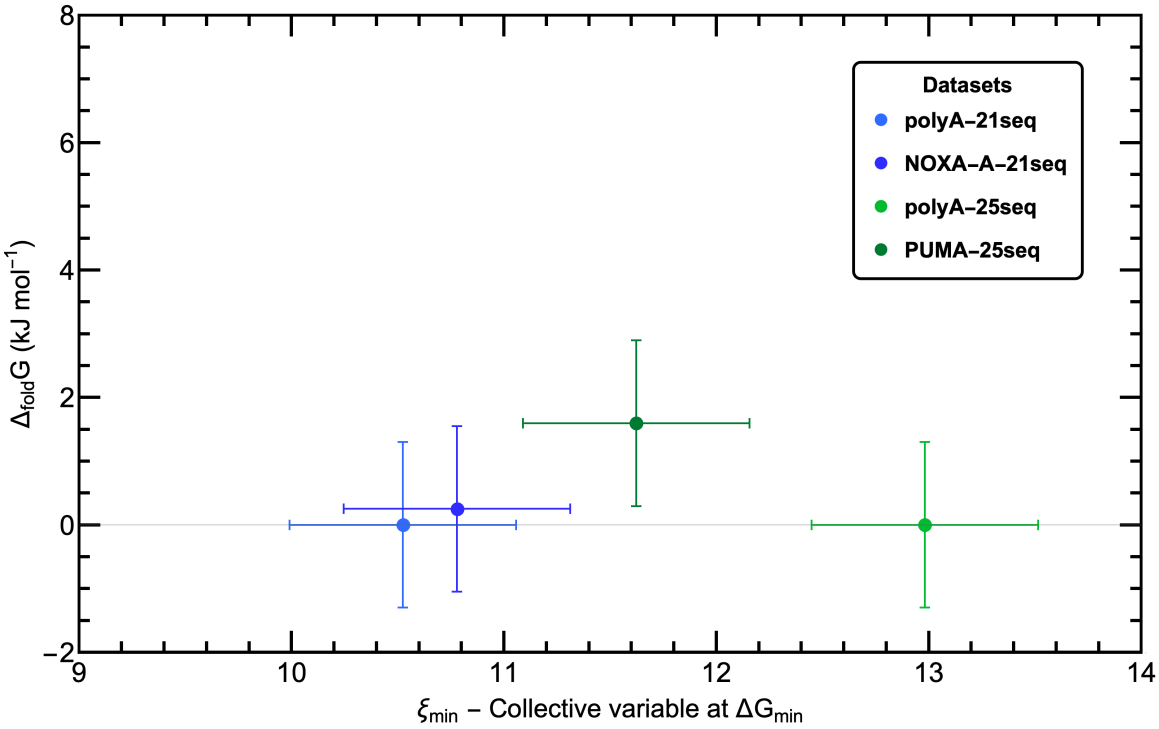

Figure 7.14: Scattered plots depicting folding free energy penalties for Mcl-1 with PUMA (green) and NOXA-A (blue)

Figure 7.14 illustrates that the folding free energy value for PUMA is 1.8 ± 1.3 kJ/mol, while for NOXA-A, it is approximately 0.5 ± 1.3 kJ/mol. These values are notably lower than the required values of 5.35 kJ/mol for PUMA and 22.68 kJ/mol for NOXA-A to match experimental observations. The binding and folding free energies for PUMA add up close

to the experimental value of -55.97 kJ/mol with an error of about 4 kJ/mol. However, for NOXA-A, this error is approximately 21 kJ/mol. Several factors could contribute to this discrepancy: i) Sequence Length: The discrepancy may arise from differences in the length of the sequences used in our simulations compared to those used in experiments. The length of the sequence can influence the folding dynamics and, consequently, the folding free energy. ii) Mutations: The use of mutated versions of the proteins in experiments, which differ from the structures in the Protein Data Bank (PDB) that we utilized, could lead to variations in the folding free energy. Mutations can significantly alter the stability and folding behavior of proteins. iii) Sequence Composition: Our results for NOXA-A might have been more accurate if we had used a longer sequence that included proline and glycine residues. These residues are known to destabilize alpha-helices and could result in a higher folding free energy penalty, bringing our simulated values closer to the experimental ones. Addressing these factors in future simulations could help to reduce the discrepancy between the simulated and experimental folding free energy values for PUMA and NOXA-A. Overall, our hypothesis that the experimental binding free energy can be represented as the sum of the calculated folding and binding free energies is partially verified. While our results provide some support for this hypothesis, additional studies involving other sequences are necessary to gather more conclusive data. Expanding the scope of our research to include a wider range of sequences will help to further validate this hypothesis and enhance our understanding of the energetics involved in protein-protein interactions.

7.5 Conclusion

In this chapter, we explored the complex thermodynamics related to protein folding and binding, focusing specifically on the intrinsically disordered regions (IDRs) and alpha-Molecular Recognition Features (α-MoRFs) present in the c-Myc protein. Additionally, we examined the interactions between Mcl-1 and BH3-only domains, which play pivotal roles in the regulation of apoptosis. Our primary aim was to quantify the folding free energy penalties that occur as proteins transition from disordered states to alpha-helical conformations. Furthermore, we sought to investigate the binding free energies of protein complexes, specifically those involving Mcl-1 in combination with PUMA and NOXA-A, which are key players in apoptotic pathways. By understanding these energetics, we aimed to shed light on the molecular mechanisms underlying protein interactions and their implications for apoptotic regulation.

Through molecular dynamics simulations, and umbrella sampling method, we first validated our protocol by analyzing sequences from the c-Myc protein, noting that the folding free energies fell within the anticipated range. This initial validation laid the groundwork for an extensive examination of naturally occurring intrinsically disordered regions (IDRs) and alpha-Molecular Recognition Features (α-MoRFs). In this broader analysis, we harnessed the capabilities of our protocol in conjunction with the structure prediction prowess of AlphaFold. Our results demonstrated that AlphaFold was adept at predicting the alpha-helical structures of α-MoRFs, particularly in the case of alpha-c-Myb-D, and could classify sequences as fully disordered up to a specified threshold.

The computation of binding free energies unveiled discrepancies when compared to experimental values, suggesting that a positive folding free energy component might be involved in the mechanism of binding induces folding. This notion was further corroborated by our investigations into protein complexes involving PUMA and NOXA-A, where

the calculated folding free energies fell significantly short of the values required to match experimental data. Such disparities could stem from various factors, including differences in sequence length between our simulations and experimental studies, the use of mutated versions in experiments that differ from the structures we used in our simulations, and variations in the overall sequence composition. These factors could have a substantial impact on the folding behavior of the proteins, leading to the discrepancies observed between our computational results and experimental findings. Addressing these issues in future studies could help to align our computational predictions more closely with experimental outcomes.

In summary, our hypothesis suggesting that the experimental binding free energy can be decomposed into the sum of computed folding and binding free energies received partial confirmation. The findings emphasize the intricate nature of protein folding and binding mechanisms and underscore the necessity for additional investigations across a broader spectrum of sequences to thoroughly substantiate this hypothesis. This work enhances our comprehension of the energetic aspects of protein-protein interactions and has potential implications for the development of pharmaceuticals and the exploration of molecular-level biological phenomena.

Chapter 8

CONCLUSION

This study employs classical molecular dynamics simulations with explicit water and the 3D-2PT analysis method to calculate spatially resolved solvation free energies of intrinsically disordered protein (IDP) K-18 Tau and for a selected water in bulk water in the presence of external electric fields. IDPs' unstructured nature poses difficulty in studying their properties experimentally. Additionally, a separate line of inquiry concentrates on the in-silico quantification of folding free energies of alpha-helical intrinsically disordered proteins (alpha-MoRFs), a task that is particularly challenging due to the difficulty of direct experimental measurement.

Solvation process has profound implications across a spectrum of biomolecular phenomena. Intrinsically disordered peptides (IDPs) remain unstructured primarily due to favorable protein-water interactions and their sequence alignments. In this investigation, all-atom MD simulations were performed on the K-18 domain of the Tau protein to assess the impact of diverse force field modifications aimed at enhancing the alignment between conformational ensembles of IDPs in simulations and experimental observations. Here, four distinct pairs of standard and modified force fields were explored, each designed to strengthen protein-water interactions and promote the stabilization of extended protein conformations. Initial investigations focused on analyzing conformational ensembles sampled over microseconds, followed by an in-depth examination of spatially resolved solvation free energy maps. It was found that modifications favoring extended conformations generally led to more favorable changes in protein solvation free energy compared to standard parameters. Regardless, the spatially resolved 3D-2PT analysis [83] highlighted notable differences in the effects of each modification strategy on protein-water interac-

tions, which are expected to influence the dynamics of conformational ensembles in MD simulations and the ability of proteins to interact, aggregate, or undergo phase separation. The solvation of polar and nonpolar functional groups in the protein is affected differently by each force field modification. Amber03ws-TIP4P/2005 (A03ws) and Amber99SBws-TIP4P2005 (A99ws) are documented for enhancing the solvation of nonpolar functional groups, whereas Amber99SB*-ILDN-TIP4P-D (A99*D) and CHARMM36m-TIPS3P* (C36m*) focus on the solvation of polar functional groups. These distinct preferences influence the sampled conformational ensembles, with A03ws and A99ws favoring extended states of nonpolar residues and A99*D and C36m* favoring extended states of polar residues. Unlike clear contrasts for global measures like the radius of gyration or end-to-end distance, these distinctions become more pronounced in proteins predominantly composed of polar or nonpolar amino acids.

In this second project, the influence of static and uniform electric fields on bulk liquid water's structure and entropy was meticulously explored through both classical molecular dynamics simulations using force-field models and *ab initio* molecular dynamics (AIMD). The study was further enhanced by applying an MD-based analytical technique known as 3D-2PT, which provided a detailed spatial analysis of entropy changes around a selected water molecule. Investigations were conducted using the TIP4P/2005 water model in classical MD simulations, alongside a flexible version of the same model for 3D-2PT analysis. The water's response in the bulk phase was probed under two distinct conditions—pinned and fixed—and subjected to a range of electric field strengths. The findings from these simulations were encouraging, revealing new insights into the behavior of water under the influence of electric fields.

The TIP4P/2005 water model effectively captures key structural properties of water and is computationally efficient. Its results align closely with those from more accurate ab initio simulations. However, both the TIP4P/2005 model and the ab initio simulations

underestimate the molecular entropy of bulk water compared to experimental values. This discrepancy might be due to the limitations of the two-phase thermodynamics (2PT) formalism. Additionally, the TIP4P/2005 model is less sensitive to polarization-induced changes than the BLYP+D3 model, likely due to the inclusion of electronic polarization in the latter.

The MD-based analysis using 3D-2PT yielded spatial insights into the local entropy and distribution of water around a reference molecule when subjected to electric fields. This analysis crafted 3D visualizations that captured changes in the water's structure, illustrating a field-induced strengthening of hydrogen bonds, akin to the formation of more ordered "ice-like" structures, particularly evident at field strengths around 1.0 $V \cdot nm^{-1}$ [200]. These modifications in water's structure and entropy were shown to be closely associated with shifts in water density, highlighting a shift towards more organized layering under the electric field's influence, with water molecules aligning their dipoles to form more structured arrangements. These observations underscore the electric field's role in shaping the microenvironment of water molecules, influencing their orientation and interactions.

The final chapter's focus was on the thermodynamic aspects of protein folding and interaction phenomena, particularly highlighting the unstructured segments and alpha-Molecular Recognition Features (α-MoRFs) within the c-Myc protein, which is overexpressed in various forms of cancer [370]. It delved into the regulatory interplay between the Mcl-1 protein and the BH3-only domains, essential elements in the apoptosis control mechanism. The work aimed at quantifying the energy barriers involved when proteins transform from a disordered to an ordered alpha-helical state. Additionally, the research investigated the thermodynamics of complex formations, especially between Mcl-1 and the pro-apoptotic proteins PUMA and NOXA-A. The purpose of this study was to shed light on the mechanisms underlying binding-induced folding and to test the supposition that the experimental binding free energies noted by Dahal *et al*. [218] are cumulative of both computed binding and folding free energies.

Molecular dynamics simulations complemented by the umbrella sampling method were employed to scrutinize sequences from the c-Myc protein, leading to an affirmation of the protocol's robustness, as evidenced by folding free energies aligning with predicted estimates. This fundamental validation established a foundation for a comprehensive investigation into the realm of intrinsically disordered regions (IDRs) and alpha-Molecular Recognition Features (α-MoRFs). Utilizing a combination of our refined protocol and the predictive capabilities of the AlphaFold algorithm, a deeper analysis was conducted on a different set of naturally occurring IDR sequences. Results from this comprehensive research indicated that AlphaFold is adept at predicting alpha-helical structures of α-MoRFs, especially for the α-c-Myb-D sequence, and is capable of precisely identifying sequences as entirely disordered based on a defined threshold. The results also exhibited a consistent result with some investigations [217, 326], where the value of folding free energies of all sampled alpha-MoRFs lie within a ΔG_{bind} range of 0-10.5 kJ/mol (2.5 kcal/mol)

Calculating binding free energies resulted in values that were more negative than those obtained experimentally, indicating the possible involvement of a positive folding free energy component in the binding-induced folding mechanism. This was underscored by our analysis of protein complexes involving Mcl-1 and proapoptotic BH3-only domains of PUMA and NOXA-A, which showed that the computed folding free energies were considerably lower than those necessary to align with experimental findings. This type of error could arise from numerous factors, such as variations in sequence length between our simulations and experimental studies, differences in the overall sequence composition, and the use of mutated versions in experiments that differ from the structures utilized in our simulations.

REFERENCES

[1] A. Ben-Naim. Hydrophobic interaction and structural changes in the solvent. *Biopolymers*, 14:1337–1355, 1975.

[2] Qinghua Liao. Chapter four - enhanced sampling and free energy calculations for protein simulations. In Birgit Strodel and Bogdan Barz, editors, *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*, volume 170 of *Progress in Molecular Biology and Translational Science*, pages 177–213. Academic Press, 2020.

[3] Viren Pattni. *Understanding Hydration via Solvent Thermodynamics and Vibrational Signatures*. PhD thesis, Ruhr-Universität Bochum Bochum, 2018.

[4] A.K. Soper. The radial distribution functions of water and ice from 220 to 673 k and at pressures up to 400 mpa. *Chemical Physics*, 258(2):121–137, 2000.

[5] Joseph M Rogers, Annette Steward, and Jane Clarke. Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'. *Journal of the American Chemical Society*, 135(4):1415–1422, 2013.

[6] Christine A Orengo, Annabel E Todd, and Janet M Thornton. From protein structure to function. *Current opinion in structural biology*, 9(3):374–382, 1999.

[7] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[8] James A Shapiro. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28, 2009.

[9] Christopher J Oldfield and A Keith Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry*, 83:553–584, 2014.

[10] A Keith Dunker, J David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, et al. Intrinsically disordered protein. *Journal of molecular graphics and modelling*, 19(1):26–59, 2001.

[11] Vladimir N Uversky. Intrinsically disordered proteins from a to z. *The international journal of biochemistry & cell biology*, 43(8):1090–1103, 2011.

[12] Peter Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27(10):527–533, 2002.

[13] A Keith Dunker, M Madan Babu, Elisar Barbar, Martin Blackledge, Sarah E Bondos, Zsuzsanna Dosztányi, H Jane Dyson, Julie Forman-Kay, Monika Fuxreiter, Jörg Gsponer, et al. What's in a name? why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins*, 1(1):e24157, 2013.

[14] Sarah C Bickers, Jonathan S Sayewich, and Voula Kanelis. Intrinsically disordered regions regulate the activities of atp binding cassette transporters. *BBA-Biomembranes*, page 183202, 2020.

[15] Snigdha Maiti, Bidisha Acharya, Veda Sheersh Boorla, Bharat Manna, Amit Ghosh, and Soumya De. Dynamic studies on intrinsically disordered regions of two paralogous transcription factors reveal rigid segments with important biological functions. *J. Mol. Biol.*, 431(7):1353–1369, 2019.

[16] Vladimir N Uversky. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Science*, 22(6):693–724, 2013.

[17] Peter Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, 37(12):509–516, 2012.

[18] Robin Van Der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, 114(13):6589–6631, 2014.

[19] Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. Introducing protein intrinsic disorder. *Chemical reviews*, 114(13):6561–6588, 2014.

[20] Vladimir N Uversky. What does it mean to be natively unfolded? *European journal of biochemistry*, 269(1):2–12, 2002.

[21] András Hatos, Borbála Hajdu-Soltész, Alexander M Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, Guillermo I Benítez, Martina Bevilacqua, Anastasia Chasapi, et al. Disprot: intrinsic protein disorder annotation in 2020. *Nucleic acids research*, 48(D1):D269–D276, 2020.

[22] Walter Basile, Marco Salvatore, Claudio Bassot, and Arne Elofsson. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS computational biology*, 15(7):e1007186, 2019.

[23] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, 16(1):18, 2015.

[24] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[25] Alex S Holehouse and Birthe B Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, pages 1–25, 2023.

[26] Ryan J Emenecker, Daniel Griffith, and Alex S Holehouse. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical journal*, 120(20):4312–4319, 2021.

[27] Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. New classification of intrinsic disorder in the human proteome, 2018. *bioRxiv* 446351, DOI: 10.1101/446351.

[28] Recep Colak, TaeHyung Kim, Magali Michaut, Mark Sun, Manuel Irimia, Jeremy Bellay, Chad L Myers, Benjamin J Blencowe, and Philip M Kim. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput. Biol.*, 9(4):e1003030, 2013.

[29] Vladimir N Uversky. The alphabet of intrinsic disorder: Ii. various roles of glutamic acid in ordered and intrinsically disordered proteins. *Intrinsically disordered proteins*, 1(1):e24684, 2013.

[30] Simone Kosol, Sara Contreras-Martos, Cesyen Cedeño, and Peter Tompa. Structural characterization of intrinsically disordered proteins by nmr spectroscopy. *Molecules*, 18(9):10802–10828, 2013.

[31] Christopher J Oldfield, Yugong Cheng, Marc S Cortese, Pedro Romero, Vladimir N Uversky, and A Keith Dunker. Coupled folding and binding with α-helix-forming molecular recognition elements. *Biochemistry*, 44(37):12454–12470, 2005.

[32] James M Krieger, Giuliana Fusco, Marc Lewitzky, Philip C Simister, Jan Marchant, Carlo Camilloni, Stephan M Feller, and Alfonso De Simone. Conformational recognition of an intrinsically disordered protein. *Biophys. J.*, 106(8):1771–1779, 2014.

[33] Ursula Jakob, Richard Kriwacki, and Vladimir N Uversky. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chemical reviews*, 114(13):6779–6805, 2014.

[34] Hedi Hegyi and Peter Tompa. Intrinsically disordered proteins display no preference for chaperone binding in vivo. *PLoS computational biology*, 4(3):e1000017, 2008.

[35] Patricia Santofimia-Castaño, Bruno Rizzuti, Yi Xia, Olga Abian, Ling Peng, Adrián Velázquez-Campoy, José L Neira, and Juan Iovanna. Targeting intrinsically disordered proteins involved in cancer. *Cell. Mol. Life Sci.*, pages 1–13, 2019.

[36] Do-Hyoung Kim and Kyou-Hoon Han. Transient secondary structures as general target-binding motifs in intrinsically disordered proteins. *Int. J. Mol. Sci.*, 19(11):3614, 2018.

[37] Alessandro Borgia, Madeleine B Borgia, Katrine Bugge, Vera M Kissling, Pétur O Heidarsson, Catarina B Fernandes, Andrea Sottini, Andrea Soranno, Karin J Buholzer, Daniel Nettels, et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555(7694):61–66, 2018.

[38] Vladimir N Uversky, Christopher J Oldfield, Uros Midic, Hongbo Xie, Bin Xue, Slobodan Vucetic, Lilia M Iakoucheva, Zoran Obradovic, and A Keith Dunker. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics*, 10(1):S7, 2009.

[39] Vladimir N Uversky. Mysterious oligomerization of the amyloidogenic proteins. *The FEBS journal*, 277(14):2940–2953, 2010.

[40] Vladimir N Uversky, Vrushank Dave, Lilia M Iakoucheva, Prerna Malaney, Steven J Metallo, Ravi Ramesh Pathak, and Andreas C Joerger. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.*, 114(13):6844–6879, 2014.

[41] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.*, 37:215–246, 2008.

[42] John Q Trojanowski and MY VIRGINIA. Parkinson's disease and related α-synucleinopathies are brain amyloidoses. *Ann. NY Acad. Sci.*, 991:107–110, 2003.

[43] Michel Goedert. Filamentous nerve cell inclusions in neurodegenerative diseases: tauopathies and alpha-synucleinopathies. *Philos. Trans. R. Soc. Lond. B*, 354(1386):1101–1118, 1999.

[44] Clifford P Brangwynne, Christian R Eckmann, David S Courson, Agata Rybarska, Carsten Hoege, Jöbin Gharakhani, Frank Jülicher, and Anthony A Hyman. Germline p granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324(5935):1729–1732, 2009.

[45] Anthony A Hyman, Christoph A Weber, and Frank Jülicher. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol*, 30(1):39–58, 2014.

[46] Simon Alberti, Amy Gladfelter, and Tanja Mittag. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*, 176(3):419–434, 2019.

[47] Simon Alberti and Dorothee Dormann. Liquid-liquid phase separation in disease. *Annu. Rev. Genet*, 53(1):171–194, 2019.

[48] Valéry Ozenne, Robert Schneider, Mingxi Yao, Jie-rong Huang, Loïc Salmon, Markus Zweckstetter, Malene Ringkjøbing Jensen, and Martin Blackledge. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J. Am. Chem. Soc.*, 134(36):15138–15148, 2012.

[49] Gregory-Neal W Gomes, Mickaël Krzeminski, Ashley Namini, Erik W Martin, Tanja Mittag, Teresa Head-Gordon, Julie D Forman-Kay, and Claudiu C Gradinaru. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. *J. Am. Chem. Soc.*, 142(37):15697–15710, 2020.

[50] Benjamin Schuler, Sonja Müller-Späth, Andrea Soranno, and Daniel Nettels. Application of confocal single-molecule fret to intrinsically disordered proteins. In *Intrinsically disordered protein analysis*, pages 21–45. Springer, 2012.

[51] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B*, 119(16):5113–5123, 2015.

[52] Robert B Best, Wenwei Zheng, and Jeetain Mittal. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.*, 10(11):5113–5124, 2014.

[53] Paul S Nerenberg, Brian Jo, Clare So, Ajay Tripathy, and Teresa Head-Gordon. Optimizing solute–water van der waals interactions to reproduce solvation free energies. *J. Phys. Chem. B*, 116(15):4524–4534, 2012.

[54] Joao Henriques, Carolina Cragnell, and Marie Skepo. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theory Comput.*, 11(7):3420–3431, 2015.

[55] Dong Song, Wei Wang, Wei Ye, Dingjue Ji, Ray Luo, and Hai-Feng Chen. ff14idps force field improving the conformation sampling of intrinsically disordered proteins. *Chem. Biol. Drug Des.*, 89(1):5–15, 2017.

[56] P. Ball. Water as an active constituent in cell biology. *Chem. Rev.*, 108(1):74–108, 2008.

[57] AK Soper. The radial distribution functions of water and ice from 220 to 673 k and at pressures up to 400 mpa. *Chemical Physics*, 258(2-3):121–137, 2000.

[58] Sheh-Yi Sheu, Dah-Yen Yang, HL Selzle, and EW Schlag. Energetics of hydrogen bonds in peptides. *Proceedings of the National Academy of Sciences*, 100(22):12683–12687, 2003.

[59] W.-L. Jorgensen, J. Chandrasekhar, J.-D. Madura, R.-W. Impey, and M.-L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[60] HJC Berendsen, JPM Postma, WF Van Gunsteren, and a J Hermans. Intermolecular forces. 1981.

[61] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[62] Vincent Kräutler, Wilfred F Van Gunsteren, and Philippe H Hünenberger. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry*, 22(5):501–508, 2001.

[63] S. Miyamoto and P.-A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[64] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

[65] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2020 manual, January 2020.

[66] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[67] Jing Huang and Alexander D. MacKerell. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *J. Comput. Chem.*, 34(25):2135–2145, 2013.

[68] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of chemical physics*, 123(23):234505, 2005.

[69] Kahled Toukan and Aneesur Rahman. Molecular-dynamics study of atomic motions in water. *Physical Review B*, 31(5):2643, 1985.

[70] Matej Praprotnik, Dušanka Janežič, and Janez Mavri. Temperature dependence of water vibrational spectrum: a molecular dynamics simulation study. *The Journal of Physical Chemistry A*, 108(50):11056–11062, 2004.

[71] Hiroki Nada and Jan PJM van der Eerden. An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of h 2 o. *The Journal of chemical physics*, 118(16):7401–7413, 2003.

[72] A Ben-Naim. Statistical thermodynamics for chemists and biochemists plenum, new york, 1992.[-11] jd bernal and rh fowler. *J. Chem. Phys*, 1:515, 1933.

[73] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The journal of physical chemistry B*, 119(16):5113–5123, 2015.

[74] Kenno Vanommeslaeghe and Alexander D MacKerell Jr. Automation of the charmm general force field (cgenff) i: bond perception and atom typing. *Journal of chemical information and modeling*, 52(12):3144–3154, 2012.

[75] Eric J Sorin and Vijay S Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical journal*, 88(4):2472–2493, 2005.

[76] Allison J DePaul, Erik J Thompson, Sarav S Patel, Kristin Haldeman, and Eric J Sorin. Equilibrium conformational dynamics in an rna tetraloop from massively parallel molecular dynamics. *Nucleic acids research*, 38(14):4856–4867, 2010.

[77] Szilárd Páll, Mark James Abraham, Carsten Kutzner, Berk Hess, and Erik Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In *International conference on exascale applications and software*, pages 3–27. Springer, 2014.

[78] Robert B Best and Jeetain Mittal. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *The journal of physical chemistry B*, 114(46):14916–14923, 2010.

[79] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of chemical physics*, 112(20):8910–8922, 2000.

[80] Seong Keun Kim, Taekjip Ha, and Jean Pierre Schermann. Water in biological systems. *Physical Chemistry Chemical Physics: PCCP*, 12(35):10145–10146, 2010.

[81] Shiang-Tai Lin, Mario Blanco, and William A Goddard III. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of lennard-jones fluids. *J. Chem. Phys.*, 119(22):11792–11805, 2003.

[82] Shiang-Tai Lin, Prabal K Maiti, and William A Goddard III. Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations. *J. Phys. Chem. B*, 114(24):8191–8198, 2010.

[83] R. A. X. Persson, V. Pattni, A. Singh, S. M. Kast, and M. Heyden. Signatures of solvation thermodynamics in spectra of intermolecular vibrations. *J. Chem. Theory Comput.*, 13(9):4467–4481, 2017.

[84] M. Chaplin. Do we underestimate the importance of water in cell biology? *Nature Rev. Mol. Cell Biol.*, 7:861–866, 2006.

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

[86] Thomas Simonson, Georgios Archontis, and Martin Karplus. Free energy simulations come of age: Protein- ligand recognition. *Acc. Chem. Res.*, 35(6):430–437, 2002.

[87] Y. Levy and J.-N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[88] M.-S. Cheung, A.-E. Garcia, and J.-N. Onuchic. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA*, 99(2):685–690, 2002.

[89] Emilio Gallicchio, Masahito Mogami Kubo, and Ronald M Levy. Entropy- enthalpy compensation in solvation and ligand binding revisited. *J. Am. Chem. Soc.*, 120(18):4526–4527, 1998.

[90] John D Chodera and David L Mobley. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Ann. Rev. Biophys.*, 42:121–142, 2013.

[91] B. Breiten, M. R. Lockett, W. Sherman, S. Fujita, M. Al-Sayah, H. Lange, C. M. Bowers, A. Heroux, G. Krilov, and G. M. Whitesides. Water networks contribute to enthalpy/entropy compensation in protein-ligand binding. *J Am Chem Soc*, 135(41):15579–15584, 2013.

[92] Eggert Rühmann, Michael Betz, Andreas Heine, and Gerhard Klebe. Fragment binding can be either more enthalpy-driven or entropy-driven: crystal structures and residual hydration patterns suggest why. *J. Med. Chem.*, 58(17):6960–6971, 2015.

[93] Jerome M Fox, Mengxia Zhao, Michael J Fink, Kyungtae Kang, and George M Whitesides. The molecular origin of enthalpy/entropy compensation in biomolecular recognition. *Ann. Rev. Biophys.*, 47:223–250, 2018.

[94] Adam Biela, Michael Betz, Andreas Heine, and Gerhard Klebe. Water makes the difference: Rearrangement of water solvation layer triggers non-additivity of functional group contributions in protein–ligand binding. *ChemMedChem*, 7(8):1423–1434, 2012.

[95] Niharendu Choudhury and B Montgomery Pettitt. Dynamics of water trapped between hydrophobic solutes. *The Journal of Physical Chemistry B*, 109(13):6422–6429, 2005.

[96] Lawrence R Pratt and Andrew Pohorille. Hydrophobic effects and modeling of biophysical aqueous solution interfaces. *Chemical Reviews*, 102(8):2671–2692, 2002.

[97] Charles Tanford. *The hydrophobic effect: formation of micelles and biological membranes 2d ed*. J. Wiley., 1980.

[98] Ken A Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[99] Dor Ben-Amotz. Water-mediated hydrophobic interactions. *Annu. Rev. Phys. Chem.*, 67:617–638, 2016.

[100] Berni Julian Alder and Thomas Everett Wainwright. Phase transition for a hard sphere system. *The Journal of chemical physics*, 27(5):1208–1209, 1957.

[101] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *nature*, 267(5612):585–590, 1977.

[102] David E Shaw, Martin M Deneroff, Ron O Dror, Jeffrey S Kuskin, Richard H Larson, John K Salmon, Cliff Young, Brannon Batson, Kevin J Bowers, Jack C Chao, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[103] Levi CT Pierce, Romelia Salomon-Ferrer, Cesar Augusto F. de Oliveira, J Andrew McCammon, and Ross C Walker. Routine access to millisecond time scale events with accelerated molecular dynamics. *Journal of chemical theory and computation*, 8(9):2997–3002, 2012.

[104] David E Shaw. Millisecond-long molecular dynamics simulations of proteins on a special-purpose machine. *Biophysical Journal*, 104(2):45a, 2013.

[105] Frank Noé. Beating the millisecond barrier in molecular dynamics simulations. *Biophysical journal*, 108(2):228–229, 2015.

[106] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[107] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics*, 97(3):1990–2001, 1992.

[108] David E Shaw, Peter J Adams, Asaph Azaria, Joseph A Bank, Brannon Batson, Alistair Bell, Michael Bergdorf, Jhanvi Bhatt, J Adam Butts, Timothy Correia, et al. Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2021.

[109] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[110] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19):8577–8593, 1995.

[111] Max Born and W Heisenberg. Zur quantentheorie der molekeln. *Original Scientific Papers Wissenschaftliche Originalarbeiten*, pages 216–246, 1985.

[112] Max Born and Robert Oppenheimer. On the quantum theory of molecules. In *Quantum Chemistry: Classic Scientific Papers*, pages 1–24. World Scientific, 2000.

[113] Uri Dinur and Arnold T Hagler. New approaches to empirical force fields. *Reviews in computational chemistry*, pages 99–164, 1991.

[114] Alexander D MacKerell Jr. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, 25(13):1584–1604, 2004.

[115] William L Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences*, 102(19):6665–6670, 2005.

[116] Matthias Heyden. Properties of water at terahertz frequencies and picosecond timescales. 2011.

[117] Alexander D MacKerell Jr. Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry*, 25(13):1584–1604, 2004.

[118] MA González. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, 12:169–200, 2011.

[119] Adri CT Van Duin, Siddharth Dasgupta, Francois Lorant, and William A Goddard. Reaxff: a reactive force field for hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, 2001.

[120] William L Jorgensen and Julian Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[121] Michael W Mahoney and William L Jorgensen. Diffusion constant of the tip5p model of liquid water. *The Journal of Chemical Physics*, 114(1):363–366, 2001.

[122] Charles L Brooks III, B Montgomery Pettitt, and Martin Karplus. Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids. *The Journal of chemical physics*, 83(11):5897–5908, 1985.

[123] Tariq A Andrea, William C Swope, and Hans C Andersen. The role of long ranged forces in determining the structure and properties of liquid water. *The Journal of chemical physics*, 79(9):4576–4584, 1983.

[124] Brad A Wells and Alan L Chaffee. Ewald summation for molecular simulations. *Journal of chemical theory and computation*, 11(8):3684–3695, 2015.

[125] Matteo Frigo and Steven G Johnson. The design and implementation of fftw3. *Proceedings of the IEEE*, 93(2):216–231, 2005.

[126] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n · log (n) method for ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[127] Dennis C Rapaport and Dennis C Rapaport Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[128] J Candy and W Rozmus. A symplectic integration algorithm for separable hamiltonian functions. *Journal of Computational Physics*, 92(1):230–256, 1991.

[129] H.-J.-C. Berendsen, J.-P.-M. Postma, W.-F. van Gunsteren, A. DiNola, and J.-R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[130] W.-G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[131] Henk Bekker, HJC Berendsen, EJ Dijkstra, S Achterop, R Vondrumen, David VANDERSPOEL, A Sijbers, H Keegstra, and MKR Renardus. Gromacs-a parallel computer for molecular-dynamics simulations. In *4th International Conference on Computational Physics (PC 92)*, pages 252–256. World Scientific Publishing, 1993.

[132] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Computer physics communications*, 91(1-3):43–56, 1995.

[133] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

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

[135] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[136] Wilfred F Van Gunsteren and Herman JC Berendsen. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023, 1990.

[137] Wilfred F Van Gunsteren, Dirk Bakowies, Riccardo Baron, Indira Chandrasekhar, Markus Christen, Xavier Daura, Peter Gee, Daan P Geerke, Alice Glättli, Philippe H Hünenberger, et al. Biomolecular modeling: goals, problems, perspectives. *Angewandte Chemie International Edition*, 45(25):4064–4092, 2006.

[138] HJC Berendsen. Simulating the physical world, a hierarchy of models for simulation, 2007.

[139] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the national academy of sciences*, 99(20):12562–12566, 2002.

[140] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2013.

[141] Yinglong Miao and J Andrew McCammon. Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. *Molecular simulation*, 42(13):1046–1055, 2016.

[142] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of chemical physics*, 151(7), 2019.

[143] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*, 23(2):187–199, 1977.

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

[145] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

[146] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020.

[147] Hui Lu and Klaus Schulten. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins: Structure, Function, and Bioinformatics*, 35(4):453–463, 1999.

[148] Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *The Journal of chemical physics*, 128(14), 2008.

[149] Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, 52(3):2893, 1995.

[150] E Weinan, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002.

[151] Luca Maragliano and Eric Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical physics letters*, 426(1-3):168–175, 2006.

[152] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[153] Xiongwu Wu and Shaomeng Wang. Self-guided molecular dynamics simulation for efficient conformational search. *The Journal of Physical Chemistry B*, 102(37):7238–7250, 1998.

[154] Xiongwu Wu, Bernard R Brooks, and Eric Vanden-Eijnden. Self-guided l angevin dynamics via generalized l angevin equation. *Journal of computational chemistry*, 37(6):595–601, 2016.

[155] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, 120(24):11919–11929, 2004.

[156] Danial Sabri Dashti and Adrian E Roitberg. Optimization of umbrella sampling replica exchange molecular dynamics by replica positioning. *Journal of chemical theory and computation*, 9(11):4692–4699, 2013.

[157] Fabian T Zeller. Adaptive biasing combined with hamiltonian replica exchange to improve umbrella sampling free energy simulations. *Biophysical Journal*, 106(2):404a–405a, 2014.

[158] Stefano Piana and Alessandro Laio. A bias-exchange approach to protein folding. *The journal of physical chemistry B*, 111(17):4553–4559, 2007.

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

[160] Marc Souaille and Benoıt Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer physics communications*, 135(1):40–57, 2001.

[161] John D Chodera, William C Swope, Jed W Pitera, Chaok Seok, and Ken A Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *Journal of Chemical Theory and Computation*, 3(1):26–41, 2007.

[162] Themis Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 1. theory. *The Journal of Physical Chemistry B*, 102(18):3531–3541, 1998.

[163] Themis Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids. *The Journal of Physical Chemistry B*, 102(18):3542–3550, 1998.

[164] Crystal N Nguyen, Tom Kurtzman Young, and Michael K Gilson. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril. *The Journal of chemical physics*, 137(4), 2012.

[165] Crystal N Nguyen, Anthony Cruz, Michael K Gilson, and Tom Kurtzman. Thermodynamics of water in an enzyme active site: grid-based hydration analysis of coagulation factor xa. *Journal of chemical theory and computation*, 10(7):2769–2780, 2014.

[166] Crystal N Nguyen, Tom Kurtzman, and Michael K Gilson. Spatial decomposition of translational water–water correlation entropy in binding pockets. *Journal of chemical theory and computation*, 12(1):414–429, 2016.

[167] Rasmus AX Persson, Viren Pattni, Anurag Singh, Stefan M Kast, and Matthias Heyden. Signatures of solvation thermodynamics in spectra of intermolecular vibrations. *J. Chem. Theory Comput.*, 13(9):4467–4481, 2017.

[168] Matthias Heyden. Disassembling solvation free energies into local contributions—toward a microscopic understanding of solvation processes. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(2):e1390, 2019.

[169] Richard H Henchman. Partition function for a simple liquid using cell theory parametrized by computer simulation. *The Journal of chemical physics*, 119(1):400–406, 2003.

[170] Richard H Henchman. Free energy of liquid water from a computer simulation via cell theory. *The Journal of chemical physics*, 126(6), 2007.

[171] Georgios Gerogiokas, Gaetano Calabro, Richard H Henchman, Michelle WY Southey, Richard J Law, and Julien Michel. Prediction of small molecule hydration thermodynamics with grid cell theory. *Journal of chemical theory and computation*, 10(1):35–48, 2014.

[172] Jonathan Higham, Szu-Yu Chou, Frauke Gräter, and Richard H Henchman. Entropy of flexible liquids from hierarchical force–torque covariance and coordination. *Molecular Physics*, 116(15-16):1965–1976, 2018.

[173] Hafiz Saqib Ali, Jonathan Higham, and Richard H Henchman. Entropy of simulated liquids using multiscale cell correlation. *Entropy*, 21(8):750, 2019.

[174] Arghya Chakravorty, Jonathan Higham, and Richard H Henchman. Entropy of proteins using multiscale cell correlation. *Journal of Chemical Information and Modeling*, 60(11):5540–5551, 2020.

[175] S. Chakraborty, S.-K. Sinha, and S. Bandyopadhyay. Low-frequency vibrational spectrum of water in the hydration layer of a protein: A molecular dynamics simulation study. *J. Phys. Chem. B*, 111(48):13626–13631, 2007.

[176] Peter H Berens, Donald HJ Mackay, Gary M White, and Kent R Wilson. Thermodynamics and quantum corrections from molecular dynamics for liquid water. *The Journal of chemical physics*, 79(5):2375–2389, 1983.

[177] Shiang-Tai Lin, Mario Blanco, and William A. Goddard III. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *J. Chem. Phys.*, 119:11792, 2003.

[178] Valeria Conti Nibali, Sthitadhi Maiti, Franz Saija, Matthias Heyden, and Giuseppe Cassone. Electric-field induced entropic effects in liquid water. *The Journal of Chemical Physics*, 158(18), 2023.

[179] M. Heyden and M. Havenith. Combining thz spectroscopy and md simulations to study protein-hydration coupling. *Methods*, 52(1):74–83, 2010.

[180] DA McQuarrie. Statistical mechanics, harper & row, new york, 1976.

[181] Norman F. Carnahan and Kenneth E. Starling. Thermodynamic plroperties of a rigid□sphere fluid. *J. Chem. Phys.*, 53(2):600–603, 1970.

[182] TP Straatsma and JA McCammon. Multiconfiguration thermodynamic integration. *The Journal of chemical physics*, 95(2):1175–1188, 1991.

[183] Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. New classification of intrinsic disorder in the human proteome. *BioRxiv*, page 446351, 2018.

[184] Eszter Hazy and Peter Tompa. Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *ChemPhysChem*, 10(9-10):1415–1419, 2009.

[185] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[186] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003.

[187] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65(3):712–725, 2006.

[188] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys.*, 123(23):234505, 2005.

[189] Robert B Best and Jeetain Mittal. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B*, 114(46):14916–14923, 2010.

[190] Robert B Best, David de Sancho, and Jeetain Mittal. Residue-specific α-helix propensities from molecular simulation. *Biophys. J.*, 102(6):1462–1467, 2012.

[191] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell Jr. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71, 2017.

[192] Herman-J-C Berendsen, J-Raúl Grigera, and Tjerk P Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[193] Paul Robustelli, Stefano Piana, and David E Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA*, 115(21):E4758–E4766, 2018.

[194] Stefano Piana, Paul Robustelli, Dazhi Tan, Songela Chen, and David E. Shaw. Development of a force field for the simulation of single-chain proteins and protein–protein complexes. *J. Chem. Theory Comput.*, 16(4):2494–2507, 2020.

[195] Paul Robustelli, Stefano Piana, and David E. Shaw. Mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *J. Am. Chem. Soc.*, 142(25):11092–11101, 2020.

[196] Saeed Izadi, Ramu Anandakrishnan, and Alexey V Onufriev. Building water models: a different approach. *J. Phys. Chem. Lett.*, 5(21):3863–3871, 2014.

[197] Parviz Seifpanahi Shabane, Saeed Izadi, and Alexey V Onufriev. General purpose water model can improve atomistic simulations of intrinsically disordered proteins. *J. Chem. Theory Comput.*, 15(4):2620–2634, 2019.

[198] Ped^3: Protein ensemble database. `http://pedb.vib.be/`. Accessed: Aug 01, 2019.

[199] Matthias Heyden. 3d-2pt, 2022. DOI: 10.5281/zenodo.7311555.

[200] Giuseppe Cassone, Jiri Sponer, Sebastiano Trusso, and Franz Saija. Ab initio spectroscopy of water under electric fields. *Phys. Chem. Chem. Phys.*, 21:21205–21212, 2019.

[201] Zdenek Futera and Niall J. English. Communication: Influence of external static and alternating electric fields on water from long-time non-equilibrium ab initio molecular dynamics. *The Journal of Chemical Physics*, 147(3):031102, 2017.

[202] Surajit Kalita, Hagai Bergman, Kshatresh Dutta Dubey, and Sason Shaik. How can static and oscillating electric fields serve in decomposing alzheimer's and other senile plaques? *Journal of the American Chemical Society*, 145(6):3543–3553, 2023.

[203] Sima Hejazi, Hassan Pahlavanzadeh, and Janet A. W. Elliott. Thermodynamic investigation of the effect of electric field on solid–liquid equilibrium. *The Journal of Physical Chemistry B*, 125(4):1271–1281, 2021. PMID: 33497220.

[204] Ali Estejab, Ricardo A. García Cárcamo, and Rachel B. Getman. Influence of an electrified interface on the entropy and energy of solvation of methanol oxidation intermediates on platinum(111) under explicit solvation. *Phys. Chem. Chem. Phys.*, 24:4251–4261, 2022.

[205] Aljaž Godec and Franci Merzel. Physical origin underlying the entropy loss upon hydrophobic hydration. *Journal of the American Chemical Society*, 134(42):17574–17581, 2012. PMID: 23003674.

[206] David Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647, Sep 2005.

[207] Shekhar Garde and Amish J. Patel. Unraveling the hydrophobic effect, one molecule at a time. *Proceedings of the National Academy of Sciences*, 108(40):16491–16492, 2011.

[208] Rasmus A. X. Persson, Viren Pattni, Anurag Singh, Stefan M. Kast, and Matthias Heyden. Signatures of solvation thermodynamics in spectra of intermolecular vibrations. *Journal of Chemical Theory and Computation*, 13(9):4467–4481, 2017. PMID: 28783431.

[209] Rita Pancsa and Monika Fuxreiter. Interactions via intrinsically disordered regions: what kind of motifs? *IUBMB life*, 64(6):513–520, 2012.

[210] Ronesh Sharma, Alok Sharma, Ashwini Patil, and Tatsuhiko Tsunoda. Discovering morfs by trisecting intrinsically disordered protein sequence into terminals and middle regions. *BMC bioinformatics*, 19(13):155–163, 2019.

[211] San Hadži, Remy Loris, and Jurij Lah. The sequence–ensemble relationship in fuzzy protein complexes. *Proceedings of the National Academy of Sciences*, 118(37):e2020562118, 2021.

[212] Christopher J Oldfield, Yugong Cheng, Marc S Cortese, Celeste J Brown, Vladimir N Uversky, and A Keith Dunker. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6):1989–2000, 2005.

[213] Peter E Wright and H Jane Dyson. Linking folding and binding. *Current opinion in structural biology*, 19(1):31–38, 2009.

[214] Jing Yan, A Keith Dunker, Vladimir N Uversky, and Lukasz Kurgan. Molecular recognition features (morfs) in three domains of life. *Molecular BioSystems*, 12(3):697–710, 2016.

[215] H Jane Dyson and Peter E Wright. Role of intrinsic protein disorder in the function and interactions of the transcriptional coactivators creb-binding protein (cbp) and p300. *Journal of Biological Chemistry*, 291(13):6714–6722, 2016.

[216] Katrine Bugge, Lasse Staby, Edoardo Salladini, Rasmus G Falbe-Hansen, Birthe B Kragelund, and Karen Skriver. $\alpha\alpha$-hub domains and intrinsically disordered proteins: A decisive combo. *Journal of Biological Chemistry*, 296, 2021.

[217] San Hadži and Jurij Lah. The free energy folding penalty accompanying binding of intrinsically disordered α-helical motifs. *Protein Science*, 31(7):e4370, 2022.

[218] Liza Dahal, Tristan OC Kwan, Jeffrey J Hollins, and Jane Clarke. Promiscuous and selective: how intrinsically disordered bh3 proteins interact with their pro-survival partner mcl-1. *Journal of molecular biology*, 430(16):2468–2477, 2018.

[219] Sthitadhi Maiti and Matthias Heyden. Model-dependent solvation of the k-18 domain of the intrinsically disordered protein tau. *The Journal of Physical Chemistry B*, 127(33):7220–7230, 2023. PMID: 37556237.

[220] John C Kendrew, G Bodo, Howard M Dintzis, RG Parrish, Harold Wyckoff, and David C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[221] Helen M Berman. The protein data bank: a historical perspective. *Acta Cryst. A*, 64(1):88–95, 2008.

[222] David S Goodsell, Christine Zardecki, Luigi Di Costanzo, Jose M Duarte, Brian P Hudson, Irina Persikova, Joan Segura, Chenghua Shao, Maria Voigt, John D Westbrook, et al. Rcsb protein data bank: Enabling biomedical research and drug discovery. *Protein Sci.*, 29(1):52–65, 2020.

[223] David R Armstrong, John M Berrisford, Matthew J Conroy, Aleksandras Gutmanas, Stephen Anyango, Preeti Choudhary, Alice R Clark, Jose M Dana, Mandar Deshpande, Roisin Dunlop, et al. Pdbe: improved findability of macromolecular structure data in the pdb. *Nucleic Acids Res.*, 48(D1):D335–D343, 2020.

[224] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[225] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[226] Ivan Anishchenko, Minkyung Baek, Hahnbeom Park, Naozumi Hiranuma, David E Kim, Justas Dauparas, Sanaa Mansoor, Ian R Humphreys, and David Baker. Protein tertiary structure prediction and refinement using deep learning and rosetta in casp14. *Proteins: Struct., Funct., Bioinf.*, 89(12):1722–1733, 2021.

[227] M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, 21(3):432–440, 2011.

[228] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[229] Hans W Horn, William C Swope, Jed W Pitera, Jeffry D Madura, Thomas J Dick, Greg L Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *J. Chem. Phys.*, 120(20):9665–9678, 2004.

[230] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Struct., Funct., Bioinf.*, 78(8):1950–1958, 2010.

[231] Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía B Chemes, Javier A Iserte, Nicolás A Méndez, Nicolás A Garrone, Tadeo E Saldaño, Julia Marchetti, Ana Julia Velez Rueda, Pau Bernadó, Martin Blackledge, Tiago N Cordeiro, Eric Fagerberg, Julie D Forman-Kay, Maria S Fornasari, Toby J Gibson, Gregory-Neal W Gomes, Claudiu C Gradinaru, Teresa Head-Gordon, Malene Ringkjøbing Jensen, Edward A Lemke, Sonia Longhi, Cristina Marino-Buslje, Giovanni Minervini, Tanja Mittag, Alexander Miguel Monzon, Rohit V Pappu, Gustavo Parisi, Sylvie Ricard-Blum, Kiersten M Ruff, Edoardo Salladini, Marie Skepö, Dmitri Svergun, Sylvain D Vallet, Mihaly Varadi, Peter Tompa, Silvio C E Tosatto, and Damiano Piovesan. Ped in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, 49(D1):D404–D411, 2021.

[232] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack Jr. Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Struct., Funct., Bioinf.*, 77(4):778–795, 2009.

[233] Samuel Wohl, Matthew Jakubowski, and Wenwei Zheng. Salt-dependent conformational changes of intrinsically disordered proteins. *J. Phys. Chem. Lett.*, 12(28):6684–6691, 2021.

[234] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[235] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[236] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52(2):255–268, 1984.

[237] Viren Pattni and Matthias Heyden. Pressure effects on protein hydration water thermodynamics. *J. Phys. Chem. B*, 123(28):6014–6022, 2019.

[238] V. Pattni, T. Vasilevskaya, W. Thiel, and M. Heyden. Distinct protein hydration water species defined by spatially resolved spectra of intermolecular vibrations. *J. Phys. Chem. B*, 121(31):7431–7442, 2017.

[239] Matthias Heyden. Heterogeneity of water structure and dynamics at the protein-water interface. *J. Chem. Phys.*, 150(9):094701, 2019.

[240] Tawny N Fajardo and Matthias Heyden. Dissecting the conformational free energy of a small peptide in solution. *J. Phys. Chem. B*, 125(18):4634–4644, 2021.

[241] Morteza M Waskasi, Aleksandar Lazaric, and Matthias Heyden. Solvent-mediated forces in protein dielectrophoresis. *Electrophoresis*, 42(20):2060–2069, 2021.

[242] Lukas S Stelzl, Lisa M Pietrek, Andrea Holla, Javier Oroz, Mateusz Sikora, Jürgen Köfinger, Benjamin Schuler, Markus Zweckstetter, and Gerhard Hummer. Global structure of the intrinsically disordered protein tau emerges from its local structure. *JACS Au*, 2(3):673–686, 2022.

[243] Marco D Mukrasch, Jacek Biernat, Martin von Bergen, Christian Griesinger, Eckhard Mandelkow, and Markus Zweckstetter. Sites of tau important for aggregation populate β-structure and bind to microtubules and polyanions. *J. Biol. Chem.*, 280(26):24978–24986, 2005.

[244] F. Hofmeister. *Zur Lehre von der Wirkung der Salze*, volume 25, pages 1–30. Springer–Verlag: Berlin, New York, 1888.

[245] Benoit Roux, Hsiang Ai Yu, and Martin Karplus. Molecular basis for the born model of ion solvation. *J. Phys. Chem.*, 94(11):4683–4688, 1990.

[246] Johan Åqvist and Tomas Hansson. On the validity of electrostatic linear response in polar solvents. *J. Phys. Chem.*, 100(22):9512–9521, 1996.

[247] Jacob I Monroe, Sally Jiao, R Justin Davis, Dennis Robinson Brown, Lynn E Katz, and M Scott Shell. Affinity of small-molecule solutes to hydrophobic, hydrophilic, and chemically patterned interfaces in aqueous solution. *Proc. Natl. Acad. Sci. USA*, 118(1):e2020205118, 2021.

[248] R. M. Lynden-Bell and J. C. Rasaiah. Mobility and solvation of ions in channels. *J. Chem. Phys.*, 105:9266–9280, 1996.

[249] Garegin A Papoian, Johan Ulander, and Peter G Wolynes. Role of water mediated interactions in protein- protein recognition landscapes. *J. Am. Chem. Soc.*, 125(30):9170–9178, 2003.

[250] Susmitha Ambadipudi, Jacek Biernat, Dietmar Riedel, Eckhard Mandelkow, and Markus Zweckstetter. Liquid–liquid phase separation of the microtubule-binding repeats of the alzheimer-related protein tau. *Nat. Commun.*, 8(1):275, 2017.

[251] Stephen D. Fried and Steven G. Boxer. Measuring electric fields and noncovalent interactions using the vibrational stark effect. *Accounts of Chemical Research*, 48(4):998–1006, 2015. PMID: 25799082.

[252] M. Gavish, J.-L. Wang, M. Eisenstein, M. Lahav, and L. Leiserowitz. The role of crystal polarity in α-amino acid crystals for induced nucleation of ice. *Science*, 256(5058):815–818, 1992.

[253] Eun-Mi Choi, Young-Hwan Yoon, Sangyoub Lee, and Heon Kang. Freezing transition of interfacial water at room temperature under electric fields. *Phys. Rev. Lett.*, 95:085701, Aug 2005.

[254] I. Braslavsky and S. G. Lipson. Electrofreezing effect and nucleation of ice crystals in free growth experiments. *Applied Physics Letters*, 72(2):264–266, 1998.

[255] Weiduo Zhu, Yingying Huang, Chongqin Zhu, Hong-Hui Wu, Lu Wang, Jaeil Bai, Jinlong Yang, Joseph S. Francisco, Jijun Zhao, Lan-Feng Yuan, and Xiao Cheng Zeng. Room temperature electrofreezing of water yields a missing dense ice phase in the phase diagram. *Nature Communications*, 10(1):1925, Apr 2019.

[256] Nina Balke, Stephen Jesse, Ben Carmichael, M Baris Okatan, Ivan I Kravchenko, Sergei V Kalinin, and Alexander Tselev. Quantification of in-contact probe-sample electrostatic forces with dynamic atomic force microscopy. *Nanotechnology*, 28(6):065704, jan 2017.

[257] Z. Hammadi, M. Descoins, E. Salançon, and R. Morin. Proton and light ion nanobeams from field ionization of water. *Applied Physics Letters*, 101(24):243110, 2012.

[258] Eric M. Stuve. Ionization of water in interfacial electric fields: An electrochemical view. *Chemical Physics Letters*, 519-520:1–17, 2012.

[259] Sason Shaik, Debasish Mandal, and Rajeev Ramanan. Oriented electric fields as future smart reagents in chemistry. *Nature Chemistry*, 8(12):1091–1098, Dec 2016.

[260] Albert C. Aragonès, Naomi L. Haworth, Nadim Darwish, Simone Ciampi, Nathaniel J. Bloomfield, Gordon G. Wallace, Ismael Diez-Perez, and Michelle L. Coote. Electrostatic catalysis of a diels–alder reaction. *Nature*, 531(7592):88–91, Mar 2016.

[261] Xiaoyan Huang, Chun Tang, Jieqiong Li, Li-Chuan Chen, Jueting Zheng, Pei Zhang, Jiabo Le, Ruihao Li, Xiaohui Li, Junyang Liu, Yang Yang, Jia Shi, Zhaobin Chen, Mindong Bai, Hao-Li Zhang, Haiping Xia, Jun Cheng, Zhong-Qun Tian, and Wenjing Hong. Electric field–induced selective catalysis of single-molecule reaction. *Science Advances*, 5(6):eaaw3072, 2019.

[262] Sason Shaik, Sam P. de Visser, and Devesh Kumar. External electric field will control the selectivity of enzymatic-like bond activations. *Journal of the American Chemical Society*, 126(37):11746–11749, 2004. PMID: 15366922.

[263] Rinat Meir, Hui Chen, Wenzhen Lai, and Sason Shaik. Oriented electric fields accelerate diels–alder reactions and control the endo/exo selectivity. *ChemPhysChem*, 11(1):301–310, 2010.

[264] Giuseppe Cassone, Fabio Pietrucci, Franz Saija, François Guyot, and A. Marco Saitta. One-step electric-field driven methane and formaldehyde synthesis from liquid methanol. *Chem. Sci.*, 8:2329–2336, 2017.

[265] Christopher A. Petroff, Giuseppe Cassone, Jiří Šponer, and Geoffrey R. Hutchison. Intrinsically polar piezoelectric self-assembled oligopeptide monolayers. *Advanced Materials*, 33(17):2007486, 2021.

[266] Antonino Marco Saitta and Franz Saija. Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences*, 111(38):13768–13773, 2014.

[267] Giuseppe Cassone, Jiri Sponer, Judit E. Sponer, Fabio Pietrucci, A. Marco Saitta, and Franz Saija. Synthesis of (d)-erythrose from glycolaldehyde aqueous solutions under electric field. *Chem. Commun.*, 54:3211–3214, 2018.

[268] Kshatresh Dutta Dubey, Thijs Stuyver, Surajit Kalita, and Sason Shaik. Solvent organization and rate regulation of a menshutkin reaction by oriented external electric fields are revealed by combined md and qm/mm calculations. *Journal of the American Chemical Society*, 142(22):9955–9965, 2020. PMID: 32369357.

[269] Miguel Saggu, Nicholas M. Levinson, and Steven G. Boxer. Experimental quantification of electrostatics in x–h···π hydrogen bonds. *Journal of the American Chemical Society*, 134(46):18986–18997, 2012. PMID: 23098379.

[270] Damien Laage, Thomas Elsaesser, and James T. Hynes. Perspective: Structure and ultrafast dynamics of biomolecular hydration shells. *Structural Dynamics*, 4(4):044018, 2017.

[271] Achintya Kundu, Fabian Dahms, Benjamin P. Fingerhut, Erik T. J. Nibbering, Ehud Pines, and Thomas Elsaesser. Hydrated excess protons in acetonitrile/water mixtures: Solvation species and ultrafast proton motions. *The Journal of Physical Chemistry Letters*, 10(9):2287–2294, 2019.

[272] A. Marco Saitta, Franz Saija, and Paolo V. Giaquinta. Ab initio molecular dynamics study of dissociation of water under an electric field. *Phys. Rev. Lett.*, 108:207801, May 2012.

[273] Giuseppe Cassone. Nuclear quantum effects largely influence molecular dissociation and proton transfer in liquid water under an electric field. *The Journal of Physical Chemistry Letters*, 11(21):8983–8988, 2020. PMID: 33035059.

[274] Arun Chattopadhyay and Steven G. Boxer. Vibrational stark effect spectroscopy. *Journal of the American Chemical Society*, 117(4):1449–1450, 1995.

[275] Sara Laporte, Fabio Finocchi, Lorenzo Paulatto, Marc Blanchard, Etienne Balan, François Guyot, and Antonino Marco Saitta. Strong electric fields at a prototypical oxide/water interface probed by ab initio molecular dynamics: Mgo(001). *Phys. Chem. Chem. Phys.*, 17:20382–20390, 2015.

[276] Zdenek Futera and Niall J. English. Water breakup at fe2o3–hematite/water interfaces: Influence of external electric fields from nonequilibrium ab initio molecular dynamics. *The Journal of Physical Chemistry Letters*, 12(29):6818–6826, 2021. PMID: 34270253.

[277] Fabrizio Creazzo and Sandra Luber. Explicit solvent effects on (110) ruthenium oxide surface wettability: Structural, electronic and mechanical properties of rutile ruo2 by means of spin-polarized dft-md. *Applied Surface Science*, 570:150993, 2021.

[278] Giuseppe Cassone, Jiri Sponer, Judit E. Sponer, and Franz Saija. Electrofreezing of liquid ammonia. *The Journal of Physical Chemistry Letters*, 13(42):9889–9894, 2022. PMID: 36255376.

[279] Sara Laporte, Fabio Pietrucci, François Guyot, and A. Marco Saitta. Formic acid synthesis in a water–mineral system: Major role of the interface. *The Journal of Physical Chemistry C*, 124(9):5125–5131, 2020.

[280] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, Tiziano Müller, Robert Schade, Manuel Guidon, Samuel Andermatt, Nico Holmberg, Gregory K. Schenter, Anna Hehn, Augustin Bussy, Fabian Belleflamme, Gloria Tabacchi, Andreas Glöß, Michael Lass, Iain Bethune, Christopher J. Mundy, Christian Plessl, Matt Watkins, Joost VandeVondele, Matthias Krack, and Jürg Hutter. Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19):194103, 2020.

[281] R. D. King-Smith and David Vanderbilt. Theory of polarization of crystalline solids. *Phys. Rev. B*, 47:1651–1654, Jan 1993.

[282] Raffaele Resta. Macroscopic polarization in crystalline dielectrics: the geometric phase approach. *Rev. Mod. Phys.*, 66:899–915, Jul 1994.

[283] Michael Victor Berry. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 392(1802):45–57, 1984.

[284] P. Umari and Alfredo Pasquarello. Ab initio molecular dynamics in a finite homogeneous electric field. *Phys. Rev. Lett.*, 89:157602, Sep 2002.

[285] Raffaele Resta. Quantum-mechanical position operator in extended systems. *Phys. Rev. Lett.*, 80:1800–1803, Mar 1998.

[286] Woo-Kyung Lee, Stanislav Tsoi, Keith E. Whitener, Rory Stine, Jeremy T. Robinson, Jonathon S. Tobin, Asanka Weerasinghe, Paul E. Sheehan, and Sergei F. Lyuksyutov. Robust reduction of graphene fluoride using an electrostatically biased scanning probe. *Nano Research*, 6(11):767–774, Nov 2013.

[287] M. Krack. Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts*, 114(1):145–152, Sep 2005.

[288] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, Sep 1988.

[289] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[290] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of Chemical Physics*, 132(15):154104, 2010.

[291] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7):1456–1465, 2011.

[292] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[293] Jianwei Sun, Adrienn Ruzsinszky, and John P. Perdew. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.*, 115:036402, Jul 2015.

[294] Haowei Peng, Zeng-Hui Yang, John P. Perdew, and Jianwei Sun. Versatile van der waals density functional based on a meta-generalized gradient approximation. *Phys. Rev. X*, 6:041005, Oct 2016.

[295] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008. PMID: 26620784.

[296] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of Chemical Physics*, 123(23):234505, 2005.

[297] J. L. Aragones, L. G. MacDowell, J. I. Siepmann, and C. Vega. Phase diagram of water under an applied electric field. *Phys. Rev. Lett.*, 107:155702, Oct 2011.

[298] M. Parrinello and A. Rahman. Study of an f center in molten kcl. *The Journal of Chemical Physics*, 80(2):860–867, 1984.

[299] Tawny N Fajardo and Matthias Heyden. Dissecting the conformational free energy of a small peptide in solution. *The Journal of Physical Chemistry B*, 125(18):4634–4644, 2021.

[300] Shiang-Tai Lin, Mario Blanco, and William A. Goddard. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of lennard-jones fluids. *The Journal of Chemical Physics*, 119(22):11792–11805, 2003.

[301] Shiang-Tai Lin, Prabal K. Maiti, and William A. Goddard. Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations. *The Journal of Physical Chemistry B*, 114(24):8191–8198, 2010. PMID: 20504009.

[302] Tod A. Pascal, Daniel Schärf, Yousung Jung, and Thomas D. Kühne. On the absolute thermodynamics of water from computer simulations: A comparison of first-principles molecular dynamics, reactive and empirical force fields. *The Journal of Chemical Physics*, 137(24):244507, 2012.

[303] Jeffrey R. Errington and Pablo G. Debenedetti. Relationship between structural order and the anomalies of liquid water. *Nature*, 409(6818):318–321, Jan 2001.

[304] Michael J. Gillan, Dario Alfè, and Angelos Michaelides. Perspective: How good is dft for water? *The Journal of Chemical Physics*, 144(13):130901, 2016.

[305] *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Linstrom, P.J. and Mallard, W.G., National Institute of Standards and Technology, Gaithersburg MD, 20899, 2000.

[306] Cui Zhang, Leonardo Spanu, and Giulia Galli. Entropy of liquid water from ab initio molecular dynamics. *The Journal of Physical Chemistry B*, 115(48):14190–14195, 2011. PMID: 21961845.

[307] Alexandru Botan, Virginie Marry, and Benjamin Rotenberg. Diffusion in bulk liquids: finite-size effects in anisotropic systems. *Molecular Physics*, 113(17-18):2674–2679, 2015.

[308] David M. Wilkins, David E. Manolopoulos, Silvio Pipolo, Damien Laage, and James T. Hynes. Nuclear quantum effects in water reorientation and hydrogen-bond dynamics. *The Journal of Physical Chemistry Letters*, 8(12):2602–2607, 2017. PMID: 28530836.

[309] Michele Ceriotti, Jérôme Cuny, Michele Parrinello, and David E. Manolopoulos. Nuclear quantum effects and hydrogen bond fluctuations in water. *Proceedings of the National Academy of Sciences*, 110(39):15591–15596, 2013.

[310] Mahdi Shafiei, Michael von Domaros, Dusan Bratko, and Alenka Luzar. Anisotropic structure and dynamics of water under static electric fields. *The Journal of Chemical Physics*, 150(7):074505, 2019.

[311] Godehard Sutmann. Structure formation and dynamics of water in strong external electric fields. *Journal of Electroanalytical Chemistry*, 450(2):289–302, 1998.

[312] Yagel Peleg, Alexander Yoffe, David Ehre, Meir Lahav, and Igor Lubomirsky. The role of the electric field in electrofreezing. *The Journal of Physical Chemistry C*, 123(50):30443–30446, 2019.

[313] Matthias Heyden and Martina Havenith. Statistically converged properties of water from ab initio molecular dynamics simulations. In *High Performance Computing in Science and Engineering, Garching/Munich 2009*, pages 687–698. Springer, 2010.

[314] M Madan Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions*, 44(5):1185–1200, 2016.

[315] Hebah Fatafta, Suman Samantray, Abdallah Sayyed-Ahmad, Orkid Coskuner-Weber, and Birgit Strodel. Molecular simulations of idps: From ensemble generation to idp interactions leading to disorder-to-order transitions. *Progress in Molecular Biology and Translational Science*, 183:135–185, 2021.

[316] Alexander Cumberworth, Guillaume Lamour, M Madan Babu, and Jörg Gsponer. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal*, 454(3):361–369, 2013.

[317] Amrita Mohan, Christopher J Oldfield, Predrag Radivojac, Vladimir Vacic, Marc S Cortese, A Keith Dunker, and Vladimir N Uversky. Analysis of molecular recognition features (morfs). *Journal of molecular biology*, 362(5):1043–1059, 2006.

[318] Daniel W Carr, Renata E Stofko-Hahn, ID Fraser, Sarah M Bishop, TS Acott, RG Brennan, and JD Scott. Interaction of the regulatory subunit (rii) of camp-dependent protein kinase with rii-anchoring proteins occurs through an amphipathic helix binding motif. *Journal of Biological Chemistry*, 266(22):14188–14192, 1991.

[319] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology*, 12(1):54–60, 2002.

[320] Ethan Garner, Pedro Romero, A Keith Dunker, Celeste Brown, and Zoran Obradovic. Predicting binding regions within disordered proteins. *Genome informatics*, 10:41–50, 1999.

[321] Bálint Mészáros, István Simon, and Zsuzsanna Dosztányi. Prediction of protein binding regions in disordered proteins. *PLoS computational biology*, 5(5):e1000376, 2009.

[322] Munehito Arai, Kenji Sugase, H Jane Dyson, and Peter E Wright. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proceedings of the National Academy of sciences*, 112(31):9614–9619, 2015.

[323] Monika Fuxreiter, István Simon, Peter Friedrich, and Peter Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *Journal of molecular biology*, 338(5):1015–1026, 2004.

[324] Do-Hyoung Kim and Kyou-Hoon Han. Presmo target-binding signatures in intrinsically disordered proteins. *Molecules and Cells*, 41(10):889, 2018.

[325] Karen Skriver, Frederik Friis Theisen, and Birthe B Kragelund. Conformational entropy in molecular recognition of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 83:102697, 2023.

[326] Kaare Teilum, Johan G Olsen, and Birthe B Kragelund. Globular and disordered—the non-identical twins in protein-protein interactions. *Frontiers in molecular biosciences*, 2:40, 2015.

[327] Tamas Lazar, Agnes Tantos, Peter Tompa, and Eva Schad. Intrinsic protein disorder uncouples affinity from binding specificity. *Protein Science*, 31(11):e4455, 2022.

[328] Igor Drobnak, Natalie De Jonge, Sarah Haesaerts, Gorazd Vesnaver, Remy Loris, and Jurij Lah. Energetic basis of uncoupling folding from binding for an intrinsically disordered protein. *Journal of the American Chemical Society*, 135(4):1288–1294, 2013.

[329] Frederik Friis Theisen, Lasse Staby, Frederik Grønbæk Tidemand, Charlotte O'Shea, Andreas Prestel, Martin Willemoes, Birthe B Kragelund, and Karen Skriver. Quantification of conformational entropy unravels effect of disordered flanking region in coupled folding and binding. *Journal of the American Chemical Society*, 143(36):14540–14550, 2021.

[330] Uroš Zavrtanik, San Hadži, and Jurij Lah. Unraveling the thermodynamics of ultratight binding of intrinsically disordered proteins. *Frontiers in Molecular Biosciences*, 8:726824, 2021.

[331] Tilman Flock, Robert J Weatheritt, Natasha S Latysheva, and M Madan Babu. Controlling entropy to tune the functions of intrinsically disordered regions. *Current opinion in structural biology*, 26:62–72, 2014.

[332] Anusha Poosapati, Emily Gregory, Wade M Borcherds, Lucia B Chemes, and Gary W Daughdrill. Uncoupling the folding and binding of an intrinsically disordered protein. *Journal of molecular biology*, 430(16):2389–2402, 2018.

[333] Joseph M Rogers, Chi T Wong, and Jane Clarke. Coupled folding and binding of the disordered protein puma does not require particular residual structure. *Journal of the American Chemical Society*, 136(14):5197–5200, 2014.

[334] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[335] Avijit Chakrabartty, Tanja Kortemme, and Robert L Baldwin. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Science*, 3(5):843–852, 1994.

[336] Shneior Lifson and A Roig. On the theory of helix—coil transition in polypeptides. *The Journal of Chemical Physics*, 34(6):1963–1974, 1961.

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

[338] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G Bolhuis, Sandro Bottaro, Davide Branduardi, et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature methods*, 16(8):670–673, 2019.

[339] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n□ log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[340] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[341] Carlo Camilloni, Davide Provasi, Guido Tiana, and Ricardo A Broglia. Exploring the protein g helix free-energy surface by solute tempering metadynamics. *Proteins: Structure, Function, and Bioinformatics*, 71(4):1647–1654, 2008.

[342] Carl B Boyer and Uta C Merzbach. *A history of mathematics*. John Wiley & Sons, 2011.

[343] Cheng Ming Lee. Generalizations of l'hôpital's rule. *Proceedings of the American Mathematical Society*, 66(2):315–320, 1977.

[344] Denis J Evans and Brad Lee Holian. The nose–hoover thermostat. *J. Chem. Phys.*, 83(8):4069–4074, 1985.

[345] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[346] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[347] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

[348] Leonardo Darré, Matías Rodrigo Machado, Astrid Febe Brandner, Humberto Carlos González, Sebastián Ferreira, and Sergio Pantano. Sirah: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *Journal of chemical theory and computation*, 11(2):723–739, 2015.

[349] Xavier Periole and Siewert-Jan Marrink. The martini coarse-grained force field. *Biomolecular simulations: Methods and protocols*, pages 533–565, 2013.

[350] Siewert J Marrink and D Peter Tieleman. Perspective on the martini model. *Chemical Society Reviews*, 42(16):6801–6822, 2013.

[351] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[352] Matías R Machado and Sergio Pantano. Sirah tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics*, 32(10):1568–1570, 2016.

[353] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. *Nucleic acids research*, 32(suppl_2):W665–W667, 2004.

[354] Monique M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, 77(9):1905, 1996.

[355] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1998.

[356] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[357] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Physical review letters*, 79(16):3090, 1997.

[358] Yuichi Togashi and Holger Flechsig. Coarse-grained protein dynamics studies using elastic network models. *International journal of molecular sciences*, 19(12):3899, 2018.

[359] Sambit Kumar Mishra and Robert L Jernigan. Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics. *PLoS One*, 13(6):e0199225, 2018.

[360] Daniel Trzesniak, Anna-Pitschna E Kunz, and Wilfred F van Gunsteren. A comparison of methods to compute the potential of mean force. *ChemPhysChem*, 8(1):162–169, 2007.

[361] Ilja V Khavrutskii, Joachim Dzubiella, and J Andrew McCammon. Computing accurate potentials of mean force in electrolyte solutions with the generalized gradient-augmented harmonic fourier beads method. *The Journal of chemical physics*, 128(4), 2008.

[362] Dudley R Herschbach, Harold S Johnston, and Donald Rapp. Molecular partition functions in terms of local properties. *The Journal of Chemical Physics*, 31(6):1652–1661, 1959.

[363] Hong Qian. An asymptotic comparative analysis of the thermodynamics of noncovalent association. *Journal of mathematical biology*, 52:277–289, 2006.

[364] John D Chodera and David L Mobley. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annual review of biophysics*, 42:121–142, 2013.

[365] Jochen S Hub, Bert L De Groot, and David Van Der Spoel. g_wham□ a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of chemical theory and computation*, 6(12):3713–3720, 2010.

[366] Sarah K Madden, Aline Dantas de Araujo, Mara Gerhardt, David P Fairlie, and Jody M Mason. Taking the myc out of cancer: toward therapeutic strategies to directly inhibit c-myc. *Molecular Cancer*, 20(1):3, 2021.

[367] Rajiv Kumar Jha, Fedor Kouzine, and David Levens. Myc function and regulation in physiological perspective. *Frontiers in Cell and Developmental Biology*, 11:1268275, 2023.

[368] Natalie Meyer and Linda Z Penn. Reflecting on 25 years with myc. *Nature Reviews Cancer*, 8(12):976–990, 2008.

[369] Marie-Eve Beaulieu, Francisco Castillo, and Laura Soucek. Structural and biophysical insights into the function of the intrinsically disordered myc oncoprotein. *Cells*, 9(4):1038, 2020.

[370] Chi V Dang. Myc on the path to cancer. *Cell*, 149(1):22–35, 2012.

[371] Jim Battey, Christopher Moulding, Rebecca Taub, William Murphy, Timothy Stewart, Huntington Potter, Gilbert Lenoir, and Philip Leder. The human c-myc oncogene: structural consequences of translocation into the igh locus in burkitt lymphoma. *Cell*, 34(3):779–787, 1983.

[372] Karyn T O'Neil and William F DeGrado. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, 250(4981):646–651, 1990.

[373] Pingchang C Lyu, Mark I Liff, Luis A Marky, and Neville R Kallenbach. Side chain contributions to the stability of alpha-helical structure in peptides. *Science*, 250(4981):669–673, 1990.

[374] C Nick Pace and J Martin Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical journal*, 75(1):422–427, 1998.

[375] Peter Y Chou and Gerald D Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2):211–222, 1974.

[376] Hao Dong, Mukesh Sharma, Huan-Xiang Zhou, and Timothy A Cross. Glycines: role in α-helical membrane protein structures and a potential indicator of native conformation. *Biochemistry*, 51(24):4779–4789, 2012.

[377] Robert B Best and Gerhard Hummer. Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides. *J. Phys. Chem. B*, 113(26):9004–9015, 2009.

[378] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[379] PINGCHIANG C Lyu, John C Sherman, Amy Chen, and Neville R Kallenbach. Alpha-helix stabilization by natural and unnatural amino acids with alkyl side chains. *Proceedings of the National Academy of Sciences*, 88(12):5317–5320, 1991.

[380] Erik J Spek. Alanine is an intrinsic alpha-helix stabilizing amino acid. *J. Am. Chem. Soc.*, 121:5571–5572, 1999.

[381] Carol A Rohl, Wayne Fiori, and Robert L Baldwin. Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. *Proceedings of the National Academy of Sciences*, 96(7):3682–3687, 1999.

[382] M Giam, DCS Huang, and P Bouillet. Bh3-only proteins and their roles in programmed cell death. *Oncogene*, 27(1):S128–S136, 2008.

[383] Richard J Youle and Andreas Strasser. The bcl-2 protein family: opposing activities that mediate cell death. *Nature reviews Molecular cell biology*, 9(1):47–59, 2008.

[384] Viacheslav V Senichkin, Alena Y Streletskaia, Anna S Gorbunova, Boris Zhivotovsky, and Gelina S Kopeina. Saga of mcl-1: regulation from transcription to degradation. *Cell Death & Differentiation*, 27(2):405–419, 2020.

[385] Thomas P Garner, Andrea Lopez, Denis E Reyna, Adam Z Spitz, and Evripidis Gavathiotis. Progress in targeting the bcl-2 family of proteins. *Current Opinion in chemical biology*, 39:133–142, 2017.

APPENDIX A

MODEL-DEPENDENT SOLVATION OF THE K-18 DOMAIN OF THE INTRINSICALLY DISORDERED PROTEIN TAU - REPRODUCTION LICENCE

Model-Dependent Solvation of the K-18 Domain of the Intrinsically Disordered Protein Tau

Author: Sthitadhi Maiti, Matthias Heyden
Publication: The Journal of Physical Chemistry B
Publisher: American Chemical Society
Date: Aug 1, 2023

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

BACK CLOSE WINDOW

APPENDIX B

ELECTRIC-FIELD INDUCED ENTROPIC EFFECTS IN LIQUID WATER - REPRODUCTION LICENCE

Electric-field induced entropic effects in liquid water

Author: Conti Nibali, Valeria; Maiti, Sthitadhi

Publication: Journal of Chemical Physics
Publisher: AIP Publishing
Date: May 8, 2023