Absolute Binding Enthalpy Calculations Using Molecular Dynamics Simulations

Süleyman Selim Çınaroğlu

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Dedicated to all who lost their lives and all impacted by the devastating effects in the Kahramanmaraş Earthquake on February 6, 2023.

Acknowledgements

With heartfelt appreciation, I would like to extend my gratitude to my supervisor, Professor Phil Biggin, for his unwavering support and guidance throughout my doctoral research journey. He provided me with valuable advice and a welcoming environment for discussions and support while also allowing me to take the lead and make decisions about the direction and details of my project. This approach allowed me to grow as a scientist and truly enjoy my experience as a PhD student.

I am also profoundly grateful to the individuals I met within the Structural Bioinformatics and Computational Biochemistry group. Although this acknowledgement section is limited, I want to express my gratitude to everyone who has positively impacted my experience in Oxford and the UK and helped shape who I am today.

I cannot forget to extend my appreciation to my friends and housemates in Oxford, especially Miguel Berbeira Santana, who made my time here memorable.

I am deeply grateful to my family, particularly my parents, for their love and unwavering support throughout my life. They have always encouraged and believed in me, and I am grateful for their presence.

Last but not least, I want to express my immense gratitude to my wife, Muazzez Ceren YILMAZ. She has been a constant source of support and understanding and has helped me in countless ways. I am incredibly lucky to have her in my life.

I express my heartfelt gratitude to my wonderful friends, Suleyman, Hazal, and Mine, who have been by my side throughout this journey. Their unwavering support and companionship have been invaluable, providing encouragement and motivation when I needed it the most. Thank you, Suleyman, Hazal, and Mine, for your constant presence and unwavering belief in me.

I want to express my gratitude to the Ministry of National Education (MoNE) of The Republic of Türkiye for their support through the International Graduate Education Scholarship (YLSY) program. This sponsorship made this work possible, and I am thankful for their support in my education.

Süleyman Selim Çınaroğlu

Abstract

Computers play an essential role in drug discovery as advancements in technology, hardware, and algorithms have allowed for improved simulations of biomolecules. The field of drug discovery stands to benefit significantly from these developments. Currently, many innovative approaches to studying drug binding and predicting binding affinity are being explored. Using computational methods to predict thermodynamic components in drug design has become routine. While progress has been made in calculating free energy, the prediction of enthalpy and entropy remains an area that requires further investigation. These components reflect the interactions and dynamics between the ligand and protein. However, despite years of research, our understanding of these components still needs to be improved. Computing the enthalpy is particularly challenging, and even the achievable accuracy of these predictions is still not precise despite the apparent simplicity of the calculations *per se*.

In my thesis, I conduct a series of studies to examine the potential utility of absolute binding enthalpy calculations using the direct method based on molecular dynamics simulations. In Chapter 3, I first assess the accuracy of water models and the host-guest force field in calculating the absolute binding enthalpy for 25 host-guest pairs. While actual protein-ligand or protein-protein data would be ideal for evaluating force fields, using very simplified test systems can be helpful for preliminary exploration of parameters. Then, in Chapter 4, I focus on predicting the binding enthalpies of small molecules to bromodomains, which are small protein modules involved in gene regulation linked to many diseases, such as cancer and inflammation. I evaluated the direct method for calculating absolute binding enthalpies by testing its ability to predict the binding enthalpies of 10 different ligands to BRD4-1. The results showed a strong correlation between the behaviour of the ZA loop and the predicted enthalpy. In Chapter 5, I extended the study by evaluating the method to include multiple protein-protein complexes essential in all cellular processes, ranging from signal transmission to enzyme activity. Understanding the thermodynamics of protein-peptide binding events is a significant challenge in computational chemistry. The complexity of both components having many degrees of freedom presents a substantial challenge for methods attempting to directly compute the enthalpic contribution to binding. Despite this, the method produced highly accurate and well-converged binding enthalpies for small protein-protein systems. Perhaps

unsurprisingly, most inaccuracies can be attributed to poor conformational sampling. Nevertheless, I have shown that this can actually be used to highlight the possibility of hidden states. Overall, my work has shown that absolute enthalpy calculations using the direct method can be performed on protein-ligand and protein-protein systems with reasonable accuracy and that this is a useful contribution to computational drug design.

Contents

List of Figures

List of Tables

List of Abbreviations

μVT	Grand Canonical Ensemble
ABFE	Absolute Binding Free Energy
APR	Attach pull Release
ATB	Automated Topology Builder
CB7	Cucurbit[7]uril
CGENFF	CHARMM General Force Field
CHARMM	Chemistry at Harvard Macromolecular Mechanics
ESP	Electrostatic Potential
GAFF	General AMBER Force Field
GROMACS	GROningen MAchine for Chemical Simulations
GROMOS	GROningen MOlecular Simulation
ITC	Isothermal Titration Calorimetry
LINCS	LINear Constraint Solver
MAE	Mean Absolute Error
MD	Molecular Dynamics
MMFF	Merck Molecular Force Field
MSE	Mean Squared Error
NMR	Nuclear Magnetic Resonance
NPT	Isothermal Isobaric Ensemble
NVE	Microcanonical Ensemble
NVT	Canonical Ensemble
OPLS	Optimized Potentials for Liquid Simulations
PBC	Periodic boundary conditions
PDB	Protein Data Bank
PDF	Probability Density Distribution
PINT	Protein–protein Interactions Thermodynamic Database
PPI	Protein–protein Interaction

QM	Quantum Mechanics
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAMPL	Statistical Assessment of the Modeling of Proteins and Ligands
SASA	Solvent Accessible Surface Area
SEM	Standard Error of the Mean
ZAFF	Zinc AMBER Force Field
βCD	β-cyclodextrin

"The only way to have real success in science, the field I'm familiar with, is to describe the evidence very carefully without regard to the way you feel it should be. If you have a theory, you must try to explain what's good and what's bad about it equally. In science, you learn a kind of standard integrity and honesty."

~ Richard Feynman

1 Introduction

This chapter will briefly introduce computational calorimetry and its applications to studying host-guest and other complex systems. We will explore how computational methods, especially the direct method, can be used to calculate the absolute binding enthalpy of these systems. We will discuss the significance of the BRD4 protein as a target for computational studies, drug discovery, and the importance of protein-protein interactions in computational studies.

1.1. Computational Calorimetry

Designing high-affinity ligands for drug discovery and other applications requires understanding how the association between the ligand, receptor and solvent contribute to overall binding affinity.[1, 2] This information helps in drug discovery and other applications to create ligands with strong binding properties. Understanding the thermodynamics of these associations at the molecular level can offer insight into the mechanisms behind various biological processes, such as enzyme function, signalling, molecular recognition, and receptor activation.[3] The primary aim of drug design and discovery in pharmaceuticals is to identify new chemical compounds that exhibit a strong affinity for specific pharmaceutical targets.[4, 5]

Developing a new compound is still challenging despite advances in understanding biological systems and numerous biotechnological improvements.[5] This is due to many factors, such as complex molecular interactions[6], varying biological responses[7], and difficulties synthesizing new compounds[8, 9]. The process of finding new drugs that have the potential to bind a specific biological target can be broken down into four phases: identifying the target, verifying the target, identifying potential lead compounds, and confirming the potential of those lead compounds. Before human clinical trials, the preclinical evaluation typically involves testing the compounds on animals or in laboratory

cultures to gather data that can inform the design and conduct of human clinical trials. This process integrates various disciplines and significant economic investments and typically takes 10-15 years from target identification to the market (**Figure 1.1**).[10, 11]

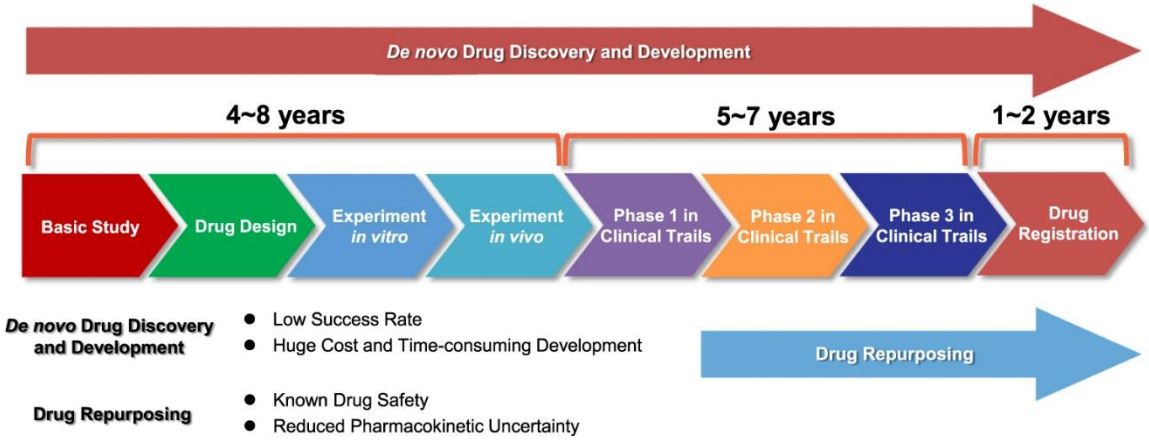

***Figure 1.1**: The estimated time and main steps in de novo drug discovery, development, and drug repurposing. Reproduced with permission from Zhang, Z., Zhou, L., Xie, N. et al. Sig Transduct Target Ther 5, 113 (2020).*[11]

Unsurprisingly, computational methods have become increasingly important in drug discovery in recent decades.[12-14] The advancements in computational techniques and hardware capabilities have made it possible to use computational methods to efficiently screen large numbers of potential drug candidates and predict their behaviour in the body.[15, 16] This involves simulating and predicting the behaviour of potential drugs using computational tools and simulations rather than relying solely on laboratory experiments, which can be time-consuming and expensive. Researchers can integrate time-consuming and expensive experiments with faster and less costly computer simulations by utilising computational algorithms to simulate biological macromolecules' structural and functional properties. This integration leads to more efficient and cost-effective identification of promising drug candidates.[17, 18] Thus, all steps in the drug discovery process are closely linked to computational methods.[19]

There is a strong interest in computational methods that can predict a molecule's docking pose and binding energy to pharmaceutical targets. Structure-based drug design utilizes the three-dimensional structure of a biomolecule to design compounds that will fit tightly into the active site and control its biological function as desired. Computational methods rely on mathematical models that capture the interactions between the target and candidate molecules and can include molecular dynamics (MD) simulations[20], molecular

docking[21], and machine learning[22] approaches. Additionally, computational tools can also be used to assess the specificity of potential ligands for various targets, enabling the development of drugs with fewer side effects and reducing the time and resources required to discover new lead compounds. Although estimating the binding affinity of potential drug candidates for a target is the main objective of computational drug development, computational studies can also describe the binding free energy and its entropic and enthalpic components.[2, 23, 24] Molecular simulations are essential for predicting the thermodynamic components, providing insight into molecular recognition and its structural and physical interactions. The design of high-affinity binders can be affected by the thermodynamic subcomponents of the free energy, which reflect the key determinants underlying molecular interactions. Researchers may better understand entropy-enthalpy compensation if we have a better grasp of their contributions at the atomic level.[25-28]

Computation's rapid increase in capacity enables MD simulations to guide our comprehension of macromolecules and their interactions at an unparalleled level.[16, 29, 30] Molecular simulations offer a detailed look at the link between atomistic detail and the thermodynamics signature of the binding. Specifically, simulating thermodynamic quantities provides atomistic details and a connection between molecular interactions and thermodynamic observations. I use MD simulations to examine the direct calculation of absolute binding enthalpy for various systems using their end states.[31] Understanding how modifications impact molecular interactions and, consequently, enthalpy/entropy might provide light on enthalpy-entropy compensation or explain small modifications to binding energy that would be challenging to identify based on experimental data. Enthalpy measures the heat absorbed or released during a chemical reaction. On the other hand, entropy measures the amount of disorder or randomness in a system. Understanding the relationship between enthalpy and entropy can provide valuable insights into the molecular interactions that lead to binding.[26]

Additionally, a method to enhance the binding affinity of lead molecules involves optimizing both enthalpy/entropy through correlation with structural features such as surface area and chemical properties.[28, 32] In some cases, understanding binding enthalpy/entropy has provided details about the molecular association that are not readily evident when focusing just on binding free energy. Enthalpy/entropy patterns can be used to identify the molecular factors that affect the molecular association. Additionally, in one

instance, a pattern not apparent from free energy alone was shown by binding enthalpy, which separated one set of ligands in a series from another.[33] As a result, a careful analysis of binding enthalpies/entropies can provide information on the physical factors that govern molecule association, with computational techniques employing physics-based models providing atomic details of the underlying interactions.[24, 34-36]

While simulation-based approaches have successfully estimated binding free energy, predicting entropy and enthalpy components is still difficult.[24, 34, 35, 37] Typically, the calculations of the enthalpy and entropy components based on the derivatives of the free energy function are less precise. Their errors are more substantial compared to the estimates of the free energy function itself. The direct method, which uses end-states and offers a clear interpretation of physical behaviour, is the most straightforward approach to estimating enthalpy. However, it determines the binding enthalpy by calculating the difference between the potential energies of the bound and unbound states produced by separate simulations. The discrepancy is much smaller than the absolute potential energy values, causing the direct method's dependability to rely on the level of sampling that can be accomplished within reasonable computation time. As a result, using the direct method to determine absolute binding enthalpies for complex systems is uncommon.[34, 35]

Knowing the binding enthalpy alongside the binding free energy offers several advantages. While binding free energy represents the overall stability of the complex, binding enthalpy specifically quantifies the energy change upon complex formation. It helps understand the underlying contributions to the binding process, especially in drug design. Medicinal chemists can assess whether modifications lead to the expected improvement in the binding free energy, thanks to knowing the enthalpy, which helps understand the impact of alterations during the optimization process.[38, 39] Accurate estimation of binding enthalpy can be challenging and may require advanced computational methods. However, this information is valuable in understanding the driving forces behind the binding process, optimizing drug design, assessing binding mechanisms and kinetics, and predicting ligand binding affinity. It enhances our understanding of thermodynamics, molecular interactions, and kinetic aspects of ligand-receptor binding, which are crucial for rational drug design and therapeutic interventions. Decomposing the enthalpic and entropic contributions within the overall binding free energy offers valuable insights for optimizing ligand efficacy in two significant ways. Firstly, it becomes feasible to prioritize

compounds for subsequent improvement by conducting comprehensive thermodynamic analyses. Enhancing the enthalpic component is often more challenging than improving the entropic component, so the compound having a more favourable enthalpic change can be selected for further optimization efforts.[38] Secondly, calorimetric data obtained from analogues of the ligand can offer a deeper understanding of the molecular forces crucial for the binding interactions, mainly when structural details are available, contributing to more informed drug discovery efforts.[40]

Recent advances in molecular simulations have allowed the study of complex systems at the atomic level. While calculating binding enthalpy can be difficult, the direct approach, which has been shown to yield numerically precise results for host-guest systems, is becoming more practical with increasing computer power.[31] This approach is favoured for its simplicity and ability to break down enthalpy contributions from various system components, such as the ligand and binding site residues, and also enables informative decompositions of the binding enthalpy into energy components.[31]

In this thesis, I have meticulously evaluated the effectiveness and versatility of this method by conducting a thorough examination across various systems. Furthermore, I have delved into the intricacies of the method and explored its limitations. Despite previous literature on the topic providing promising results, I have aimed to provide a more comprehensive analysis to advance our understanding of this method's applicability in complex systems.

1.2. Binding Enthalpy Calculation using The Direct Method

Computing the binding enthalpy changes through simulations has proven to be a difficult task despite the availability of various computational methods for enthalpy estimation. The direct method that uses end-states is simple and allows for an immediate understanding of physical behaviour. The accuracy of the direct method depends on the sampling level that can be achieved with a huge amount of simulation data.[35, 41] The direct method calculates the binding enthalpy by subtracting the energies of the bound and unbound states obtained from individual simulations. It's important to note that the direct enthalpy method provides valuable insights into the decomposition of binding enthalpies, including connections with conformational preferences, which other enthalpy methods may not achieve.[31, 35] This ability to decompose binding enthalpies could help resolve conflicting findings in calorimetric studies or be used to identify hidden states. Additionally, by combining the enthalpy calculation with a precise estimate of binding free energy, the binding entropy can be immediately obtained, which has been challenging to get with high precision from simulations.

Other enthalpy estimation methods typically require many simulations at different temperatures to determine the enthalpy change from the initial to the final state. One of the most widely used methods for computing binding enthalpy is the finite difference calculating binding free energy at various temperatures and then utilising the Van't Hoff equation to estimate the binding enthalpy at a specific temperature.[42, 43] However, the reliability of the finite difference method is affected by the assumptions regarding temperature dependence and heat capacity.[44] The finite difference approach calculates the entropy/enthalpy without considering variations in heat capacity, while the direct method determines the enthalpy at a specific temperature. Furthermore, the finite difference method demands an accurate force field across the whole temperature range selected for the simulation. In summary, the direct binding enthalpy method requires less simulation time to achieve sufficient statistical uncertainty. It is easier to implement in small host-guest systems than the Van't Hoff method.[2]

In addition to the Van't Hoff method, end-point methods have been extensively used in structure-based drug design.[45-47] Combined with the change in conformational entropy ($-T\Delta S$) usually calculated by normal-mode analysis, these techniques provide binding free energies.[24, 48, 49] However, the entropy remains challenging to estimate and is often

neglected in endpoint methods. As the name suggests, end-point methods rely on sampling the end states of a system. Molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) and molecular mechanics Generalized–Born surface area (MM/GBSA) are well-known endpoint methods that balance computational efficiency and accuracy.[45, 46] Since the PB solution is computationally expensive, more efficient approximation methods based on the GB model have been developed and gained popularity. Another popular technique is LIE, which only requires complex and free ligand simulation.[50-52] MM/PBSA and MM/GBSA have been widely used to assess docking poses, determine structural stability, and estimate binding affinities. They also enable decomposition analysis to provide residue-specific contributions, identify critical interactions, and facilitate targeted drug design. While discussing endpoint methods extensively here is beyond the scope, their contribution should be acknowledged.

As computer power continues to increase, computational calorimetry studies are expected to become more common for complex biomolecular systems. To the best of our knowledge, a small number of studies have been performed thus far that use the direct method to calculate the enthalpies of complex systems containing proteins. Roy *et al.* reported the first study showing binding enthalpy calculations using MD simulations with the direct method.[34] It showed the relative binding enthalpies of the Src SH2 domain and three small peptides using 400 ns total simulation time. Similarly, Li and Gilson made a significant contribution by calculating the relative binding enthalpies of four peptidic ligands and the Grb2 SH2 domain through MD simulations using the direct method.[35] These studies are notable as they are the only ones known to have employed this method for calculating binding enthalpies in complex systems involving proteins, highlighting the potential of this approach in the field. Another aspect of computational calorimetry is using host-guest systems. The direct method was used for the first time to calculate the absolute binding enthalpies of host-guest systems.[31]

In a study conducted by Fenley *et al*., the binding enthalpies for cucurbit[7]uril (CB7) were computed with eight guests using two different water models (TIP3P and TIP4P-Ew).[31] The results showed that the TIP3P water model was more accurate than the TIP4P-Ew model. This realization of the impact of different water models on the enthalpy result prompted further testing using other water models. Gao et al. then investigated the binding enthalpies across four water models (TIP3P, SPC/E, TIP4P-Ew, and OPC) and found

considerable variations in the salt dependency of the binding enthalpy.[53] The results also revealed sensitivity to the choice of parameters for sodium and chloride.

Further, Henriksen et al. performed highly accurate calculations of binding free energy and binding enthalpy values using the attach–pull–release (APR).[2, 54] The method was applied to determine the thermodynamic profiles for binding nine guests to either CB7 or β-cyclodextrin (βCD) host. This study introduced a concept called computational calorimetry, allowing for the calculation of both the binding free energy and binding enthalpy from a single set of simulations in a self-consistent manner for the first time.[2] Previously, Fenley *et al*.[31] applied the direct method using a multi-box approach, which requires four separate simulations of the free host, free guest, host-guest complex, and solvent to precisely balance the stoichiometry of both the bound and unbound simulations. Instead, they employed a single-box approach, which simply subtracts the mean energy of the host-guest system in the pulled-free state from that of the bound complex. Henriksen and Gilson later employed this computational calorimetry approach to evaluate and compare the precision of force fields and several water models in the computation of binding thermodynamics.[54] Overall, they suggest that no particular combination of force fields was found to be superior among those tested, and the thermodynamic values obtained varied significantly.

Moreover, the same computational calorimetry approach using APR was utilized to calculate binding free energy and enthalpy values for the SAMPL5 challenge, which resulted in good correlations with the experimental values by Yin *et al.*[55, 56] However, the deviations from the observed binding affinities remained substantial compared to the objective of forecasting protein-ligand binding affinities with a precision of approximately 1 kcal/mol. The binding enthalpy calculations only showed moderate correlations, even though they demonstrated good convergence through high numerical accuracy. The TIP3P water model was more effective than the OPC model for the SAMPL5 systems.

Yin et al. demonstrated the initial application of host-guest binding data to optimize a water model. They developed a new water model called Bind3P by adjusting the Lennard-Jones parameters of oxygen atom in TIP3P.[57] This new model proved to be more effective than TIP3P for determining binding-free energies between host and guest molecules and calculating hydration-free energy. This led to a consistent enhancement in precision across a comprehensive test set of host-guest binding thermodynamics.

The findings indicated a striking sensitivity in the binding enthalpies of host-guest systems to the chosen force field and water model. Moreover, it was observed that the enthalpies are more affected by force field parameters than free energies, suggesting an in silico entropy-enthalpy compensation case.[35] Despite the progress that has been made, the need for a comparative analysis of more force fields and water models remains. Therefore, in Chapter 3, I carried out a comprehensive evaluation of eight different water models (TIP3P, TIP4P, TIP4P-Ew, SPC, SPC/E, OPC, TIP5P, Bind3P) and five force fields (GAFFv1, GAFFv2, CGenFF, Parsley, and SwissParam) for the binding enthalpies of 25 CB7-guest pairs.[58]

1.3. Host-guest Systems

The Nobel Prize-winning discoveries of Lehn, Cram, and Pedersen in 1987 concerning host-guest systems sparked widespread interest in supramolecular chemistry among chemists, biologists, and material scientists.[59] The host-guest systems have been extensively explored in creating various host-guest systems and examining their uses. In recent decades, much effort has been devoted to developing and applying host-guest systems.[60-63] A host-guest system, also known as an inclusion complex, is a chemical compound in which one molecule, known as the guest, is incorporated into the cavity or pores of another molecule, known as the host. The guest molecule can be an ion[64], a small organic molecule[65], or even a larger biomolecule such as DNA[66]. Host-guest systems have many applications, including drug delivery[67], catalysis[68], separation[69] and purification[70], and sensing[71]. They are particularly interested in supramolecular chemistry, which focuses on the design and synthesis of complex molecular systems through non-covalent interactions. MD simulations can also be used to optimize host-guest systems for specific applications, such as drug delivery[72] or catalysis[73]. Using MD simulations, researchers can design more effective and efficient systems by understanding the underlying mechanisms of host-guest interactions.

Host-guest systems are attractive models for understanding the mechanics of noncovalent binding due to their small size and simple chemistry.[74] They are helpful for both experimental and computational studies and can be used to evaluate force fields.[54, 58] The host molecules act as small-scale receptors, and their binding affinity for guest molecules is similar to that of protein-ligand systems. Additionally, they share many characteristics with their larger biomolecular counterparts. Using host-guest systems as miniature models can help evaluate and improve force fields before they are applied to more complex biomolecular systems.[57] With the emergence of new options, it can be challenging to determine which force field to use in molecular dynamics simulations for biomolecular systems. Simulators often rely on the most widely used or well-established force fields, but the ability of these force fields to accurately reproduce experimental data is a source of debate.[75] While protein-ligand systems can provide an ideal benchmark for testing force fields, the complexity of these systems, including factors such as simulation conditions, long-term protein dynamics, and pH effects, can make it difficult to conduct these tests. On the other hand, host-guest systems offer a simpler benchmark

series that can alleviate many of these difficulties when assessing force fields.[76] These systems help identify appropriate force fields for MD simulations and aid in optimizing force fields for specific cases.

The SAMPL project, funded by the National Institutes of Health, is a blind crowdsourcing initiative that aims to improve and assess the accuracy of computational methods used in drug design. It has been running since 2008 and involves challenges that test the ability of these methods to predict the physical properties of small drug-like molecules and binding energies between proteins and ligands.[77] The first SAMPL challenge (SAMPL0) was run as an informal blind test for solvation-free energies of 17 small molecules.[77] It also includes occasional challenges focused on protein-ligand interactions.[78, 79] Host-guest systems provide valuable test systems for SAMPL challenges, such as solvation-free energies and binding affinities.[55, 80-84] These systems are more straightforward than proteins due to their smaller size and limited flexibility, making it easier to generate reliable computational results. They also offer the advantage of testing solute-water interactions and solute-solute interactions. They have concave binding surfaces like proteins, which can result in structured water. Host-guest systems are widely recognized for validating computational methods and gaining a deeper understanding of molecular recognition and its physical chemistry. As a result, it is not surprising that I used host-guest systems to compare force fields for the binding enthalpies before going to complex systems in this thesis.

1.4. BRD4-1 as a Test System for Computational Calorimetry

Bromodomains regulate gene expression and involve various biological processes, including cell division, development, and differentiation.[85] The human genome contains 61 bromodomains, associated with 46 different proteins belonging to eight different protein families. This diversity of bromodomains and their association with various proteins highlights the importance of these domains in regulating cellular processes. The Bromodomains and Extraterminal (BET) family has been identified as crucial targets for the modulation of protein-protein interactions.[86] BET family, including BRD2, BRD3, BRD4, and the testis-ovary specific BRDT, is known for its two repeated bromodomains, BD1 and BD2, which bind to acetylated lysine residues on histone proteins.[87] BRD4 and other BET proteins have a stronger binding capability for proteins that contain multiple acetylated residues. Due to their higher binding affinity for proteins with multiple acetylated residues, BRD4 and other BET proteins interact with highly acetylated regions of histones along the chromatin. This accumulation of transcriptionally active regulatory elements promotes gene transcription both at the initiation and elongation steps.

BRD4 has been linked to a range of diseases, such as multiple myeloma[88, 89], acute myeloid leukaemia[90], NUT carcinoma[91], Burkitt's lymphoma[92], and inflammatory diseases[93, 94]. It also serves as a prognostic indicator for metastatic breast cancer[95]. Moreover, BRD4 controls the transcription and cell cycle of HIV[96], oncogenes[97] and human papillomavirus[98]. Out of all the BET proteins, inhibiting BRD4 has become a leading therapeutic approach to lower the transcription levels of genes in a specific way for cell type and disease. Numerous efforts have been made to discover and create inhibitors of BRD4 for cancer treatment.[99, 100] Especially, the first bromodomain of the BRD4 (BRD4-1) is known to play a crucial role in various cellular processes such as cell growth, proliferation, and differentiation. Because of these roles in various cellular processes and diseases, BRD4-1 has been the subject of many biophysical and computational studies.

Biophysical studies of BRD4-1 typically focus on understanding the structural and dynamic properties of the protein, such as its 3D structure, stability, and interactions with other molecules. These studies can be conducted using techniques such as X-ray crystallography[101], NMR spectroscopy[102], and thermodynamics[103]. They can provide valuable insights into how the protein functions and how it might be targeted for

therapeutic purposes. Computational studies of BRD4-1 aim to understand the interactions with other molecules and develop methods to predict these interactions using techniques like molecular docking[104], MD simulations[105], and machine learning[106]. These studies provide important information on the binding interactions between BRD4-1 and other molecules and can be used to identify potential drug candidates.[107] For instance, the study by Lucas et al. utilized virtual screening of over 7 million small molecules to identify new inhibitors of BRD4-1 and found seven compounds with strong binding ability.[108] Through computer simulation screening, Allen et al. utilized machine learning and structure-based drug design to identify BRD4 inhibitors and a dual EGFR-BRD4 inhibitor.[109] Xue et al. employed structure-based virtual screening of 10,000 compounds, leading to the discovery of two novel BRD4-1 inhibitors.[110] Ali et al. conducted a docking-based virtual screening using a fragment-like database and discovered a novel scaffold as an inhibitor of BET.[111] These studies showcase the potential of computational techniques in finding new BRD4-1 inhibitors.

Furthermore, BRD4-1 is a valuable model for testing and building computational methods for predicting protein-ligand binding affinity due to its well-defined structure and well-characterized binding properties. Moreover, the availability of crystal structures and binding affinities for many compounds makes BRD4-1 an ideal candidate for evaluating and building computational techniques for selecting and designing ligands, particularly in predicting binding free energies. Aldeghi et al. demonstrated this by achieving a mean absolute error of 0.6 kcal/mol in their absolute binding free energy calculations for inhibitors binding to BRD4-1.[112] In addition, they applied the same method to predict the affinity of a drug-like ligand across multiple bromodomains without prior knowledge of the complex structures or ligand affinities.[113] Heinzelmann et al. used the attach-pull-release (APR) method to bind seven ligands to BRD4-1 in their free energy calculations.[114] They tested the influence of different water models and ligand parameters. On the other hand, Huggins assessed the performance of various AMBER force fields, partial charge methods, and water models for absolute binding free energy calculations using three benchmark sets, including BRD4-1.[115] Dickson et al. introduced an efficient implementation of adaptive biasing potential (ABP) in molecular dynamics simulations to improve the speed of free energy computation and tested it using BRD4-1.[116] Therefore, given the wealth of biophysical and computational data and our

lab's familiarity with this system, I was motivated to utilize it in binding enthalpy calculations using BRD4-1 in Chapter 4.

1.5. Protein-Protein Systems

Interactions between biological macromolecules are crucial for all biological processes in the cell. Protein-protein interactions (PPIs) play a vital role in biological processes, as they are fundamental for many essential functions like signal transduction[117, 118], transcriptional regulation[119], and the inhibition of enzymes[120]. Some common examples of PPIs are hormone-receptor complexes, protease-inhibitor interactions, or antibody-antigen interactions. Even weak PPIs can have a functional significance for the cell during the transmission of signals or the regulation of metabolism. This field has gained significant attention in drug discovery and is being heavily researched experimentally and computationally in the biomedical sciences[121-123]. A comprehensive understanding of the key elements in PPIs and the precise calculation of their binding free energies could result in more accurate predictions of protein-protein associations and assist in understanding cellular pathways and the impact of crucial mutations. Thus, comprehending PPIs and recognizing specific interactions are of both fundamental and practical significance.

A comprehensive understanding of cellular processes necessitates knowledge of all potential PPIs and a quantitative understanding of the structure and stability of the protein-protein complexes. Theoretical and computational techniques offer a molecular perspective of the structural and thermodynamics of PPIs. These techniques are able to reveal the basis of binding through the examination of contributions from electrostatic and van der Waals forces, as well as solvation. Additionally, rigorous computational methods can be utilized to direct new experimental studies. Despite the advancements in computational methods, predicting the thermodynamics of PPIs is difficult.[124, 125] There is often a low correlation between predicted and experimentally determined values. The poor correlation is due to the accuracy of experimental data quality and computational methods. Frequently, experimental values reported for the same PPIs by different researchers do not match. This is usually caused by different experimental conditions or techniques that are not adequately reported in the corresponding publications. On the other hand, computational methods are affected by structural imperfections, insufficient sampling, an inability to include appropriate experimental conditions, and so on.

The forces behind PPIs are linked to the change in binding affinity associated with the structural and chemical properties of the binding partners. It has been commonly observed

that binding partners undergo significant structural changes upon the association, leading to the "induced fit" concept of binding partners. During binding, proteins appear to trigger structural changes in their partners essential for forming specific complexes. To accurately compute protein-protein binding free energy, it is crucial to consider structural changes and the flexibility of the structure (conformational entropy) during binding. The computation of protein-protein binding affinity typically demands the experimental 3D structure or a predicted model of the complex. Despite the recent progress, experimentally determining protein-protein structures remains a difficult task, and it will be extremely difficult to predict all potential and transient PPIs in the cell in the near future. A further complication arises because many PPIs involve structural changes or disordered segments' simultaneous folding or unfolding.[126] Specifically, transient PPIs in cells often contain disordered protein regions. Such situations worsen the estimation of complex structures and the prediction of binding free energies.

A variety of methods can be employed to measure experimental binding affinities. It is crucial to have accurate and consistent experimental reference data to assess the accuracy of methods for predicting protein-protein binding affinities. It has been discovered that the experimental measurement of protein-protein binding affinities can vary greatly depending on the method used. Furthermore, the data for protein complexes is obtained under varying experimental conditions, such as differences in ionic strength, pH, and temperature. Conversely, computational techniques generally assume identical conditions for all complexes being evaluated. Despite this, curated benchmark sets for protein-protein binding affinity include both the complexes' structures and the unbound protein partners. Databases with thermodynamic information on protein-protein complexes obtained from experiments are crucial in gaining insights into the factors affecting binding affinity and developing effective prediction methods. One example of such databases is the Protein-protein Interactions Thermodynamic Database (PINT), which includes information such as the dissociation constant (Kd), binding free energy (ΔG), enthalpy, heat capacity change, as well as other relevant details such as experimental conditions, sequences, structures, and references in the literature.[127] These benchmark sets can be used as a basis for evaluating computational methods; however, it is important to consider the experimental conditions and the methods used to determine binding affinity in comparison to the computational setup. Kastritis and Bonvin assessed the predictive ability of existing affinity prediction methods and found poor performance on a validation set. They

attributed this to inaccuracies in the experimental data and emphasized the need to develop efficient and reliable experimental methods to obtain high-quality data.[125, 128]

Precise calculation of the binding free energy between proteins could enhance the prediction of PPIs and shed light on cellular processes and the impact of mutations. It is crucial to gain insight into the thermodynamics of protein-protein interaction in order to comprehend the basics of the molecular recognition process. However, there is a limited amount of computational analysis on the thermodynamic signatures of PPIs. Several methods have been used for computing protein-protein binding free energies, and they can be grouped into three categories.[124, 125, 129] The first group is free energy simulation methods, which explicitly use an atomic force field to model solute and solvent atoms. These methods have a solid statistical mechanics foundation. They can accurately compute the difference in binding free energy but are less reliable for ligand-protein or PPIs due to difficulty sampling relevant problems. The second group is empirical methods that use parameters from a "training set" of known interactions, taking the binding free energy as a sum of various terms such as hydrophobic contacts, hydrogen bonds, and conformational entropy loss. These methods are fast but rely heavily on the quality of the training dataset. The third group is methods that treat solute atoms explicitly and the solvent as a continuum dielectric medium, using finite difference Poisson-Boltzmann or Generalized Born methods for electrostatic contribution and an empirical function of solvent-accessible surface area for non-electrostatic contribution.[129] These methods balance speed and accuracy by avoiding solvent conformational sampling and incorporating long-range electrostatics, ionic strength, and polarization effects. However, the limited knowledge of how two proteins bind energetically and dynamically poses a major challenge for creating small, non-peptide inhibitors to control protein-protein interactions rationally.

MD simulations enable researchers to study the behaviour and interactions of molecules at the atomic level. Additionally, they can be used to estimate thermodynamic quantities and relate them to physical interactions. While computational methods for estimating binding free energy have advanced, predicting the ΔS and ΔH components of the binding remains challenging. Although deriving the ΔS and ΔH components from the free energy function is less accurate than estimating the binding free energy, the direct method for the binding enthalpy prediction is the simplest. It provides direct insight into the physical nature. However, the direct method relies on calculating the difference in energy between

the bound and unbound states, which requires running four separate simulations and a significant amount of data to ensure reliability.

In this thesis, I also investigate the calculation of absolute binding enthalpy through MD simulations of various protein-protein complexes using the direct method. Despite the long-standing existence of this theory, its application to complex systems beyond host-guest systems has been rare, possibly due to the assumption that the error from sampling would make the results challenging to interpret. However, given the advances in computational power and access to high-quality biophysical data, I believe it is worthwhile to assess the performance of this method for protein-protein systems and to highlight areas for improvement to the scientific community.

1.6. Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is the most precise and direct technique for measuring the enthalpy under constant temperature and pressure.[130] This approach stands out as the sole technique capable of concurrently acquiring the enthalpy (ΔH), entropy (TΔS), and binding free energy (ΔG) in a single titration trial. ITC has been extensively utilized in areas like pharmaceutical discovery to quantify the thermodynamic factors of molecular interactions.[131, 132]

An ITC analysis simultaneously ascertains both ΔG and ΔH for the binding reaction. Using these components, TΔS can be computed via the fundamental equation $\Delta G = \Delta H - T\Delta S$. The notable feature of ITC is that it enables the highly accurate calculation of thermodynamic variables without needing further chemical alterations like labelling or immobilization. In an ITC, two compounds (in this case, the protein and the ligand) are titrated at known concentrations at a given temperature, and the heat signal is directly quantified. The cumulative heat signal yields insights into the binding enthalpy (ΔH) and the stoichiometry. Additionally, the equilibrium binding constants (K_a or K_d) are extracted from the shape of the titration graph, facilitating the direct computation of the Gibbs free energy (ΔG) through the equation $\Delta G = -RTlnKd$, where R is the ideal gas constant, and T represents the absolute temperature.[130]

The technique involves monitoring the heat generated (exothermic event) or absorbed (endothermic event) during the interaction between a protein and a ligand. The titration process typically includes gradually introducing a concentrated ligand solution with a syringe into a solution containing the protein within the calorimeter cell (**Figure 1.2**). The protein becomes progressively saturated until all available binding sites are engaged. As a result, the heat change displays variation throughout the titration process. The outcome is a binding isotherm that accurately determines binding affinity, enthalpy, and stoichiometry, contingent upon the specific experimental conditions.

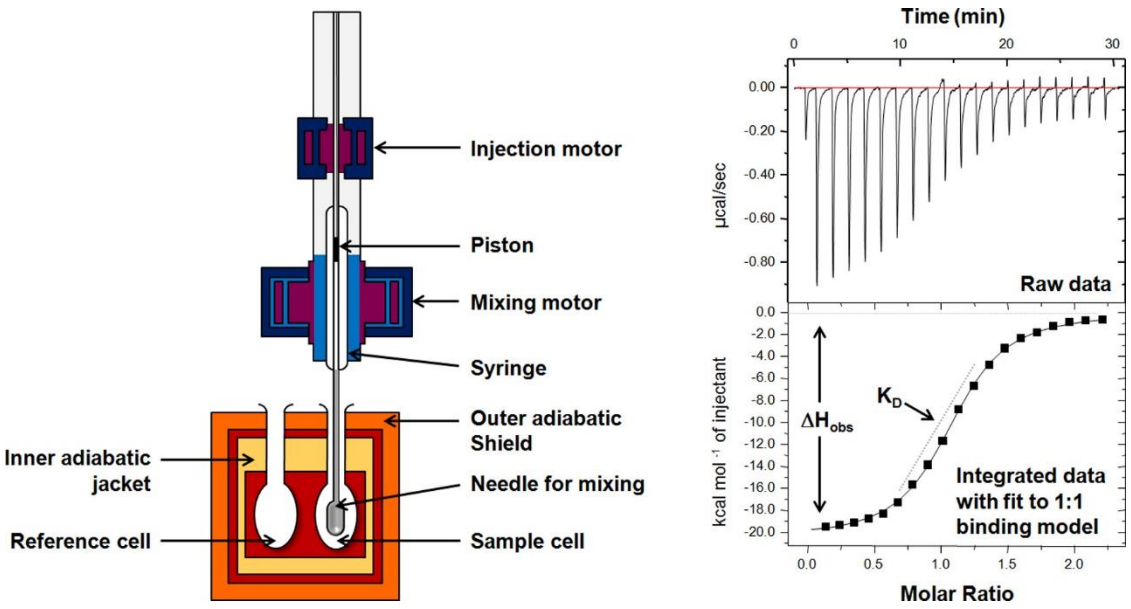

Figure 1.2: *The left part of the figure shows a schematic of an ITC. A motor-driven piston can inject precise amounts of a ligand solution into the sample cell. The upper right shows examples of raw data from multiple ligand injections into the macromolecule solution. The lower right shows the integrated binding isotherm that is generated from the raw injection data. From the slope and maximum amplitude of the binding isotherm, the binding affinity* K_d *and observed binding enthalpy* ΔH_{obs} *can be extracted. Adapted with permission from J. Med. Chem. 2015, 58, 16, 6321–6335. Copyright 2015 American Chemical Society.*

The errors in determining binding affinity are typically lower with ITC than with traditional biochemical assays, making ITC the gold standard for measuring affinity.[28] Nevertheless, some calorimetric data's precision, accuracy, and repeatability may raise concerns.[26] Some reported data may not be as precise or accurate as claimed. To ascertain the reliability of ITC data, one approach involves repeating measurements of the same reaction across multiple times within various laboratories.

The reliability of ITC data has been questioned, and its accuracy has been investigated in instances where the thermodynamic parameters of inhibitors binding to carbonic anhydrase were compared across different laboratories. These ITC results were also cross-referenced with findings from alternative techniques like surface plasmon resonance, serving as a comparative benchmark. This notable instance investigating the precision of ITC data integrity is the ABRF-MIRG'02 study.[133] This initiative involved 14 distinct laboratories independently characterizing identical samples' thermodynamics. The fact that these measurements were executed autonomously across diverse research settings highlights the robustness of ITC as a technique and provides a thorough assessment of its accuracy. The diverse sources of error impacting an ITC measurement are now well comprehended. However, despite this understanding, the errors or uncertainties inherent

in ITC measurements often go unreported.[134] The error in the ABRF-MIRG'02 study was shockingly high and considerably higher than those commonly documented for ITC measurements.[133] Researchers engaged in ITC studies examining protein-ligand interactions should validate their ITC accuracy.

The primary limitation of ITC is the low protein concentration in aqueous solution, making it problematic to derive precise data from ITC. Another limitation of the technique is its suitability solely for detecting interactions with an enthalpic contribution, as pure entropic interactions cannot be observed.[26]

In the thesis, I utilized ITC data from the literature to compare the binding enthalpy calculations with experimental enthalpies. There are two primary prerequisites to guarantee the experimental information represents the physics-based binding calculable from the simulations.[135] First, the quantified output should reflect or closely correlate with the binding. Second, the experimental setup and molecular system utilized in the simulation should match as nearly as possible. The first point is associated with selecting the appropriate experimental data type for comparison. Ideally, biophysical binding data is determined from ITC. I chose ITC data with a similar experimental setup since different ITC setups may produce significant errors compared to experimental data to compare computed binding enthalpies.

1.7. Aims and Motivation

This thesis focuses on examining the calculation of absolute binding enthalpy through MD simulations. Its main goal is to assess the direct method for its use in complex systems for drug design purposes. With the advancement of computer power, MD simulations based on statistical mechanics have become increasingly accessible, as seen in the successful application of relative binding enthalpy calculations in complex systems containing proteins.[34, 35] However, the absolute binding enthalpy calculation based on the direct method has not been applied to large conformational systems due to its demanding computational requirements.

Despite the limitations of large-scale performance analysis, it is crucial to evaluate the technique's potential through careful retrospective testing. This evaluation will provide insight into the accuracy and precision of the technique and identify potential sources of errors that need to be addressed for further advancements.

In Chapter 3, the performance of various water models and force field schemes on absolute binding enthalpy will be examined using host-guest systems. Chapter 4 will evaluate the absolute binding enthalpy calculations for ten inhibitors against the bromodomain BRD4-1. Finally, in Chapter 5, the direct method will be applied to calculate the binding enthalpy of protein-protein interactions using a set of 11 protein-peptide complexes with available structural and thermodynamic data.

2 Theory and Methods

This chapter discusses the use of computer simulation techniques and provides a brief overview of the fundamental and technical concepts. It is not possible to fully describe all methods in this chapter due to more comprehensive coverage and the availability of comprehensive textbooks on the subject.[14, 136-140]

2.1. Molecular Dynamics Simulations

The molecular dynamics (MD) simulations describe the movement of atoms and acquire the atomistic system's dynamic properties using classical Newtonian mechanics.[136, 138] MD simulations accurately calculate atomic motions based on physical theory. So, MD simulations can simultaneously provide a system's statistical and dynamic properties and be used in various molecular systems. MD simulations are less complicated and computationally efficient than quantum mechanics (QM) calculations. MD enables simulations of more complex molecular systems with the increased processing capability of computational architectures.

2.1.1. Theoretical Foundation

N number of atoms in a molecular system have vectors for their positions and momentums assigned by $r_i = (x_i,\ y_i,\ z_i)$ and $p_i = \left(p_{i,x},\ p_{i,y},\ p_{i,z}\right)$, respectively. The Hamiltonian H of a system, corresponding to the total energy of that system, is written as

$$H(R^N, P^N) = \sum_{i}^{N} \sum_{\alpha} \frac{p_{i,\alpha}^2}{2m_i} + U(R^N) \tag{2.1}$$

R and P represent coordinates and momenta for N number of atoms in the system, respectively. The first term describes the system's kinetic energy, while the second describes its potential energy. α represents the *x*, *y*, and *z* directions; m_i is the mass of the ith atom. The force acting on the system is conserved, allowing the force as a vector on each particle to be determined by taking the derivative of the system's potential energy with the particle's position. The direction of the derivative of the potential energy's first

order reveals the position of the lowest energy, and the gradient's magnitude indicates the steepness of the local slope. The system's energy can be decreased by shifting each particle according to the force affecting it. As a result, the force is equal to the negative gradient and is written as[138, 141]

$$F_i(R^N) = -\frac{\partial U(R^N)}{\partial r_i} \tag{2.2}$$

Thus, the motion of an atomic particle can be defined by Newton's second law as

$$m_i \ddot{r}_i = F_i(R^N) \tag{2.3}$$

where $\ddot{r}_i$ represents the second derivative of r_i to time. The integration of Eq. (2.3) to time, using the initial R^N and P^N of particles, produces their trajectories. MD simulation is deterministic, meaning that the subsequent time evolution can be determined given predefined initial conditions. Solving the initial value problem is demonstrated by looking at a basic one-dimensional (1D) harmonic oscillator. The classical Hamiltonian for an oscillator is written as

$$H = \frac{p(t)^2}{2m} + \frac{1}{2} k_s x(t)^2 \tag{2.4}$$

The Hamiltonian for this oscillator is a function of its momentum ($p(t)$), position ($x(t)$), mass (m), and spring constant (k_s). First-order differential equations can describe it to time:

$$\frac{dx}{dt} = \frac{\partial H}{\partial p} = \frac{p}{m}, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial x} = -k_s x \tag{2.5}$$

The oscillator's path is calculated numerically using the finite difference method, which transforms the differential equations into algebraic equations.[138] Using the forward difference scheme with an integration timestep of Eq. (2.5) can be rewritten as

$$\frac{dx}{dt} = \frac{x(t+\Delta t) - x(t)}{\Delta t} = \frac{p(t)}{m}, \qquad \frac{dp}{dt} = -\frac{p(t+\Delta t) - p(t)}{\Delta t} = -k_s x(t) \tag{2.6}$$

Eq. (2.6) can be easily manipulated to solve for Eq. (2.4) as

$$x(t+\Delta t) = x(t) + \frac{p(t)\Delta t}{m}, p(t+\Delta t) = p(t) - k_s x(t)\Delta t \tag{2.7}$$

Using Eq. (2.7) and the initial position and momentum of the oscillator, we can find the oscillator's position and momentum at each point in time through a series of recurrence equations.

Solving Eq. (2.1) in MD simulations is more complex than solving Eq. (2.4) because all interatomic interactions in the system influence each atom's movement. This is a many-body problem with 6N variables (3N positions and 3N momenta) for a system with N particles.[138] It is almost impossible to find an analytical solution because they are closely connected. However, the finite difference method can be used to decouple the 6N variables. The crucial step in solving the Hamiltonian in Eq. (2.1) is to find an expression for $F_i(R^N)$ in Eq. (2.3). In MD simulations, the interactions between atoms are described using a force field that has an associated potential energy function $U(R^N)$, and $F_i(R^N)$ is determined through the use of Eq. (2.2). In MD simulations, atomic trajectories (the evolution of R^N and P^N) are often not the final product. Instead, physical quantities of the system are often deduced from these trajectories. This is possible because MD simulations are statistical mechanical approaches, even though classical mechanics describes the atoms' movement. The atomic trajectories can be seen as arrangements that follow a specific statistical pattern. For instance, in a situation where the temperature is constant, like, in a canonical ensemble, the likelihood of the system's arrangements is proportional to the Boltzmann distribution.

$$\exp\left[-\frac{H(R^N, P^N)}{k_B T}\right] \tag{2.8}$$

where k_B is the Boltzmann constant (1.380649×10^{-23} *J/K*), and T indicates the temperature. Therefore, physical properties are calculated by averaging over all possible configurations. Arithmetically, a physical quantity $Q(R^N, P^N)$ can be expressed as

$$\langle Q(R^N, P^N)\rangle = \int dR^N \int dP^N \, p(R^N, P^N) Q(R^N, P^N) \tag{2.9}$$

where $p(R^N, P^N)$ represents the probability density distribution (PDF) of the system's arrangements. The angle bracket signifies that the physical quantity is calculated as the average over all possible arrangements of the system specified by $R^N$and P^N. The normalized Boltzmann function gives the PDF for configurations, which is a function of the Hamiltonian.[138]

$$p(R^N, P^N) = \frac{\exp\left[-\frac{H(R^N, P^N)}{k_B T}\right]}{\int dR^N \int dP^N \exp\left[-\frac{H(R^N, P^N)}{k_B T}\right]} \tag{2.10}$$

It is almost impossible to sample this phase space fully. The ergodic hypothesis states that the system can encounter all potential states in space if enough time is provided. As a result, quantities averaged over the phase space can be estimated by their equivalents obtained through averaging over time as

$$\langle Q(R^N, P^N)\rangle = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^{\tau} Q(R^N, P^N) dt \tag{2.11}$$

Practically, MD simulation cannot continue forever, but ensuring that the phase space is properly sampled is essential. MD simulation struggles to model a system with many atoms because of the high demands for computation and data storage resources. To increase the precision of the modelled system and decrease statistical errors, the simulation time step should be made as small as possible.

2.1.2. Technical Aspects

In recent years, various techniques for enhancing MD simulation have been developed and accumulated by researchers. These techniques are now critical components of numerous widely used modern MD software programs.[142-146] This section will offer brief information on the fundamentals of MD simulation, including fundamental techniques such as periodic boundary conditions (PBCs), time integration algorithms, and neighbour lists.

Procedure

The typical procedure for conducting a classical MD simulation is outlined in the flowchart (**Figure 2.1**).

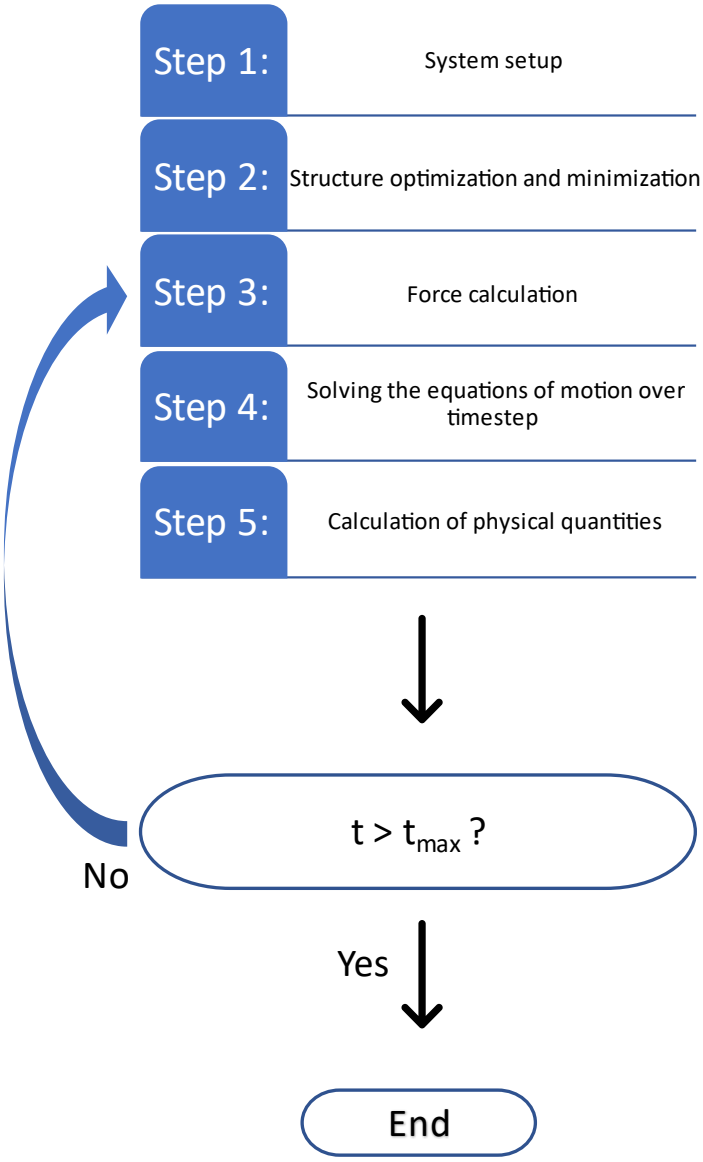

Figure 2.1: *Flowchart of MD simulation.*

Step 1 involves setting up the system's initial conditions, such as the positions, velocities, and constraints of atoms, and defining the potential energy function. Step 2 involves optimizing the starting structure to reflect an equilibrium state based on the chosen possible energy function. Structural optimization and energy minimization results in a more effective "starting point" of the simulation and is generally helpful for successfully executing a simulation. Step 3 involves calculating the forces on each particle based on

the chosen potential energy function and then solving the equation of motion for each particle over a given timestep Δt. Step 4 involves updating the particle positions and velocities. In step 5, specific physical properties are computed using information from step 4. The simulation continues by repeating steps 3-5 until the pre-set maximum simulation time is reached or certain conditions are met; at this point, the simulation will end. Step 2 may not always be necessary and can sometimes be omitted. For instance, if an MD simulation is being resumed from a previous simulation where the system has stabilized, there would be no need for structural optimization or energy minimization. An MD simulation can also yield accurate results without these steps if the starting atomic structure is already nearly equal to the equilibrium structure that corresponds to the chosen potential energy function and the simulation runs smoothly. However, this scenario is unlikely to occur as initial atomic structures are typically obtained from experiments or previous MD simulations that utilized different potential functions or were imposed with varying constraints.[138]

System Initialization

An MD box must be defined to run a simulation, and each particle's initial positions and velocities must be initialized. There are two ways to initiate an MD simulation: starting from scratch and continuing from a previous simulation. Initial velocities can be set to any value as long as the system remains stable. At the end of a simulation, the velocities of the atoms will be adjusted to suitable values based on the simulation conditions and specific settings.

Periodic Boundary Conditions

An issue that arises during the initialization stage of an MD simulation is how to establish the boundary conditions of the simulation box, which is of finite size. The dynamic behaviours of atoms or molecules located at the boundaries differ significantly because they have fewer neighbouring atoms or molecules than those situated inside the box. To accurately simulate the macroscopic scale, it is generally best to avoid surface effects, which can cause deviation from real results and depend on the ratio of surface atoms to total atoms. This is especially important when simulating a cluster of atoms. PBCs are often used in MD simulations to avoid surface effects. The simulation space is filled with atoms to simulate macroscopic materials by selecting a primary simulation box from a

representative atomic system, avoiding surface atoms (**Figure 2.2a**), and then repeating and arranging the primary cell in all three dimensions (**Figure 2.2b**).

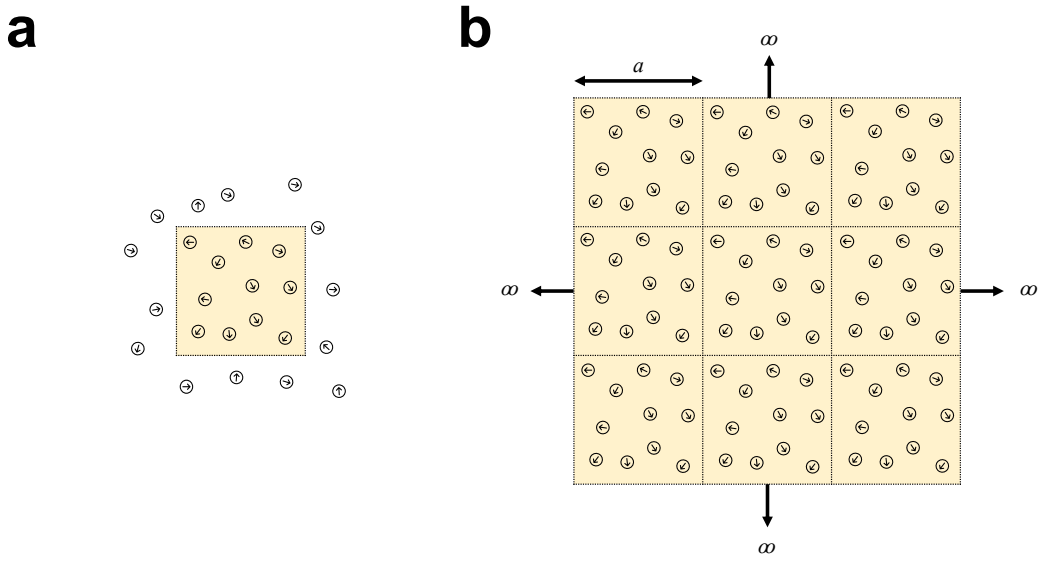

Figure 2.2: *A realistic atomic system versus A simulated atomic system* ***(a)*** *A realistic atomic system with black arrows indicating the movement direction of atoms. The atoms enclosed by dotted lines form a simulation cell.* ***(b)*** *A simulated atomic system is established by repeating the simulation cell in all three dimensions and extending it indefinitely.*

In actual MD simulations, the repetition of the primary cell is not carried out in a physical sense. Instead, the interactions between the representative atoms inside the primary cell and their "image" atoms outside it are calculated. At each timestep, the simulated atoms' positions are updated first, followed by the positions of the image atoms, which are updated by implementing translational periodicity. The PBCs offer two key advantages. Firstly, the surface effects are minimised by ensuring that each atom in the primary simulation cell interacts with both atoms inside and outside the cell. Secondly, as atoms move in and out of the simulation cell, the number of atoms remains constant, which is essential for MD simulations.[138]

A problem with PBCs is that the number of interactions becomes arbitrarily large due to infinite image atoms. However, in actual simulations, most atomic interactions diminish rapidly with distance and can be cut off beyond a specific critical distance R_c. In some cases, the *i*th atom in the primary simulation cell can interact with atom *j* and multiple of its image atoms if they are within a distance of R_c, resulting in increased complexity in the MD simulation and potentially unrealistic outcomes. The minimum image criterion is applied to fix this, stating that the *i*th atom only interacts with the nearest atom *j* or one of

its images. **Figure 2.3** shows a simulation box containing atoms A, B, C, D, E, and F. Eight replicated images surround the initial simulation system. To determine the force acting on atom A, the contributions of atoms C and D and the image versions of atoms B and E from the adjacent cells must be taken into account, as all of atom A's neighbours within the critical radius R_c are located outside of the primary cell.

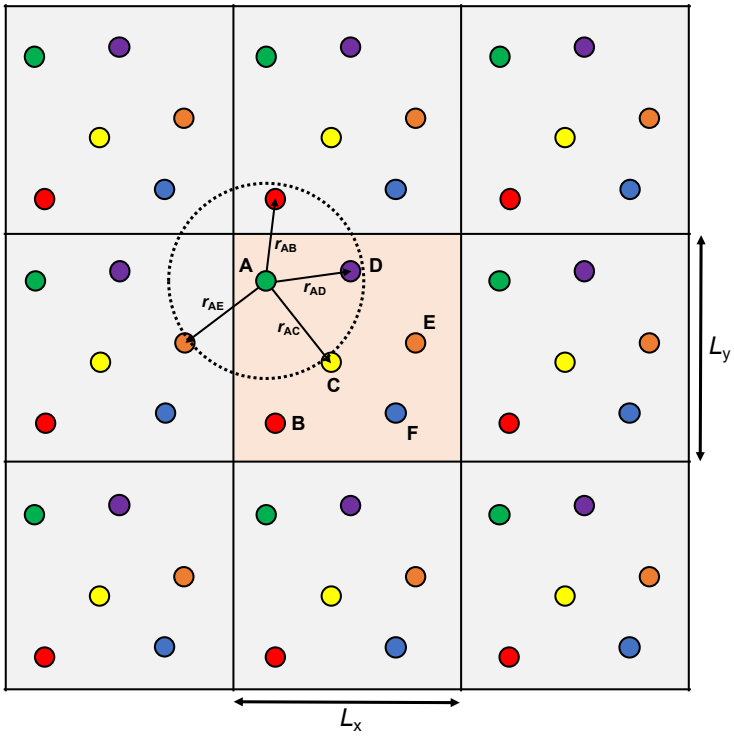

Figure 2.3: *The minimum image criterion is used to identify atom A's closest neighbours in the central simulation cell. The size of each simulation cell is specified as $L_x \times L_y$. Any atoms within the critical distance R_c from atom A are deemed its closest neighbours and can interact with it.*

Energy Minimization and Structure Optimization

The starting structure built from scratch typically deviates from the balanced structure that is calculated using the given potential energy function. Before the simulation, structural minimization is usually required. In some instances, MD simulations aim to get the most stable atomic conformation. In these situations, an energy minimization procedure produces the most stable or the lowest energy state. In the process of energy minimization, only the potential energy $U(R^N)$ of the system at 0 K is considered, not its kinetic energy. The focus is on the lowest energy state of the system. A molecular system's potential energy is based on its N atoms' position. One common approach to finding a function's minimum is to utilize its derivatives. In MD simulations, the steepest descent and conjugate gradient method are the most widely employed techniques for finding the

minimum energy by utilizing the first derivative of potential energy. These methods involve calculating the potential energy for the initial structure, then moving each atom in a specific direction in small steps to maximize the decrease in potential energy, continuing until a minimum potential energy is found. The main difference between the two methods is how they determine the direction of movement. In practical MD simulations, the energy minimization process has two stages. In the first stage, derivative techniques like the steepest descent and conjugate gradient methods are employed to swiftly decrease the starting system's high energy. Second derivative methods are used in the second stage to refine the configuration of the atoms further to find the lowest minimum energy state. It's not possible to reach the true minimum energy state in MD simulations, only estimates. There are two methods for determining the termination of the energy minimization process. The first is when the energy difference between two consecutive iteration steps falls below a specified value, signifying that the minimum energy has been reached. The second is based on the magnitude of atomic movements during two successive steps.

Force Calculation

The force on an atom is found by calculating the first derivative of potential energy using a given potential energy function, as shown in Eq. (2.2).[138] Analytical expressions for the first derivative of potential energy are available for many common potential energy functions used in MD simulations and are implemented in most MD codes. A chain molecule is utilized to demonstrate the force calculation procedure in MD simulations (**Figure 2.4**). Its potential energy arises from bond stretching, angle bending and torsion, and interactions with other molecules, as depicted in **Figure 2.4**.

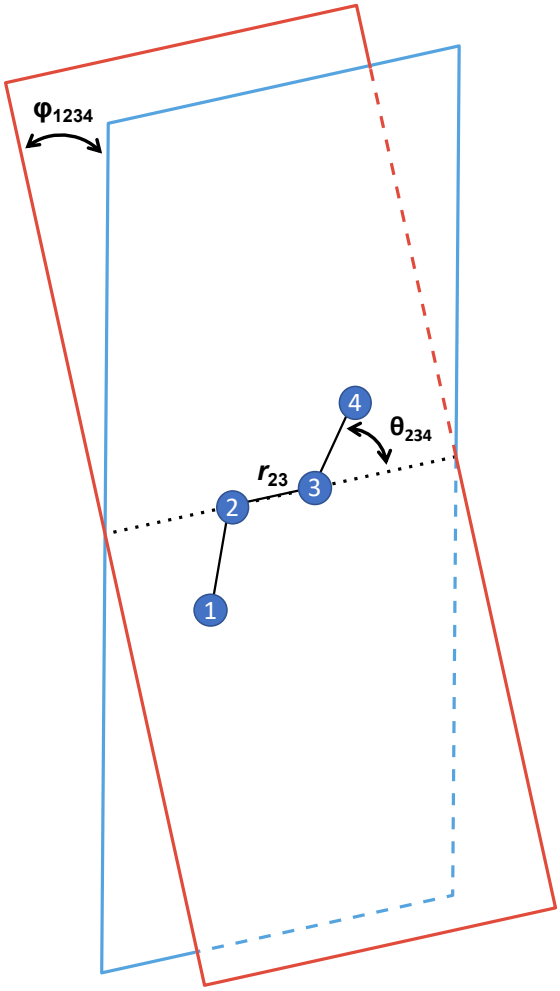

Figure 2.4: *A simple chain molecule with labelled interatomic distance r_{23}, bond angle θ_{234}, and torsion angle ϕ_{1234}.*

The bond bending potential component is written as:

$$U_{bond-bending} = \sum_{i \neq j \neq k} k_{ijk}^{\theta} cos\theta_{ijk} \tag{2.12}$$

where k_{ijk}^{θ} is a constant. The angle is between two consecutive bond vectors and is represented by:

$$cos\theta_{ijk} = \left(r_{ij}.r_{ij}\right)^{-\frac{1}{2}}\left(r_{jk}.r_{jk}\right)^{-\frac{1}{2}}\left(r_{ij}.r_{jk}\right) \tag{2.13}$$

Here, I define the bond vector as $r_i - r_j$ rather than $r_j - r_i$.

The chain consisting of atoms 2, 3, and 4 is utilized as an illustration. To determine the forces resulting from the bending of the angle θ_{234} on these atoms, the derivatives listed below are required based on Equation (2.2):

$$\begin{aligned}\frac{\partial}{\partial r_2}(r_{23}.r_{34}) = r_{34}, \frac{\partial}{\partial r_3}(r_{23}.r_{34}) = r_{23} - r_{34}, \frac{\partial}{\partial r_4}(r_{23}.r_{34}) \\ = -r_{23}, \frac{\partial}{\partial r_2}(r_{23}.r_{23}) = 2r_{23}, \frac{\partial}{\partial r_3}(r_{23}.r_{23}) \\ = -2r_{23}, \frac{\partial}{\partial r_4}(r_{23}.r_{23}) = 0\end{aligned} \tag{2.14}$$

The forces on atoms 2, 3, and 4 arise from the angle bending based on the following:

$$\begin{aligned}f_2 = \frac{\partial}{\partial r_2} k_{234}^{\theta} cos\theta_{234} = k_{234}^{\theta} r_{23}^{-1} r_{34}^{-1} \left(r_{34} - \frac{r_{23}.r_{34}}{r_{23}^2} r_{23} \right), \\ f_3 = \frac{\partial}{\partial r_3} k_{234}^{\theta} cos\theta_{234} \\ = k_{234}^{\theta} r_{23}^{-1} r_{34}^{-1} \left(\frac{r_{23}.r_{34}}{r_{23}^2} r_{23} - \frac{r_{23}.r_{34}}{r_{34}^2} r_{34} + r_{23} - r_{34} \right), \\ f_4 = \frac{\partial}{\partial r_4} k_{234}^{\theta} cos\theta_{234} = k_{234}^{\theta} r_{23}^{-1} r_{34}^{-1} \left(\frac{r_{23}.r_{34}}{r_{34}^2} r_{34} - r_{23} \right)\end{aligned} \tag{2.15}$$

The forces resulting from the torsion of the bond between atoms 2 and 3 can be obtained through a comparable method for all four atoms involved (1, 2, 3, and 4). The example mentioned above encompasses the covalent interaction of connected atoms, a type of short-range interaction. The total force on an atom from short-range interactions can be found by summing the forces from a certain number of nearby atoms. For long-range interactions, the force calculation must account for the interaction between all pairs of atoms. The computing demand rises significantly with system size. Atomic interactions tend to decline rapidly with increasing interatomic distances. To minimize the computational cost of force calculation, the interaction between two atoms is often terminated at a specified distance.[138]

Smooth Particle Mesh Ewald

The accuracy and computational expense of biomolecular simulations are frequently dependent on the electrostatic interactions between atoms. Calculating the Coulomb forces between all possible atom pairs to model the long-range electrostatic effects is often prohibitively expensive for typical biomolecular systems. Therefore, many approximate methods have been developed, including techniques based on distance cutoffs, multipole approximations, and approaches utilizing Ewald summation. Choosing an appropriate electrostatics approximation scheme is crucial for obtaining accurate results in a feasible timeframe when performing computational simulations of biomolecules.

The Ewald decomposition approximates the infinite summation over all Coulomb interaction pairs by two finite sums. These summations intelligently account for all pairwise electrostatic interactions and are thoughtfully structured to achieve rapid and smooth convergence.[147] The electrostatic energy E for a defined system containing N distinct electric charges q_{ij} positioned at r_{ij} inside a periodic cell with a volume V can be calculated as

$$E = \frac{1}{2}\sum_{k}^{\infty}\sum_{i,j}^{N}{}' \frac{q_i q_j}{|r_i - r_j - k|} \tag{2.16}$$

where k represents a vector that characterizes the translation connecting a set of periodic unit cells. Using two finite sums, the Ewald summation method estimates the infinite sum over all Coulomb interactions between pairs of atoms. These selected sums account for all electrostatic pairs in a smooth and quickly convergent manner. The smooth particle-mesh Ewald (SPME) technique utilizes such an approximation.[148] SPME provides an efficient and reliable approach to modelling long-range electrostatics in large biomolecular systems. We express the following using the Ewald decomposition technique per the process outlined in the SPME derivation.

$$E = E_r + E_k + E_{corr} \tag{2.17}$$

$$E_r = \frac{1}{2}\sum_{k}^{\infty}\sum_{i,j}^{N}\frac{q_i q_j erfc\left(\beta|r_{ij} + k|\right)}{|r_{ij} + k|} \tag{2.18}$$

$$E_k = \frac{4\pi}{V}\sum_{k\neq 0}^{\infty}\sum_{i,j}^{N}\frac{q_i q_j}{|k|^2}e^{i\mathbf{k}.r_{ij}-\frac{|k|^2}{4\beta^2}} \tag{2.19}$$

$$E_{corr} = -\frac{1}{2}\sum_{(i,j)\in M}\frac{q_i q_j\,\mathrm{erf}(\beta|r_{ij}|)}{|r_{ij}|} - \frac{\beta}{\sqrt{\pi}}\sum_{i}^{N} q_i^2 \tag{2.20}$$

The vectors r_{ij} refers to the distance vector between particles i and j. β is a parameter called the Ewald width parameter. The Ewald summation splits the electrostatic energy into the components of real-space (r-space) and reciprocal-space (k-space). There is also a correction term E_{corr} that accounts for double counting, but this has a negligible effect on the forces and accuracy so can be ignored for this discussion. The r-space sum converges quickly with increasing $|r_{ij}|$ distance between particle pairs thanks to the Ewald decomposition. The SPME implementation makes a list of particle pairs within a cutoff distance r_c and only calculates interactions between these pairs. To obey the minimum image convention in GROMACS, r_c must be less than half the smallest box dimension.[147] This means that no term of E_r with k≠0 contribute, the r-space part reduces to a simple sum over the short-range pair interactions.

$$E_r = \frac{1}{2}\sum_{i,j|(i,j)\notin M, r_{ij}<r_c}\frac{q_i q_j(\beta r_{ij})}{r_{ij}} \tag{2.21}$$

Truncating r-space sum at the cutoff distance r_c introduces an approximation in calculating energies and forces. The quality of this approximation improves as the Ewald parameter width β and the cutoff radius r_c increase. However, increasing r_c means more particle interactions must be computed, so there is a trade-off between accuracy and speed. In GROMACS, the cost of computing van der Waals interactions can also rise with r_c. The pair list's minimum size handles van der Waals and SPME r-space.[147]

Time Integration Algorithms

The MD simulation program is powered by a time integration algorithm that calculates the movement of interacting atoms by integrating the equations of motion, resulting in atomic trajectories. The finite difference method serves as the foundation for it, dividing time into

smaller segments with a set distance between each point, known as the timestep, on a grid. The integration scheme takes in the atomic positions, velocities, and accelerations at a specific time, *t*, and calculates these same quantities at a future time. The Verlet algorithm is commonly used in MD simulations. The atomic position in the original Verlet method is expressed using a second-order Taylor polynomial.[149] The Verlet method is efficient in terms of force calculations and time-reversible but can introduce errors with small timesteps and cannot couple to an external thermal bath due to the independence of atomic positions and velocities.

Verlet method was later developed the leap-frog method, to address the limitations of the original Verlet method.[150] It calculates velocities and positions as:

$$v\left(t+\frac{1}{2}\Delta t\right)=v\left(t-\frac{1}{2}\Delta t\right)+a(t)\Delta t \tag{2.22}$$

$$r(t+\Delta t)=r(t)+v\left(t+\frac{1}{2}\Delta t\right)\Delta t \tag{2.23}$$

The acceleration vectors $a(t)$ are obtained by calculating the forces on the atoms based on their positions at time *t*. The velocity at time $t+\frac{\Delta t}{2}$ is obtained using the velocity at time $t-\frac{\Delta t}{2}$ and Eq. (2.22). The velocity at time t is then calculated using the obtained velocity at time $t+\frac{\Delta t}{2}$ as

$$v(t)=\frac{1}{2}v\left(t+\frac{1}{2}\Delta t\right)+\frac{1}{2}v\left(t-\frac{1}{2}\Delta t\right) \tag{2.24}$$

The leap-frog method, which is more precise than the original Verlet method, enables the MD system to be connected to an external thermal bath. Additionally, it results in a dependent atomic trajectory. However, it requires more computing and storage resources and has a lag in velocity updates. The velocity Verlet method is an improvement on the leap-frog method, and simultaneously calculates positions and velocities at time $t+\frac{\Delta t}{2}$ using their values at time t using the following equations.[138, 151]

$$r(t+\Delta t)=r(t)+v(t)\Delta t+\frac{1}{2}a(t)\Delta t^2 \tag{2.25}$$

$$v\left(t+\frac{1}{2}\Delta t\right)=v(t)+\frac{1}{2}a(t)\Delta t \tag{2.26}$$

$$v(t+\Delta t)=v\left(t+\frac{1}{2}\Delta t\right)+\frac{1}{2}a\left(t+\frac{1}{2}\Delta t\right)\Delta t \tag{2.27}$$

Neighbour List

Nonbonded atomic interactions can be limited to a certain distance to lower computing expenses. This technique is useful for large-scale MD simulations; however, all atomic pairs require distance calculations. Calculating pair separation distances in an *N*-particle system takes time proportional to $\frac{N(N-1)}{2}$. In MD simulations, the timestep is typically small, and the movement of atoms is limited, so the neighbours of a given atom often remain unchanged. To save time, a neighbour list can be established and updated for each atom, allowing distance and force calculations to be done only between an atom and the atoms on its neighbour list.

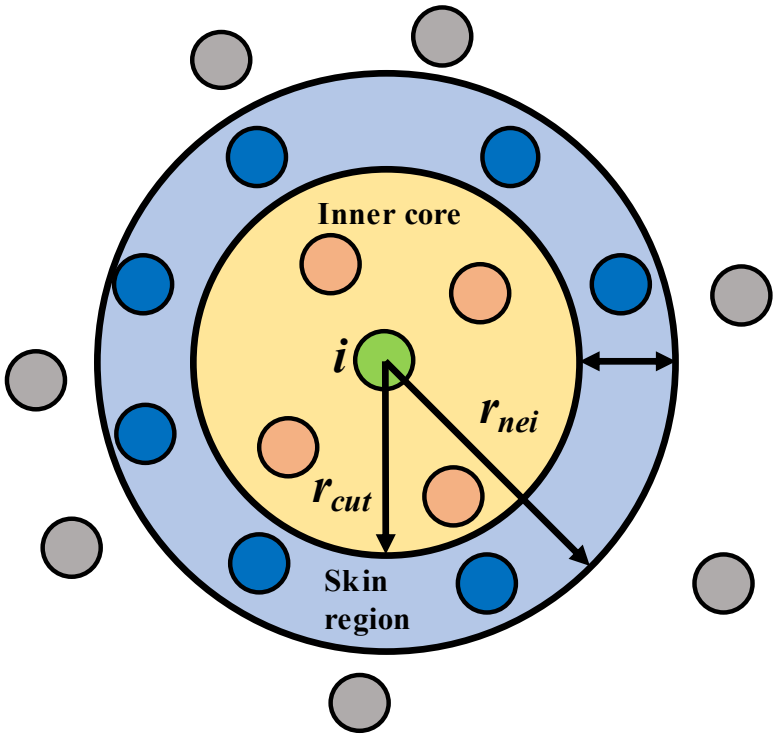

Figure 2.5: *The diagram shows atoms in proximity to atom i using the Verlet neighbour list method. It displays the inner core area (represented by a circle with radius r_{cut}), the skin region (the area outside the inner core but within a larger circle with radius r_{nei}), as well as the atoms located in each region and the atoms that are excluded from the neighbour list. Green circles represent atom i, orange circles represent atoms in the inner core, blue circles represent atoms in the skin region, and grey circles represent atoms that have been omitted from the neighbour list.*

Two methods exist for generating a list of neighbours. The first, called the Verlet method, involves defining a distance for the list based on the potential cutoff distance plus a skin distance. Atoms are added to the list of a given atom if the distance between them is less than this defined distance (**Figure 2.5**). At each step of an MD simulation, the distances between a given atom and the atoms on its neighbour list are computed, and their interactions with the atom are only considered if the distance is shorter than a specified cut-off distance. r_{nei} for the neighbour list is chosen based on a balance between the frequency of updates and the number of atomic pairs checked for force calculations. To update the list every N_t time steps, the skin distance must be set accordingly.

$$r_{nei} - r_{cut} > N_t \langle v_{max} \rangle \Delta t \tag{2.28}$$

The Verlet method for generating a neighbour list is easy to implement. It uses a row vector to store the neighbours of each atom. The method is likely the most straightforward option. The Verlet method's efficiency is affected by the frequency at which it updates, which is often kept low for accuracy. As a result, the number of distance calculation operations remains roughly proportional to N^2. The cell index method is more efficient and better suited for large systems with dimensions greater than r_c.[138]

Ensemble and Statistical Observables

Classical statistical thermodynamics focuses on the average behaviours or properties of a system's particles rather than individual particles. Multiple duplicates of the system are created, and the average values are computed based on these copies. The duplicates must have equivalent values for specific thermodynamic properties such as the number of atoms, temperature, pressure, volume, chemical potential, and total energy. Still, they do not have to be the same at the atomic level. A collection of independent duplicates is referred to as a statistical ensemble. There are four statistical ensembles commonly used in MD simulations: the canonical ensemble (with constant N, V, T), the microcanonical ensemble (with constant N, V, E), the isothermal-isobaric ensemble (with constant N, P, T), and the grand canonical ensemble (with constant V, T, and chemical potential). MD

simulations produce duplicates of a molecular system by capturing its microscopic states at regular intervals over an extended period of time. This is done under different conditions to measure its properties in various ensembles.

2.1.3. Control Techniques

Types of Constraints

MD simulations accurately depict space, time, and energy, which can help understand experimental phenomena and supplement experimental techniques. Certain constraints are applied in MD simulations to mimic practical experimental conditions to control the dynamic process. There are three types of constraints employed in MD simulations.[140, 152]

The first type of constraint used in MD simulations is the thermodynamic constraint, which encompasses three quantities that describe the macroscopic state of the MD system. These three quantities are typically selected from one of three pairs: (1) the number of atoms N and the chemical potential μ, (2) the volume V and the pressure P, (3) the energy E and the temperature T.[140] The particular quantities and their reference values determine the thermodynamic ensemble sampled during the simulation. In a conventional MD simulation, the number of atoms N, the volume V, and the energy E are typically held constant, corresponding to the microcanonical ensemble. Maintaining other thermodynamic quantities constant and sample other ensembles besides the microcanonical ensemble is feasible by adjusting or restricting the Hamiltonian or Newtonian equations of motion. The second type of constraint used in MD simulations relates to the system's spatial aspect and surroundings. This involves defining the shape and environment of the molecular system being simulated. Different types of spatial boundary conditions include periodic boundaries, vacuum surroundings, fixed borders, repulsive walls, and semi-rigid atomic boundaries. The third type of constraint used in MD simulations involves specific molecular geometry, such as maintaining a fixed bond length within a molecule to accurately represent its high-frequency bond vibrations.

SHAKE algorithm

The SHAKE algorithm takes a set of unconstrained coordinates r' and converts them into coordinates r'' that satisfy a given list of distance constraints.[153] It does this by using a reference set of coordinates r.

$$SHAKE(r' \rightarrow r''; r) \tag{2.29}$$

SHAKE works by solving Lagrange multipliers for the constrained equations of motion. To use SHAKE, specify a relative tolerance, the allowed deviation from the distance constraints. SHAKE will keep iterating until all the constraints are satisfied within that tolerance level. If the constraints are unmet because the variation is too large or SHAKE exceeds the maximum iterations, it will stop and output an error message. So, in essence, SHAKE repeatedly adjusts the coordinates until the distance restraints are satisfied within a given tolerance, or it fails if the constraints can't be adequately met after excessive iterations.

Presume the equations of motion need to satisfy K holonomic constraints written as:

$$\sigma_k(r_1 \dots r_N) = 0; k = 1 \dots K. \tag{2.30}$$

For example, $(r_1 - r_2)^2 - b^2 = 0$. Then the forces are defined as

$$-\frac{\partial}{\partial r_i}\left(V + \sum_{k=1}^{K} \lambda_k \sigma_k\right) \tag{2.31}$$

The equations of motion contain lambda terms λ_k, which represent Lagrange multipliers. These Lagrange multipliers must be solved to satisfy the constraint equations. The second part of the sum in the equations gives the constraint forces G_i. These constraint forces are calculated from an expression containing the Lagrange multipliers and the gradients of the constraint functions.

$$G_i = -\sum_{k=1}^{K} \lambda_k \frac{\partial \sigma_k}{\partial r_i} \tag{2.32}$$

In the leapfrog or Verlet algorithm, the displacement caused by the constraint forces is proportional to $(G_i/m_i)(\Delta t)^2$. A coupled set of quadratic equations needs to be solved to determine the Lagrange multipliers and resulting displacements. SHAKE uses an iterative approach to solve for the Lagrange multipliers and displacements required to maintain the constraints. Specifically, SHAKE iteratively finds the Lagrange multiplier values, which give the constraint forces and associated displacements to satisfy the distance restraints at each timestep.

The LINCS algorithm

LINCS algorithm corrects bond lengths after an unconstrained update to match specified constraints.[154] Unlike SHAKE, LINCS always uses just two steps rather than iterating. LINCS uses matrices internally but does not require slow matrix-matrix multiplications. Compared to SHAKE, LINCS is more stable and faster but can only handle bond length and isolated angle constraints like the OH proton angle. The stability of LINCS makes it very useful for Brownian dynamics simulations.

Let's consider a system of N particles, where the positions of each particle are given by the $3N$ vector $r(t)$.

$$\frac{d^2r}{dt^2} = M^{-1}F \tag{2.33}$$

where F represents the $3N$ force vector, and M is a $3N\ x\ 3N$ diagonal matrix containing the masses of the N particles. Additionally, there are K time-independent constraint equations that must be satisfied. These constraint equations impose restrictions on the positions of the particles, such as fixing bond lengths or angles.

$$g_i(r) = |r_{i1} - r_{i2}| - d_i = 0, i = 1, \dots, K \tag{2.34}$$

Like SHAKE, LINCS with a numerical integration algorithm is applied after an unconstrained position update. LINCS works in two steps: The projections of the new bonds onto the old bond directions are set to zero. This removes the components of the displacement along the bonds. A correction is applied for the lengthening of the bonds due to rotations. The mathematics for the two steps is very similar. The complete derivation of the LINCS algorithm is given in reference.[154]

Temperature Control

In a typical MD simulation, the behaviour of N atoms in volume V is governed by classical Newton's laws, leading to a microscopic distribution that aligns with the NVE ensemble. Energy is conserved in the system. However, this does not reflect the constant temperature conditions commonly seen in experiments, so to better approximate these conditions, the NVT, NPT, and μVT ensembles, which maintain a constant temperature, are more widely used. Temperature control is important for several reasons, such as studying temperature-dependent properties or processes, preventing unrealistic heating due to numerical errors, and improving sampling efficiency in some cases.

Temperature control is achieved in an MD simulation by connecting the molecular system to a thermostat, which modifies the motion of the atoms or changes their equations of motion to add or subtract energy. Different techniques for controlling the temperature in MD simulations include adjusting the motion of atoms or modifying their Newtonian equations of motion through a thermostat. These thermostats can utilize random or systematic methods and be applied to the entire system or individually to each atom/molecule.

Pressure Control

In many cases, maintaining a consistent pressure is crucial for MD simulations to be in agreement with experiments that involve the measurement of thermodynamic properties in open air or under conditions that differ from atmospheric pressure. Experiments can achieve this by using a piston-filled container with an inert gas, while simulations can use a barostat to maintain constant pressure. If only a barostat is employed, it will only regulate pressure, leading to an NPH ensemble. To attain an NPT ensemble, a barostat must be combined with a thermostat.[140]

There are several benefits to controlling pressure in an MD simulation. For instance, the NPT ensemble, which is maintained at constant pressure, allows for fluctuations in energy and density. This allows researchers to study processes involving significant energy fluctuations, such as crystal nucleation[155, 156] and the phase separation[157, 158]. Having control over the pressure also enables researchers to study the system's response to finite heating[159], cooling[160], compression[161], and expansion rates[162], such as the glass transition[163] and the formation of bubbles[164].[165] Furthermore, by

maintaining control over the pressure, MD simulations can examine nonequilibrium processes resulting in large energy releases[166] and substantial volume changes, such as chemical reactions and phase transitions.[167] The pressure of a molecular system is typically calculated using the virial equation.

$$PV = Nk_BT + \frac{2}{3}W \tag{2.35}$$

where V is the volume, and the virial term W is defined as

$$W = \frac{1}{2}\sum_{i=1}^{N} r_i . F_i \tag{2.36}$$

where F_i denotes the total force on atom i. As it depends on atomic positions and interaction forces, the pressure is a time-dependent variable. Calculating the pressure at a single time step gives the instantaneous pressure, which fluctuates around the target pressure.

2.2. Force-Fields

A force field is a mathematical representation that outlines the relationship between a system's energy and its particles' positions.[168] It is comprised of a formula for calculating the potential energy between atoms and a collection of parameters. These parameters are typically obtained through either quantum mechanical calculations or experimental data. The force field substitutes the actual potential with a convenient model that can be quickly calculated but still provides accurate results for the desired properties.[139] There are many force fields available with varying complexity and intended for different systems, and a typical force field expression may look like this:

$$U = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)] + \sum_{improper} V_{imp} + \sum_{LJ} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}} \quad (2.37)$$

The first four terms describe internal energy contributions (bond stretching, angle bending, torsions), and the last two terms describe non-bonding interactions (repulsion, Van der Waals, Coulombic).

2.2.1. Intramolecular Terms

Bond stretching is described using a simple harmonic function in equation (2.37). The value for r_0 can be gathered from experiments, and the spring constant can be calculated from spectra. However, the harmonic function becomes unreliable for significant changes in bond length and does not consider chemical reactions. Sometimes, a more accurate functional form, like the Morse potential, is used, but the harmonic form is more commonly used due to its simplicity and lower computation cost. Angle bending is generally described using a harmonic potential, although a trigonometric potential is sometimes used.

$$U_{bending} = \frac{1}{2} k_a (cos\theta - cos\theta_0)^2 \quad (2.38)$$

The Urey-Bradley[169] potential is sometimes used to improve the accuracy of vibrational spectra.

$$U_{UB} = \sum_{angles} \frac{1}{2} k_{UB}(s - s_0)^2 \tag{2.39}$$

where *s* is a distance between two atoms forming an angle.

For molecules with more than four atoms in a row, a dihedral or torsional term must be taken into account. Torsional motions, which are less rigid than bond stretching and play a key role in determining the molecular conformations and evaluating the stability of different conformations, are essential for accurately modelling the local structure of a macromolecule. Torsional energy is often described using a cosine function with parameters for the torsional angle (φ), phase (δ), number of minima or maxima (n), and potential barrier height (V_n). A dihedral potential with varying depths can be created using multiple terms with different *n*. Alternative representations, such as the OPLS[170] potential below, also exist.

$$U_{tors} = \sum_{torsions} k_0 + \frac{k_1}{2}(1 + \cos\varphi) + \frac{k_2}{2}(1 - \cos 2\varphi) + \frac{k_3}{2}(1 + \cos 3\varphi) \tag{2.40}$$

The torsional parameters are frequently obtained from a combination of *ab initio* calculations and experimental data. An additional term may be required to maintain planarity for certain groups, such as sp^2 hybridized carbons in carbonyl groups or aromatic rings.[168] This additional term accounts for the energy contributions from deviations from the plane. Formulas for the improper torsion term include:

$$U_{imp} = \sum_{impropers} \frac{k_{imp}}{2}[1 + \cos(2\omega - \pi)], or$$
$$U_{imp} = \sum_{impropers} \frac{k_{imp}}{2}(\omega - \omega_0) \tag{2.41}$$

where the improper angle (ω) represents deviation from planarity.

2.2.2. Intermolecular terms

Van der Waals interactions between atoms result from a balance between repulsion due to overlapping electron clouds and attraction from induced dipoles, which varies as r^{-6}. Van der Waals forces can occur between atoms of different molecules or within the same molecule if separated. Parameters for each atom pair can be defined, but most force fields use individual atomic parameters and rules for combining them. The LJ potential depth of the well for an atom pair is calculated using the geometric mean, and the point at which the potential becomes zero is determined by the geometric or arithmetic mean, depending on the force field.

The final term in equation (2.37) describes electrostatic interactions, which are typically calculated using partial atomic charges and Coulomb's law. These charges can be determined through experimental data, fast methods based on electronegativities, or *ab initio* calculations[136]. However, partial charges are not experimental observables and may differ between methods.[171, 172] Polarization effects in condensed phases are often taken into account using effective charges. Electrostatic interactions are long-ranged and may require special treatment when calculating forces, and partial charges are often only assigned to atomic sites for computational efficiency. However, it is also possible to place partial charges outside of atomic positions, as demonstrated by water models with varying numbers of water models.[173-175]

Interactions between atoms within a molecule that are separated by more than three bonds are generally handled similarly to intermolecular interactions. Meanwhile, interactions between pairs of atoms 1-2 and 1-3 are omitted to prevent numerical difficulties and because intramolecular terms have already considered them. Handling 1-4 interactions can vary, but they are often partially considered using a scaling factor. This scaling factor can vary between force fields and should be taken into account when combining parameters from different sources. The combination of the dihedral potential and the 1-4 Van der Waals and electrostatic interactions determine rotational barriers, so modifying or mixing parameters from different sources is generally not advisable.[168]

2.3. Water Models

Water is the most studied molecule, essential for life.[174, 176] However, we have yet to fully understand how this seemingly simple molecule of just three atoms gives rise to its many extraordinary liquid phase properties. Hundreds of theoretical and computational models have been proposed for water due to its complex properties and multiple theoretical and computational approximation levels.[177] Rigid non-polarizable models define water as a set of points that are the simplest and computationally efficient among water models. The figures below show the general geometries of the vastly used rigid non-polarizable models (**Figure 2.6**). The exact parameters vary depending on the model. These models are reasonably accurate and fast, but they're not perfect. In particular, none accurately reproduces all of the characteristics of bulk water. Even minor errors in water models can cause unexpected effects on the results of atomistic biomolecular simulations.

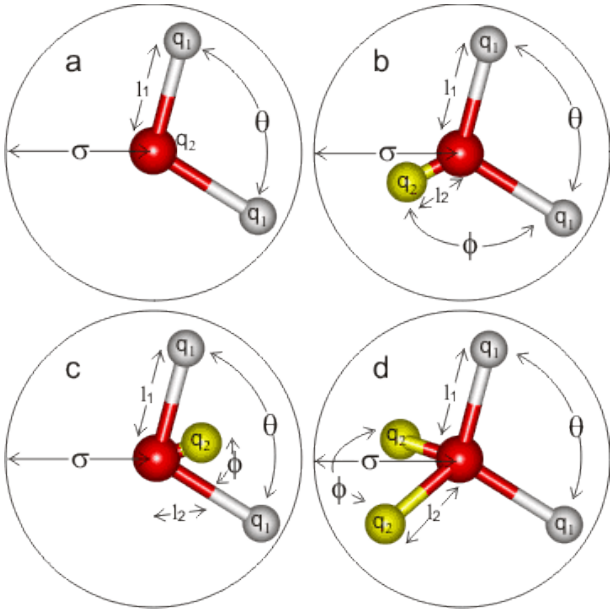

Figure 2.6: *The general shape of water models. Model types* ***a***, ***b***, *and* ***c*** *are all planar, whereas type* ***d*** *is almost tetrahedral. Reproduced with permission from "Water Models"[177] © Martin Chaplin, under CC BY-NC-ND 2.0 licence: https://creativecommons.org/licenses/by-nc-nd/2.0/*

The computational cost has always been crucial, especially in biomolecular simulations. In order to fully fill a biomolecular system with water in a molecular dynamics simulation, thousands of water molecules are required. This results in computing millions of water interactions at each time step. Rigid non-polarizable models are widely used in biomolecular simulations due to their simplicity and computational efficiency (**Table 2.1**).

This type of model was first proposed by Bernal and Fowler with a straightforward pairwise energy function.[175, 178]

$$E = \sum_{pairs} \left(\frac{A_{LJ}}{r_{\infty}^{12}} - \frac{B_{LJ}}{r_{\infty}^{6}} + \mathfrak{k} \frac{q_i q_j}{r_{ij}} \right) \tag{2.42}$$

where A_{LJ} and B_{LJ} are the corresponding LJ parameters. r_{∞} is oxygen–oxygen distance, r_{ij} is distance between the charged sites with charges q_i, q_j, and $\mathfrak{k}$ is the constant in Coulomb's law. This function does not consider quantum effects and electronic polarizability.

Table 2.1. Parameters for some water models used in Chapter 3.

Model	Type	σ (Å)	ε (kJ/mol)	l_1 Å	l_2 Å	q_1 (e)	q_2 (e)	θ°	φ°
SPC[179]	**a**	3.166	0.650	1.0000	-	+0.410	-0.8200	109.47	-
SPC/E[180]	**a**	3.166	0.650	1.0000	-	+0.4238	-0.8476	109.47	-
TIP3P[181]	**a**	3.15061	0.6364	0.9572	-	+0.4170	-0.8340	104.52	-
TIP4P[181]	**c**	3.15365	0.6480	0.9572	0.15	+0.5200	-1.0400	104.52	52.26
TIP4P-Ew[182]	**c**	3.16435	0.680946	0.9572	0.125	+0.52422	-1.04844	104.52	52.26
OPC[174]	**c**	3.1666	0.8903	0.8724	0.1594	+0.6791	-1.3582	103.6	51.8
TIP5P[181]	**d**	3.12000	0.6694	0.9572	0.70	+0.2410	-0.2410	104.52	109.47

Parameters and the model's geometry are typically adjusted for a chosen set of experimental water properties to produce the best fit.

2.4. Absolute Binding Enthalpy Calculations

Enthalpy (H) is a measure of the total energy of a system. H is given by summing of internal energy and the product of pressure and volume (Eq. 2.43).

$$H = U + PV \tag{2.43}$$

where U is internal energy, P and V are pressure and volume, respectively. PV contribution to the binding enthalpy is negligibly small (except at very high pressure)[183] compared to the other terms.[184] The system is typically simulated in the constant pressure and temperature in MD simulations, with a thermostat and a barostat. In this ensemble, the system's pressure and volume fluctuate dynamically to maintain a constant temperature. As a result, the PV term's average contribution over the simulation tends to be small compared to other energetic contributions, such as intermolecular interactions. It has been shown that the transformations of glucopyranose ring conformers, which involve changes in volume and pressure, have little effect on the enthalpy differences between them.[185] Therefore, the differences in enthalpy between the conformers are mainly due to energy factors rather than changes in pressure and volume.

The binding enthalpy (ΔH) represents the energy change of the system after binding event. This energy change comes from the formation or disruption of interactions (hydrogen bonds, van der Waals forces, and all other interactions) between the receptor and the binder. Nevertheless, ΔH is a measure of global change, including solute and solvent contributions due to forming and disrupting many interactions between all molecules in the system, including, for example, protein, ligand, solvent, and ions. All these elements can drive either favourable or unfavourable enthalpic contributions, and the absolute binding enthalpy derives from combining all these contributions. We can compute ΔH as the potential energy difference between the bound and the unbound states of system (**Figure 2.7**).[31, 34]

$$\Delta H = E_{complex} + E_{water} - E_{host} - E_{guest}$$

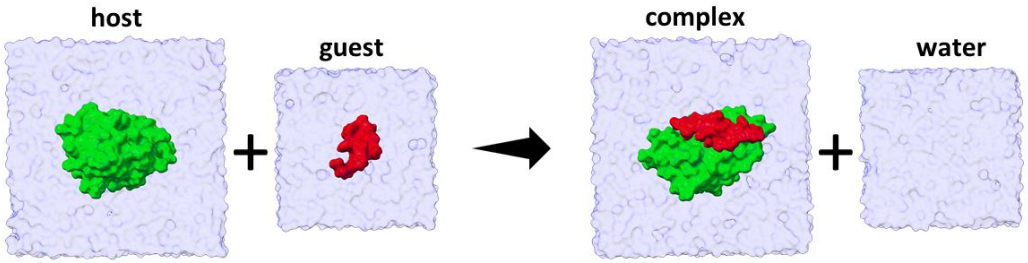

Figure 2.7: *Solvent-balance method for getting absolute binding enthalpies.*

We can write the reaction of the noncovalent association of host and guest to build a complex:

$$host + guest \leftrightharpoons complex \tag{2.44}$$

A solvent mixture plays a crucial role in this reaction in the biological environment. In the reaction solvent, each component has a chemical potential given by

$$\mu_i = \mu_i^\circ + RTln\left(\frac{\gamma_i C_i}{C^\circ}\right) \tag{2.45}$$

where i = host, guest, and complex components in solution. μ_i°, γ_i and C_i is the standard chemical potential, the activity coefficient, and the concentration of each component, respectively. R is the gas constant, T is absolute temperature. C° is the standard concentration.[184] The standard chemical potential for each component at standard concentration is

$$\mu_i^\circ = -RTln\left(\frac{8\pi^2}{\sigma_i C^\circ}\frac{W_i}{W^\circ}\right) \tag{2.46}$$

where W_i and W° are the partition functions of water molecules with and without the solute, respectively. σ_i is the symmetry number in case the partition function of each

component.[186] A detailed derivation of this expression may be found elsewhere.[187] We can write the binding free energy using the standard chemical potentials of each component at the standard concentration in solution.

$$\Delta G^{\circ} = \mu^{\circ}_{complex} - \mu^{\circ}_{host} - \mu^{\circ}_{guest} \tag{2.47}$$

The standard chemical potential can be also used to write the partial molar enthalpic and entropic terms.

$$\mu_i^{\circ} = h_i^{\circ} - Ts_i^{\circ} \tag{2.48}$$

We can use the partial molar entropy

$$s_i^{\circ} = -\frac{\partial \mu_i^{\circ}}{\partial T} \tag{2.49}$$

to extract the partial molar enthalpy as

$$h_i = \mu_i^{\circ} - T\frac{\partial \mu_i^{\circ}}{\partial T} \tag{2.50}$$

Using the standard chemical potential (Eq. 2.46) in this equation (Eq. 2.50) gives

$$h_i = RT^2\left(\frac{\frac{\partial}{\partial T}W_i}{W_i} - \frac{\frac{\partial}{\partial T}W^{\circ}}{W^{\circ}}\right) \tag{2.51}$$

The partition functions in the equation (Eq. 2.51) give $W = \int e^{-\frac{E(r)}{RT}}dr$, where $E(r)$ is the potential energy of the system as a function of r, which includes all atomistic coordinates in the system. Then, we can get that

$$\frac{1}{Z}\frac{\partial Z}{\partial T} = \frac{\langle E \rangle}{RT^2} \tag{2.52}$$

where $\langle E \rangle$ is the Boltzmann average of the potential energy. Substituting Eq. 2.52 into Eq. 2.51 gives the partial molar enthalpy as

$$h_i = \langle E_i \rangle - \langle E^{\circ} \rangle \tag{2.53}$$

Here, $\langle E_i \rangle$ and $\langle E^{\circ} \rangle$ are the mean potential energies of system with and without the solute, respectively. ΔH is given by using the standard chemical potential of each component.

$$\Delta H = h_{complex} - h_{host} - h_{guest} \tag{2.54}$$

Consequently, we can obtain the solvent-balance method (**Figure 2.7**) for getting absolute binding enthalpies by combining the partial molar enthalpies of each component with Eq. 1.11 as

$$\Delta H = \langle E_{complex} \rangle - \langle E_{host} \rangle - \langle E_{guest} \rangle - \left(\langle E^{\circ}_{complex} \rangle - \langle E^{\circ}_{host} \rangle - \langle E^{\circ}_{guest} \rangle \right) \tag{2.55}$$

$$\Delta H = \langle \mathrm{E} \rangle_{complex} + \langle \mathrm{E} \rangle_{water} - \langle \mathrm{E} \rangle_{receptor} - \langle \mathrm{E} \rangle_{ligand} \tag{2.56}$$

where $\langle \cdots \rangle$ is the time-averaged value obtained from multiple independent MD trajectories. This derivation of ΔH was also described previously by Fenley et al.[31]

A more comprehensive sampling of the conformational space can be achieved by conducting multiple simulations compared to a single, extended simulation.[34] This allows for the determination of the average values that are distributed throughout the potential energy surface. The mean energy of each simulation is estimated by computing the average over a time period corresponding to N number of snapshots, while the expected energy value of the entire set of simulations, ⟨E⟩, is found by calculating the mean of each individual simulation. When there are K simulations, the ensemble mean is calculated by taking the sum of the mean values for each simulation and dividing by K (Eq. 2.57).

$$\langle E \rangle = \frac{1}{KN} \sum_{t=1}^{K} \sum_{n=1}^{N} E_{t,n} \tag{2.57}$$

where $E_{t,n}$ is the energy value of the n'th snapshot of the t'th trajectory. K is the number of trajectories, and N is defined as the number of snapshots taken from a trajectory.

2.5. Absolute Binding Free Energy Calculations

A ligand (L) can bind reversibly to a protein (P) to form a complex (PL).

$$P + L \rightleftharpoons PL \tag{2.58}$$

When in equilibrium, the binding constant (K_b°) reflects the ratio of the concentration of the reactants and products in the solution. The binding constant is a dimensionless value that depends on the selected standard state. It is related to the binding free energy (ΔG_b°) through the equation.

$$\Delta G_b^\circ = -k_B T ln K_b^\circ \tag{2.59}$$

where k_B is the Boltzmann constant and T is the temperature. The binding free energy is determined in reference to the selected standard state. There are many methods for calculating binding free energies, with varying accuracy and computational efficiency levels. Scoring approaches are quick but less accurate, while free energy calculations based on atomistic simulations (alchemical or geometrical) are more accurate but slower. Alchemical free energy calculations involve multiple steps to calculate the binding free energy by inserting or removing the ligand from different environments. Geometrical calculations use a physical pathway, pulling the ligand away from the protein and calculating the work required to extract it. Endpoint methods, which are based on the analysis of snapshots from simulations using implicit solvent models, fall between the two in terms of accuracy and efficiency. Alchemical and geometrical free energy approaches are based on statistical mechanics. However, scoring functions and endpoint methods are not as accurate, as they estimate the free energy rather than rigorously calculating it.

2.5.1. Thermodynamic Cycle

The perturbation and thermodynamic integration methods require intermediate states to accurately determine the free energy difference. If the end states are a ligand that is bound to a protein and one that is unbound, a series of intermediate states must be established to calculate the binding free energy.[188] The free energy difference is determined by establishing a path that links the two end states of interest and dividing it into separate simulations (windows). The binding free energy is the total of the free energy differences

between each state and the state that follows it. A thermodynamic cycle with nonphysical intermediate states can be used to connect the end states, as shown in **Figure 2.8**.[189] Six stages are involved in computing the absolute binding free energy, consisting of two significant end states (bound and unbound) and four intermediary states where the ligand is isolated from its surroundings (with no interaction with other molecular entities). The ligand is transformed from a fully interacting ligand in solution to a noninteracting solute by scaling its electrostatic and van der Waals interactions (ΔG_{elec}^{solv} and ΔG_{vdw}^{solv}) to zero through intermediate states. The ligand is then restrained while still not interacting with the environment and has its interactions (ΔG_{elec}^{prot} and ΔG_{vdw}^{prot}) turned back on in complex with the protein. Finally, the restraints are removed (ΔG_{rest}^{prot}), and the cycle is closed, reaching the bound protein-ligand state.

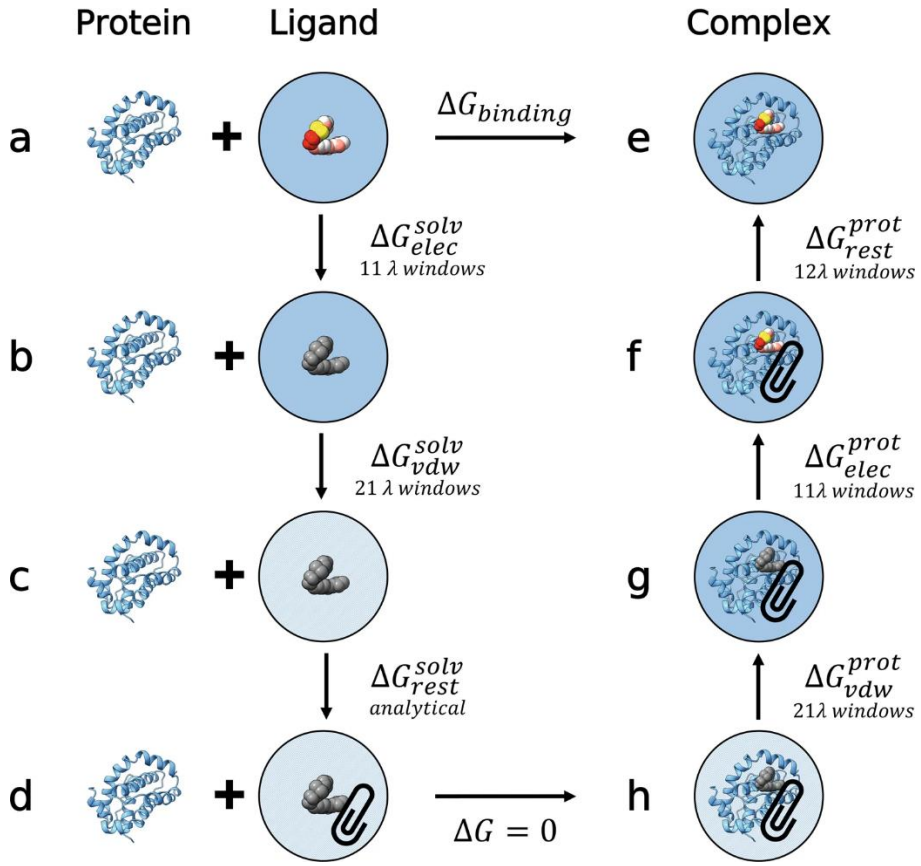

Figure 2.8: *The binding free energy is measured through a non-physical process (a-e). The electrostatics are eliminated in 11 windows (**b**), followed by 21 windows (**c**) that separate the ligand's van der Waals interaction from the solvent. The ligand is then restricted (**d**) and considered equivalent to a non-interacting ligand in a complex (h). The ligand's interactions with the environment are restored, starting with the van der Waals interactions in 21 windows (**g**), followed by the electrostatics in 11 windows (**f**). The restraints on the ligand's orientation are then removed in 12 windows (**e**), resulting in a fully interacting complex. Reproduced with permission from Alibay, I., Magarkar, A., Seeliger, D. et al. Commun Chem 5, 105 (2022).[189]*

2.5.2. Restraints

Restraints prevent the ligand from leaving the binding pocket during simulations, ensuring the conformations correspond to a bound state. Without restraints, the ligand would have a larger configurational space and cause convergence issues. Restraints help improve phase space overlap and convergence.[188]

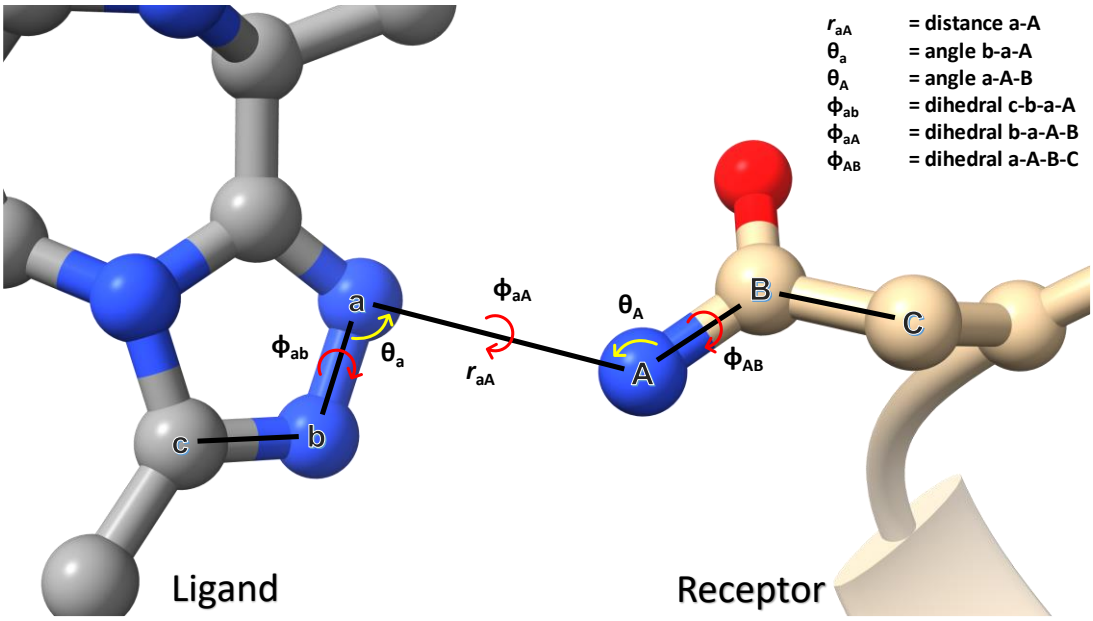

Figure 2.9:. *Boresch restraints for free energy calculations involving atoms "a," "b," and "c" (Ligand) and atoms "A," "B," and "C" (Receptor).*

Restraints are necessary to prevent the ligand from having a larger configurational space, which can cause convergence issues. Ligand restraints, like the ones proposed by Boresch et al., can keep the ligand in a specific orientation, reducing the number of simulations needed and providing an analytical solution for ΔG_{rest}^{solv}.[190] The set consists of one distance, two angles, and three dihedrals between three ligand atoms and three protein atoms and includes the standard state correction. The ΔG_{rest}^{prot} can be obtained by simulating various intermediate states and interpolating the force constants of the six

harmonic restraints. The ΔG_{rest}^{solv} can be calculated using the formula in Boresch et al., which takes into account the standard state correction.[188, 190]

$$\Delta G_{rest}^{solv} = k_B T ln \left[\frac{8\pi^2 V^{\circ}}{r_{aA}^2 \, sin\,\theta_a \, sin\,\theta_A} . \frac{\left(k_{r_{aA}} k_{\theta_a} k_{\theta_A} k_{\varphi_{ab}} k_{\varphi_{aA}} k_{\varphi_{AB}}\right)^{\frac{1}{2}}}{(2\pi k_B T)^2} \right] \qquad (2.60)$$

where V° is the standard volume (1660 $Å^3$), r_{aA} is the distance restraint value, θ_a and θ_A are the angle restraint values, and k are the force constants for the distance, angle, and dihedral restraints (ϕ_{ba}, ϕ_{aA}, and ϕ_{AB}).

2.6. Error analysis

When we sample data (e.g., potential energy, pressure, RMSD, etc.) using MD simulations, we obtain a time series of data. We can say $x_1, x_2 \ldots, x_N$ which is a result of N successive measurements (correlated data) from MD simulations. The fact that uncorrelated data-points are desirable allows us to avoid a lot of hassle. We can quickly calculate the average value, μ,

$$\mu = \frac{x_1, x_2 \ldots, x_N}{N} \tag{2.61}$$

of the uncorrelated data as well as its standard deviation, σ.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \tag{2.62}$$

Correlated data's standard error of the mean can be estimated using several approaches such as computing the correlation length, reblocking, and resampling techniques like the jackknife and bootstrap methods.[191-194] Reblocking was used in this case because it is computationally efficient and reduces the memory needed. Using the standard error expression for independent data points (blocking length of 1) leads to overestimating the number of independent data points and, therefore, underestimating the standard error. The standard error increases as the blocking length increases until it reaches a level where the block averages are no longer dependent on each other, which is when the blocking length surpasses the data's decorrelation period. Having a large blocking length results in small datasets with an unreliable standard error estimate (**Figure 2.10**). There are techniques available to determine the ideal blocking length for a specific dataset.

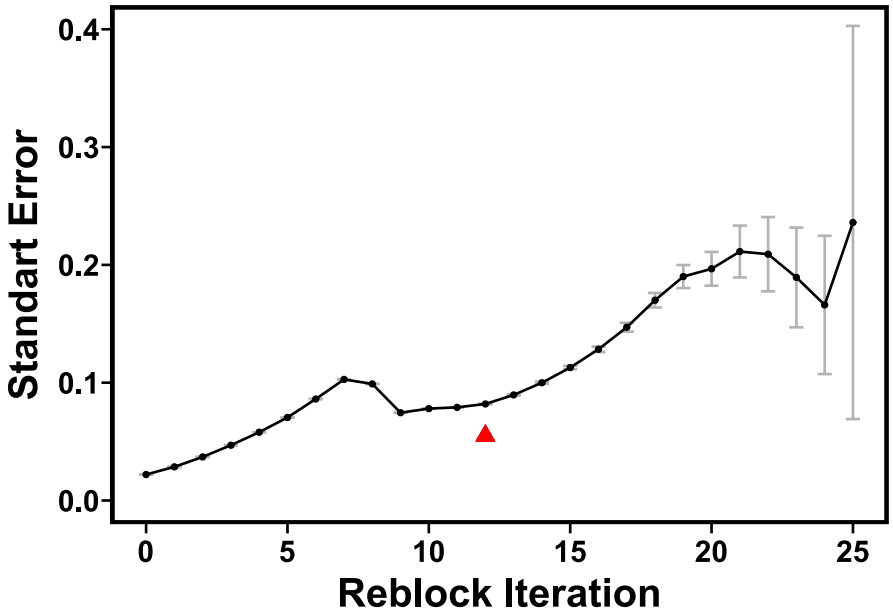

Figure 2.10:. *A sample reblocking analysis of MD data. The optimal block size (red triangle) is chosen by the algorithm described in eq 2.54.*

This method can be used to estimate the standard error of a data set. It is often used in simulations and other Monte Carlo methods to reduce the variance of estimates and improve the accuracy of the results. Suppose we have a set of data points, $x_1, x_2 \ldots, x_N$, and we want to estimate the mean of the data, denoted as μ. We can divide the data into B blocks, with the bth block containing $x_{[(b-1)n+1]}, x_{[(b-1)n+2]}, \ldots, x_{bn}$, where n is the size of each block and $1 <= b <= B$. The estimate of μ for the bth block is given by:

$$\mu_b = \left(\frac{1}{\mathrm{n}}\right)\sum\left(x_{[(b-1)n+1]}, x_{[(b-1)n+2]}, \ldots, x_{bn}\right) \tag{2.63}$$

The overall estimate of μ is then given by:

$$\mu = \left(\frac{1}{\mathrm{B}}\right)\sum(\mu_1, \mu_2, \ldots, \mu_B) \tag{2.64}$$

The standard error of the overall estimate, denoted as se, can be estimated using the following formula:

$$se = \frac{\sigma_b}{\sqrt{B}} \tag{2.65}$$

One challenge in understanding this data type is determining the correct block size. In the example presented, a simulation of 10 microseconds (10 million data points) was used,

which resulted in a plateau in the reblock plot. However, manually reviewing the reblocked data for each calculation to decide by eye is neither objective nor efficient. An algorithm to automatically choose the best block size can be simple and robust as shown below.[195]

$$B_{opt} = \sqrt[3]{2nn_{corr}^2} \tag{2.66}$$

where n is the number of data points and n_{corr} is the correlation length described elsewhere[195, 196]. By using the blocking method to estimate standard error, we can reduce the variance of the overall estimate and obtain a more accurate estimate of the mean of the data. However, the blocking method does have some limitations. It can be sensitive to the choice of block size, as the standard error may vary depending on how the data are divided into blocks. Additionally, the method may be inaccurate for small sample sizes. Other methods, such as bootstrap or jackknife, may be more suitable in these cases.

Here, I used the pyblock tool that automatically performs reblocking analysis, which removes serial correlation from a dataset to get a more accurate estimate of the standard error.[197] pyblock also includes functions for further analyzing, interpreting, and manipulating the resulting mean and standard error estimates.

3 Evaluating the Performance of Water Models with Host–Guest Force Fields in Binding Enthalpy Calculations for Cucurbit[7]uril–Guest Systems

In this chapter, I will report the results of a study where I performed absolute binding enthalpy calculations for twenty-five different host-guest systems. This work aimed to assess the water models' performance with host–guest force fields. The content of this chapter has been published in the following journal article[58]:

Çınaroğlu, S. S., & Biggin, P. C. (2021). Evaluating the performance of water models with host–guest force fields in binding enthalpy calculations for cucurbit[7]uril–guest systems. The Journal of Physical Chemistry B, 125(6), 1558-1567. doi: 10.1021/acs.jpcb.0c11383

3.1. Introduction

Host–guest systems are promising for studying calorimetric calculations because of their size, enabling long-time-scale molecular dynamics (MD) simulations.[198, 199] They mimic many features of larger protein–ligand systems that require more computational resource.[55, 200] Importantly, the binding affinities of host–guest systems are generally in the same range as protein–ligand systems.[201, 202] Thus, their small size, simple structure, and predictable protonation states make them ideal miniature models of molecular recognition.[201] Their simple structures overcome setup problems in MD simulations, and their small sizes make it possible to reach numerical convergence faster than protein–ligand systems in explicit solvent for binding free-energy calculations.[2, 35]

Host–guest systems as tiny models are alternative for informative testing and improving force fields before using them in protein–ligand systems.[54] It may be complicated to choose which force field to use in MD simulations for protein–ligand systems, as new options are continually being offered. Simulators generally prefer to select the most cited or popular force field in their case, but the ability of all force fields to replicate experimental data is controversial.[203, 204] Protein–ligand systems as a benchmark would perhaps seem ideal to test force fields. However, many factors, such as simulation

setups, long-time-scale protein dynamics, pH effects, and so on, make these tests complicated. Host–guest systems offer a benchmark series that overcome many of these difficulties for the assessment studies of force fields. Furthermore, they may help not only to select appropriate force fields for MD simulation but also to optimize force fields for a specific case.[203, 204]

In addition to general force fields, the force-field parameters of water are crucial factors in molecular simulations using explicit water models.[173] Although water is a tiny and simple molecule, replicating its properties with MD simulations is still challenging.[205-209] The choice of water model significantly affects the simulation results since all water models represent water and its interactions according to various approaches or paradigms.[205, 206, 210] Rigid water models with a fixed charge distribution are popular and commonly used, but their ability to accurately predict entropic and enthalpic components still needs to be improved. Thus, assessment studies of the water models have been required to determine which model is capable of producing structural and dynamic properties of water.[211]

Some studies have reported how the choice of water model and force fields affects the accuracy of MD simulations for the prediction of thermodynamic properties.[31, 53, 54, 57, 212] Fenley et al.[31] compared TIP3P to TIP4P-Ew water models in binding enthalpy calculations for eight cucurbit[7]uril–guest pairs, showing that the TIP3P water model yields more accurate results than TIP4P-Ew. Subsequently, Gao et al.[53] investigated the salt dependency on the binding enthalpy across four water models, TIP3P, SPC/E, TIP4P-Ew, and OPC, with the choice of parameters for sodium and chloride using a neutral host–guest pair. Henriksen and Gilson[54] examined the influence of four water models (TIP3P, TIP4P-Ew, SPC/E, and OPC), ion parameters, partial charge assignment methods, and the cyclodextrin force field on thermodynamic entities. Yin et al.[57] introduced a new TIP3P derivative called Bind3P, using the sensitivity analysis approach. They evaluated Bind3P by comparing it with the TIP3P, TIP4P-Ew, OPC, and TIP4P-D in thermodynamics calculations. In the most recent work, TIP3P and Bind3P were compared in a representative set of organic micropollutants binding β-cyclodextrin. Another recent work reports a comprehensive study of the diffusivity of α-, β-, and γ-cyclodextrin in aqueous solutions using TIP3P, SPC/E, TIP4P/2005, and Bind3P water models.[213] Thus, a significant amount of work has already been done in this area.

However, a comparison of widely used methods for general force-field parameterization has not yet been presented. In this chapter, I use the direct method with the multi-box approach to determine the binding enthalpy and comparatively assess eight rigid models for water along with five different parameterization methods for the hosts and guests. The results further highlight the sensitivity to force-field parameter choice in estimating these values.

3.2. Methods

3.2.1. Host–Guest Systems and Force Fields

I used 25 host–guest pairs obtained from various publications to calculate the binding enthalpy.[214-219] I selected isothermal titration calorimetry (ITC) experiments without salts or buffers since they may yield significant errors relative to experimental data when used to compute host–guest binding enthalpies.[53] Structures are represented in **Figure 3.1** and are listed in **Table 3.1.** Also, the acyclic cucurbituril-derived TrimerTrip clip molecule with a guest molecule (heptane-1,7-diamine) was used for further investigations (**Figure 3.5A**).[220]

I performed MD simulations using five different force fields (CGenFF[221, 222], GAFFv1[223], GAFFv2[223], Parsley[224], and SwissParam[225]) and eight different water models including OPC[174], SPC[226], SPC/E[180], TIP3P[227], TIP4P[227], TIP4P-Ew[182], TIP5P[228], and Bind3P[57]. I only used the TIP3P water model for the TrimerTrip simulations. Strictly speaking, these water models are also force fields, but the term “water model” is commonly used in the literature, and we adhere to that convention hereafter. Detailed description about water models can be found in section **2.3**.

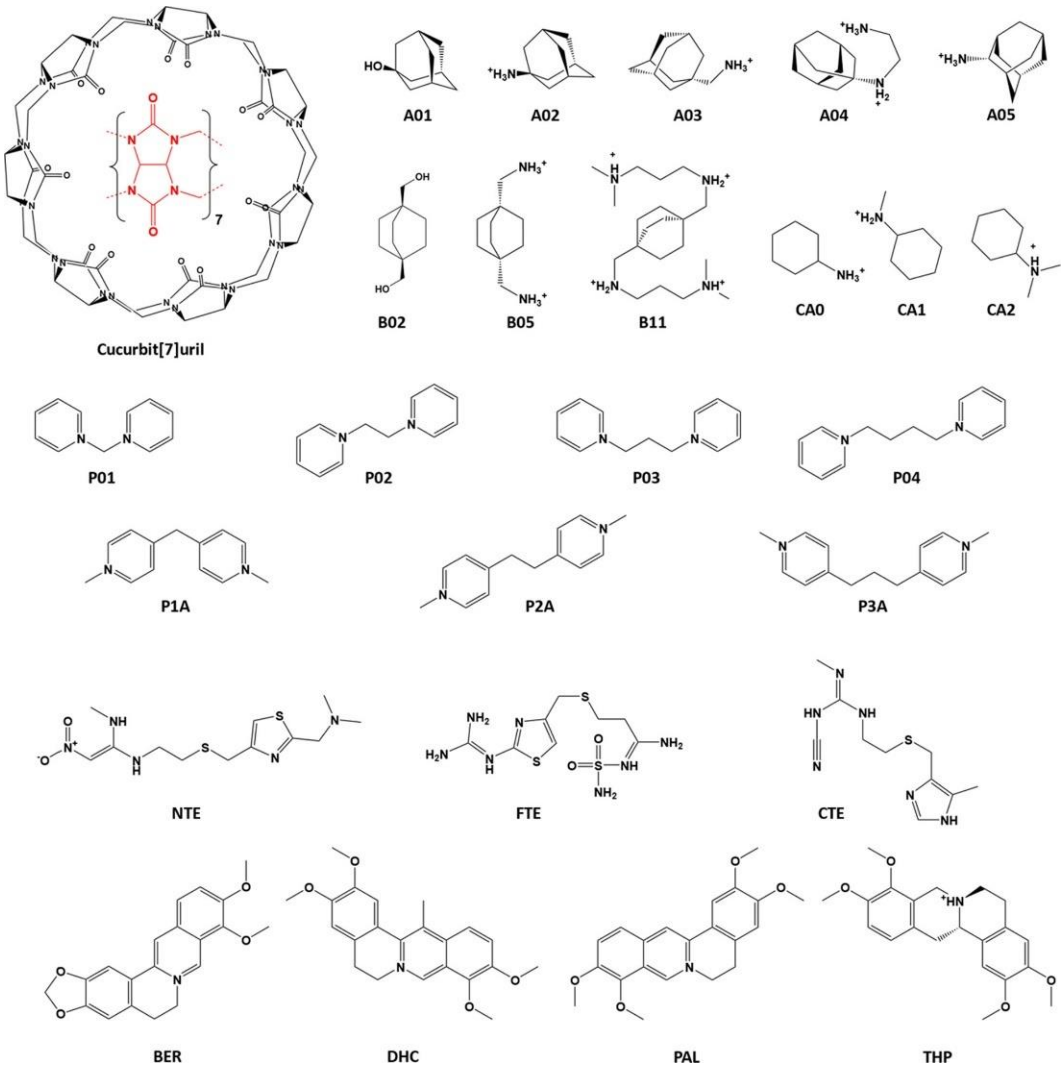

Figure 3.1: *Chemical structures of the host (cucurbit[7]uril) and guest molecules.*

GAFFv1 and GAFFv2 parameters and partial atomic charges obtained with the AM1-BCC method[229, 230] were generated by AMBER's antechamber tool[231], while CGenFF parameters and atomic charges were obtained using the CGenFF web server[221, 222, 232]. SwissParam[225] topologies and parameters with MMFF charges were generated by using the SwissParam server. In this work, I mainly focused on "out-of-the-box" performance, but others have reported for GAFF that consideration of different charge models (RESP and AM1-BCC) appears to make little difference.[54]

Table 3.1: *Guest Molecules with Their Experimental Binding Affinities Including Entropic and Enthalpic Components**

Guest	ΔG	–TΔS	ΔH	Reference
A01	–14.1 ± 0.2	4.9 ± 0.4	–19.0 ± 0.4	Moghaddam et al.[214]
A02	–19.4 ± 0.1	–0.1 ± 0.5	–19.3 ± 0.4	Moghaddam et al.[214]
A03	–20.3	1.7	–21.9 ± 0.4	Moghaddam et al.[214]
A04	–21.5	–1.4	–20.1 ± 0.4	Moghaddam et al.[214]
A05	–19.1 ± 0.2	0.4 ± 0.5	–19.5 ± 0.4	Moghaddam et al.[214]
B02	–13.4 ± 0.1	2.4 ± 0.2	–15.8 ± 0.2	Moghaddam et al.[214]
B05	–19.5 ± 0.2	–3.9 ± 0.5	–15.6 ± 0.4	Moghaddam et al.[214]
B11	–20.6 ± 0.4	–4.3 ± 0.5	–16.3 ± 0.4	Moghaddam et al.[214]
BER	–12.95 ± 0.36	–3.87 ± 0.28	–9.08 ± 0.5	Miskolczy and Biczók[218]
CA0	–10.79 ± 0.22	–0.88 ± 0.09	–9.91 ± 0.2	Yu et al.[217]
CA1	–9.17 ± 0.16	1.65 ± 0.07	–10.82 ± 0.1	Yu et al.[217]
CA2	–8.76 ± 0.16	2.82 ± 0.07	–11.58 ± 0.1	Yu et al.[217]
CTE	–5.69 ± 0.16	7.57 ± 0.01	–13.2 ± 4.0	Yin et al.[215]
DHC	–8.55 ± 0.28	–0.43 ± 0.14	–8.12 ± 0.24	Miskolczy et al.[219]
FTE	–6.40 ± 0.04	4.68 ± 0.01	–11.1 ± 0.4	Yin et al.[215]
NTE	–5.64 ± 0.09	9.96 ± 0.01	–15.6 ± 2.2	Yin et al.[215]
P01	–8.43 ± 0.02	–4.45 ± 0.01	–3.98 ± 0.02	Tcyrulnikov et al.[216]
P02	–8.15 ± 0.11	–1.69 ± 0.04	–6.45 ± 0.10	Tcyrulnikov et al.[216]
P03	–8.86 ± 0.04	–3.70 ± 0.0	–5.16 ± 0.04	Tcyrulnikov et al.[216]
P04	–8.83 ± 0.15	–4.29 ± 0.05	–4.55 ± 0.15	Tcyrulnikov et al.[216]
P1A	–8.28 ± 0.06	–1.30 ± 0.05	–6.98 ± 0.03	Tcyrulnikov et al.[216]
P2A	–8.98 ± 0.03	–2.39 ± 0.01	–6.58 ± 0.02	Tcyrulnikov et al.[216]
P3A	–8.75 ± 0.10	–1.67 ± 0.06	–7.08 ± 0.08	Tcyrulnikov et al.[216]
PAL	–9.83 ± 0.64	–1.00 ± 0.43	–8.84 ± 0.64	Miskolczy et al.[219]
THP	–6.88 ± 0.54	0.43 ± 0.36	–7.31 ± 0.41	Miskolczy et al.[219]

**All values are in kcal/mol and to the number of significant digits originally reported. Uncertainties are given as standard deviations of the mean.*

3.2.2. Molecular Dynamics Simulations

All simulations were performed using the Gromacs v2019 software package.[30, 142, 233] A periodic cubic water box was used for all systems with 1500 water molecules while 3000 water molecules for TrimerTrip simulations. A 3-step steepest descent energy minimization with a maximum force of 10 kJ/(mol nm) was applied to all systems.[234] In the first step, all heavy atoms were implemented position restraints with the harmonic potential at a force constant of 1000 kJ/(mol nm2), after those restraints on water molecules were removed, and then the final step removed all position restraints. All systems were equilibrated with a 5 ns NVT and NPT simulations. The V-rescale[235] and Parrinello–Rahman[236] algorithms were used for temperature and pressure coupling to control the temperature at 300 K and an isotropic pressure at 1.0 bar, respectively. Unbonded interactions were calculated up to a cutoff of 1.0 nm with a potential-shift modifier for GAFFv1, GAFFv2, and Parsley simulations, while the calculation was done using a cutoff of 1.2 nm with a force-switching for CGenFF and SwissParam simulations. A dispersion correction was only applied to energy and pressure for GAFFv1, GAFFv2, and Parsley. All H-bond lengths were constrained with a LINear Constraint Solver (LINCS)

algorithm.[237] Coulomb interactions were evaluated with the fast smooth particle-mesh Ewald (SPME) electrostatics method with an initial short-range cutoff of 1.0 (GAFFv1, GAFFv2, and Parsley) and 1.2 (CGenFF and SwissParam) nm.[148] The leap-frog algorithm was used to integrate the equations of motion for 100 ns MD simulation runs with the periodic boundary conditions. An integration time step was set to 2 fs for all simulations.

3.2.3. Binding Enthalpy Calculations

The binding enthalpy (ΔH) is calculated by **Eq. 3.1**, where $\langle \mathrm{E} \rangle_{complex}$, $\langle \mathrm{E} \rangle_{water}$, $\langle \mathrm{E} \rangle_{host}$, and $\langle \mathrm{E} \rangle_{guest}$ are the averaged potential energies of the system from four separate simulations of the host–guest complex, water, host, and guest, respectively. This multibox approach was implemented by Fenley et al.[31] In this approach, the number of solvent molecules should exactly balance the stoichiometry of the simulations

$$\Delta H = \langle \mathrm{E} \rangle_{complex} + \langle \mathrm{E} \rangle_{water} - \langle \mathrm{E} \rangle_{host} - \langle \mathrm{E} \rangle_{guest} \quad (3.1)$$

Bonded and unbonded state simulations of the host–guest complex should have an equal number of water molecules. If ions, such as sodium and chloride, are present in the solution, then these ions must similarly be balanced between these states. Note that the pressure–volume contribution is negligible for the binding enthalpy.[31] A detailed description was written in section **2.4**.

3.2.4. Error Analysis

The standard error of the mean (SEM) is calculated using the blocking method without running time-consuming additional simulations.[192, 196] The pyblock python module was used to estimate the SEM for time-correlated data series. It implements a reblocking analysis to remove serial correlation from a simulation data and obtain an improved SEM. The blocking method iteratively calculates averages of the simulation data series into consecutively larger blocks and calculates the SEM for each block. The experimental SEM listed here was obtained by dividing observed standard deviations by the square root of the number of replicates for a fair comparison. A detailed description was written in section **2.6**.

3.3. Results and Discussion

Numerous examples of binding affinity calculations using MD simulations are available in the literature.[2, 53, 198, 199, 203] This study provides additional insight into comparing the force fields and water model combinations in the binding enthalpy calculations. I report a comprehensive comparison of the various combinations tested in this work, showing which combination of parameters has good agreement with the experiment for the enthalpy calculation. For this purpose, 2080 independent MD simulations were performed on 25 host–guest systems having experimental data available (**Table 3.1**), using the direct, multi-box approach for the binding enthalpy calculations from the difference in mean potential energies.

3.3.1. Overall Statistics

Overall, the force-field combinations produced a diverse agreement with the experimental enthalpy values (**Table 3.2** and **Figure 3.2**). I may define the agreement with experimental values in two ways: (i) accuracy and (ii) ranking ability. Here, I report the accuracy of the calculation using slope/intercept, root-mean-square error (RMSE), MSE, and mean absolute error (MAE). The ranking ability is reported by the coefficient of determination (R^2) and Kendall's rank correlation coefficient (τ). The smallest RMSE value 2.07 (kcal/mol), relative to the experiment, is obtained with the GAFFv1–OPC combination. Yin et al.[57] also reported that the OPC water model outperformed Bind3P, TIP3P, TIP4P-Ew, and TIP4P-D water models in binding enthalpy calculations for octa acid and guests parameterized by GAFFv1. In another work by the same research group, TIP4P-Ew produced more accurate results than TIP3P, SPC, and OPC water models for binding enthalpies of the hosts (α- and β-cyclodextrin) and guests.[54] In the first study, GAFFv1 was used for the parameterization of the host molecule, while the Q4MD-CD force field[238] was preferred to generate proper conformations matching experimental data of the host molecule, illustrating that careful treatment of all components should always be considered. The largest RMSE, 56.25 kcal/mol, was observed for CGenFF-TIP5P. No single combination is superior to the others for the binding enthalpy. However, GAFFv1 may be a preferable default force-field choice for cucurbit[7]uril since GAFFv1, with all water models, except Bind3P, outperforms all other force fields. Previous works showed that Bind3P with GAFFv1 produced better results than TIP3P in the binding enthalpy

calculations for α- and β-cyclodextrin, octa acid, and tetra-endo-methyl octa acid (TEMOA).[57, 65]

Table 3.2: *Statistics between the Computed and Experimental Binding Enthalpy for Each Force Field*

	OPC	SPC	SPC/E	TIP3P	TIP4P	TIP4P-Ew	TIP5P	Bind3P
				GAFFv1				
Slope	0.99	1.18	1.11	1.23	1.06	1.01	0.57	1.21
Intercept	0.42	1.22	-0.24	3.2	-2.55	-3.69	-9.4	4.76
R^2	0.88	0.93	0.89	0.93	0.86	0.86	0.38	0.93
Kendall's τ	0.8	0.85	0.8	0.82	0.78	0.78	0.36	0.79
RMSE	2.07	2.22	2.6	2.22	4.02	4.32	6.3	3.15
MSE	4.28	4.92	6.78	4.93	16.15	18.67	39.7	9.89
MAE	1.52	1.89	2.25	1.9	3.71	3.86	5.31	2.65
				GAFFv2				
Slope	1.30	1.43	1.37	1.46	1.41	1.35	0.84	1.42
Intercept	3.86	3.96	2.14	5.01	0.47	0.28	-7.93	7.91
R^2	0.61	0.79	0.79	0.80	0.70	0.69	0.45	0.79
Kendall's τ	0.55	0.61	0.62	0.64	0.59	0.56	0.33	0.59
RMSE	5.75	4.73	4.81	4.62	6.94	6.52	7.86	5.46
MSE	33.10	22.36	23.16	21.34	48.17	42.57	61.83	29.84
MAE	4.27	4.04	4.09	4.05	6.14	5.83	6.75	4.40
				CGenFF				
Slope	0.09	0.35	0.26	0.4	0.32	0.21	-0.09	0.45
Intercept	-13.56	-10.37	-13.26	-7.57	-13.28	-14.63	43.01	-5.74
R^2	0.03	0.43	0.24	0.49	0.28	0.12	0.01	0.54
Kendall's τ	0.1	0.45	0.35	0.54	0.37	0.2	-0.06	0.54
RMSE	6.32	4.95	6.51	3.94	6.98	7.48	56.25	3.76
MSE	39.92	24.52	42.38	15.49	48.68	55.97	3,163.95	14.17
MAE	5.23	4.27	5.38	3.45	5.83	5.98	55.8	3.22
				Parsley				
Slope	0.53	0.76	0.70	0.81	0.72	0.58	0.11	0.83
Intercept	-8.86	-6.85	-9.79	-6.52	-10.62	-11.67	-21.57	-3.09
R^2	0.53	0.73	0.67	0.79	0.67	0.55	0.02	0.81
Kendall's τ	0.56	0.68	0.68	0.71	0.68	0.60	0.10	0.71
RMSE	4.94	4.92	6.96	4.92	7.98	7.59	12.98	2.56
MSE	24.41	24.17	48.51	24.25	63.65	57.68	168.44	6.56
MAE	4.02	4.08	6.24	4.26	7.34	6.69	11.21	1.83
				SwissParam				
Slope	0.62	0.76	0.69	0.77	0.76	0.68	0.39	0.79
Intercept	-9.96	-8.07	-10.43	-3.24	-11.22	-12.92	-13.23	-0.49
R^2	0.38	0.54	0.48	0.61	0.51	0.52	0.12	0.61
Kendall's τ	0.43	0.51	0.50	0.51	0.50	0.47	0.22	0.51
RMSE	7.25	6.58	7.95	3.59	9.37	9.97	8.95	4.08
MSE	52.50	43.35	63.19	12.86	87.85	99.37	80.03	16.63
MAE	6.27	6.07	7.27	2.68	8.85	9.19	7.04	2.86

R^2 is the correlation coefficient of determination. RMSE is the root-mean-square error. MSE is the mean squared error. MAE is the mean absolute error.

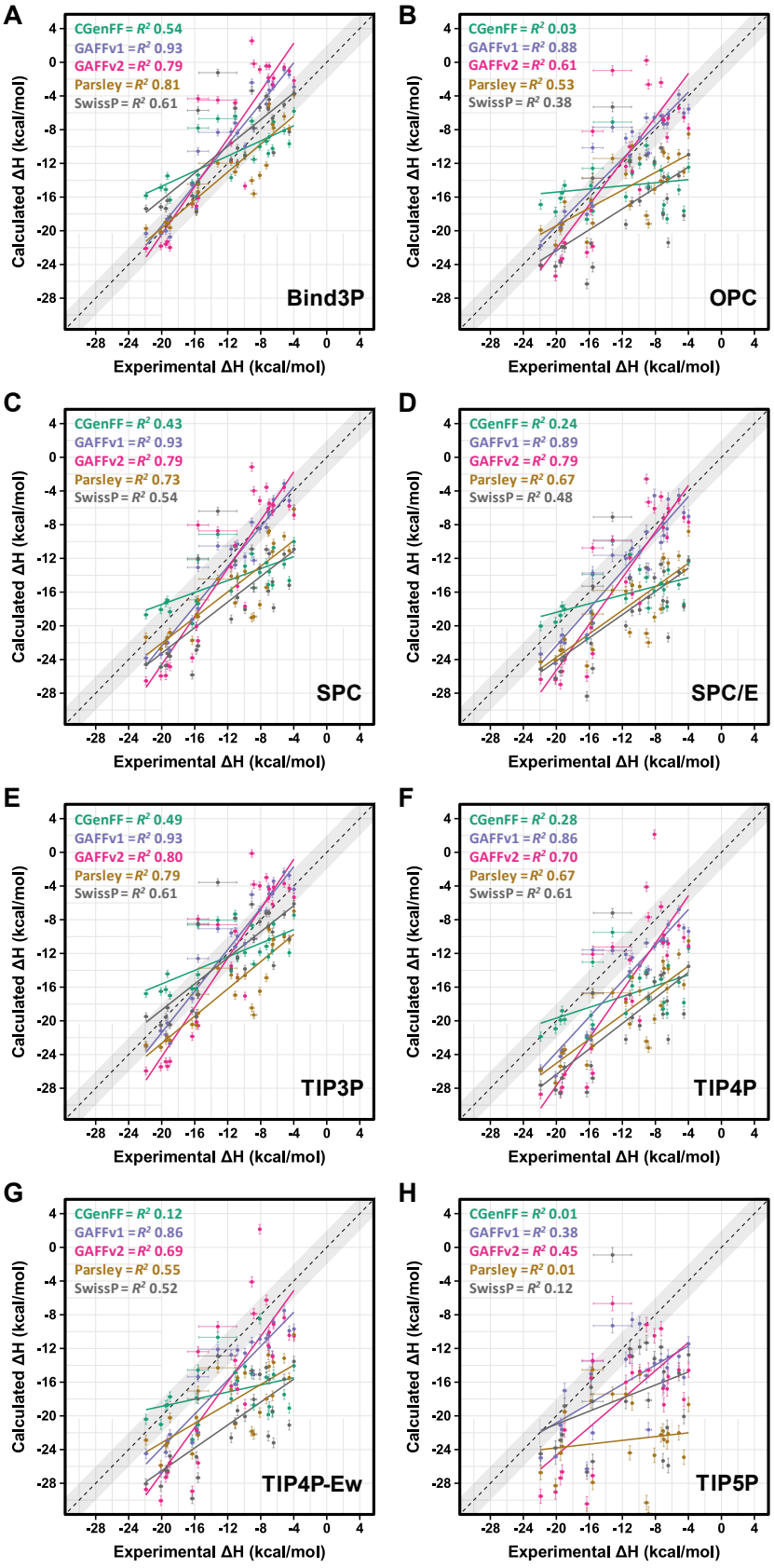
A
CGenFF = R^2 0.54
GAFFv1 = R^2 0.93
GAFFv2 = R^2 0.79
Parsley = R^2 0.81
SwissP = R^2 0.61
Bind3P
B
CGenFF = R^2 0.03
GAFFv1 = R^2 0.88
GAFFv2 = R^2 0.61
Parsley = R^2 0.53
SwissP = R^2 0.38
OPC
C
CGenFF = R^2 0.43
GAFFv1 = R^2 0.93
GAFFv2 = R^2 0.79
Parsley = R^2 0.73
SwissP = R^2 0.54
SPC
D
CGenFF = R^2 0.24
GAFFv1 = R^2 0.89
GAFFv2 = R^2 0.79
Parsley = R^2 0.67
SwissP = R^2 0.48
SPC/E
E
CGenFF = R^2 0.49
GAFFv1 = R^2 0.93
GAFFv2 = R^2 0.80
Parsley = R^2 0.79
SwissP = R^2 0.61
TIP3P
F
CGenFF = R^2 0.28
GAFFv1 = R^2 0.86
GAFFv2 = R^2 0.70
Parsley = R^2 0.67
SwissP = R^2 0.61
TIP4P
G
CGenFF = R^2 0.12
GAFFv1 = R^2 0.86
GAFFv2 = R^2 0.69
Parsley = R^2 0.55
SwissP = R^2 0.52
TIP4P-Ew
H
CGenFF = R^2 0.01
GAFFv1 = R^2 0.38
GAFFv2 = R^2 0.45
Parsley = R^2 0.01
SwissP = R^2 0.12
TIP5P
Calculated ΔH (kcal/mol)
Experimental ΔH (kcal/mol)
4
0
-4
-8
-12
-16
-20
-24
-28

Figure 3.2: *Comparison of calculated binding enthalpies with experimental values for different water models. (A) TIP4P, (B) SPC, (C) OPC, (D) TIP4P-Ew, (E) SPC/E, (F) TIP5P, (G) TIP3P, and (H) Bind3P for five different force fields: CGenFF, GAFFv1, GAFFv2, Parsley, and SwissParam. The R^2 values are also indicated.*

Overall, the Parsley force field in combination with Bind3P gives the most accurate result (RMSE = 2.56 kcal/mol). Parsley is a better force field in terms of reproducing quantum mechanical (QM) conformer energies and geometries than both GAFF versions[239]. Still, it only produces better enthalpy predictions when used in conjunction with the Bind3P water model. Generally, TIP3P and Bind3P have more accurate results (average RMSE = ∼3.8 kcal/mol) than other water models and also show a good correlation among all force fields, especially in GAFFv1. Surprisingly, GAFFv1 produces better results than GAFFv2 even though GAFFv2 was released to improve its performance and to cover a broader chemical space.[240] GAFFv1 and GAFFv2 are in the same force fields and most likely tend to be more consistent with each other. However, GAFFv1 is slightly more successful in reproducing QM conformer energies and geometries than GAFFv2.[239] Also, a problematic "over-definition" of the dihedrals in the GAFFv2 has been identified by a SAMPL7 study.[241] TIP5P gives the worst results regardless of force fields. It produces very inaccurate enthalpy values with CGenFF (**Figure 3.2**), making this combination one of the worst I tested. SwissParam (average RMSE = ∼6.97 kcal/mol) is the worst force field if TIP5P is ignored because of the result in CGenFF. Binding enthalpies computed with GAFFv1 in combination with TIP3P, Bind3P, and SPC water models (**Table 3.2**) correlate very well with the experiment (**Figure 3.2**): linear regression analysis yields a good correlation coefficient R^2 = 0.93. On the other hand, CGenFF gives the worst correlations for all water models. Overall, there is a tendency to overestimate the binding enthalpies. The CGenFF server produces default host parameters, which clearly require further optimization, at least in our hands. Both GAFFv1 and CGenFF can provide reasonable force fields for reproducing experimental data even though CGenFF performs worse overall in the binding enthalpy calculations.[242, 243]

In this assessment, it is also important to consider the uncertainties associated with both the experimental and computational values. The reported experimental uncertainties of the mean range between 0.01 and 2.31 kcal/mol and most of them are lower than 0.50 kcal/mol. Based on blocking analysis, the uncertainties of the calculated binding enthalpies have a range of 0.40–0.89 kcal/mol. Given these uncertainties, it would seem unlikely that the differences between experiment and computation can be attributed to undersampling

and there must be other reasons for the discrepancies. This is supported by the extremely flat convergence graphs shown in **Figure 3.3**, suggesting that the simulation trajectories have good conformational sampling.

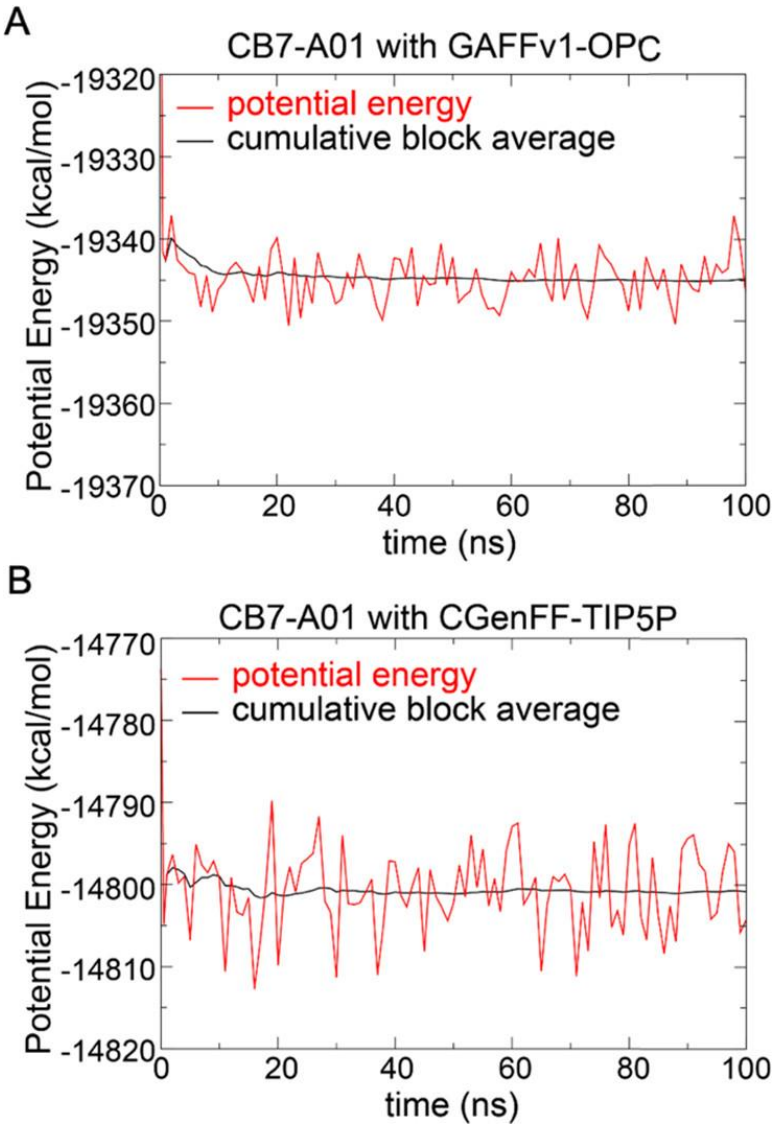

Figure 3.3: *Convergence plot of the binding enthalpy for the systems CB7-A01. The most and worst accurate combinations of force fields are included here. Instantaneous potential energy (at every 1 ns) is shown in red, while the black line shows the cumulative convergence in increasing blocks of 1 ns.*

3.3.2. Impact of the Force Fields on the Host Molecule

Strategies for force-field development depend on covering chemical space, training data, and the approach to optimizing parameters.[244, 245] Given the challenge of force-field development, there is often huge variability in their performance across a wide change of thermodynamic properties. As part of our efforts to understand the differences in force fields to predict enthalpy changes, I examined the conformation of the host in simulations

of the apo state. Parsley and CGenFF result in more structural deviation than other force fields (**Figure 3.4**), leading to conformations that exhibit an evident distortion away from the regular circular conformation observed in the crystal. It is important to note that this kind of distortion has been observed for CB7 when in complex with guests.[246] Parsley produced the highest structural fluctuation with 1 ± 0.1 Å, and GAFFv2 had the lowest one with 0.30 ± 0.10 Å (**Figure 3.4**).

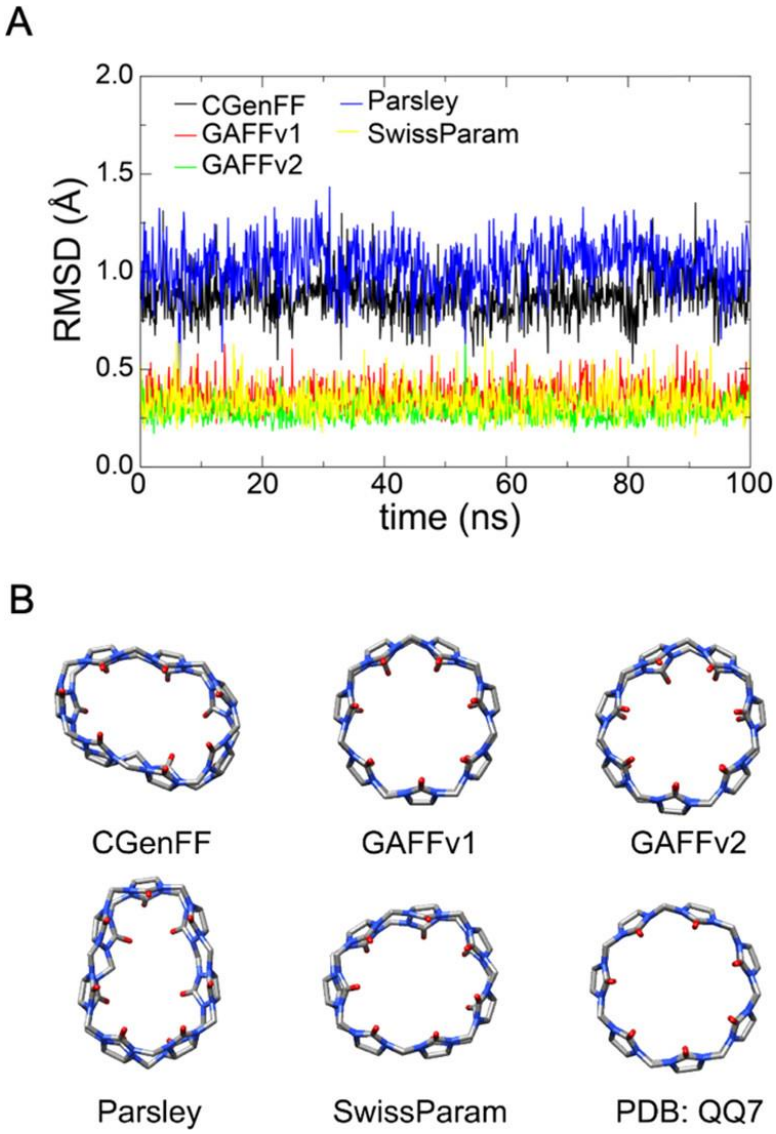

Figure 3.4: *Impact of the force fields on the host molecule (A) Root-mean-square deviation (RMSD) of heavy atoms in the unbound state of the host molecule through the simulations solvated with TIP3P for all force fields. (B) Final frames of the unbound state of the host solvated by TIP3P and the crystal structure (PDB: QQ7).*

Parsley, also known as the open force field, is the most recent of the force fields examined here and was designed to overcome some known limitations of older force fields, including, for example, the calculation of hydration-free energies and partition

coefficients.[223, 224, 247] CGenFF was initially developed based on the CHARMM biomolecular force fields, and it covers a wide range of chemical groups and drug-like molecules.[232] Another force-field generation tool, SwissParam, generates topologies and parameters derived from MMFF, compatible with the CHARMM force fields, but SwissParam has been developed for rapid calculations, i.e., docking, minimization, etc. Thus, it may not be suitable for long MD simulations, despite being used in many MD studies.[225, 248, 249] Accordingly, it is useful to include this comparison here. Finally, I have included GAFFv1 and its second generation, GAFFv2, compatible with the AMBER force fields. These two force fields introduce similar dynamics patterns to the host molecule, but the performance is not the same for the enthalpy calculation. Indeed, the presence of the large distortions observed for CGenFF and Parsley does not seem correlated to the performance of enthalpy prediction.

Table 3.3: *Binding enthalpies (ΔH, kcal/mol) for TrimerTrip and the heptane-1,7-diamine guest molecule.*

Method	**Enthalpy (kcal/mol)**
Experimental ΔH	-10.1 ± 0.04
CgenFF	-1.10 ± 0.60
GAFFv1	-9.83 ± 0.63
GAFFv2	-9.36 ± 0.62
Parsley	-7.22 ± 0.63
SwissParam	-5.57 ± 0.60

The observations of the potential flexibility of CB7 prompted me to assess enthalpy predictions on a host predicted to have a larger range of conformational fluctuation. The acyclic cucurbituril-derived TrimerTrip clip molecule is more flexible compared to the CB7 host molecule (**Figure 3.5**). In this analysis, I randomly picked a guest molecule (heptane-1,7-diamine) having a known experimental binding enthalpy value.[220] MD simulations using only the TIP3P water model were performed since water models have no obvious effect on the host dynamics, and TIP3P is considered a "default" water model. All simulations nicely converged the potential energy illustrated by **Figure 3.5B.** GAFFv1 and GAFFv2 successfully computed the binding enthalpy very near the experimental value (−10.10 ± 0.04 kcal/mol), but the performance of CGenFF was relatively poor. Both SwissParam and Parsley underestimated the binding enthalpies (**Table 3.3**). The pattern of conformational flexibility observed for CB7 where Parsley and CGenFF have more structural deviation compared to the others (**Figure 3.4**) was not observed for this system (**Figure 3.5B**).

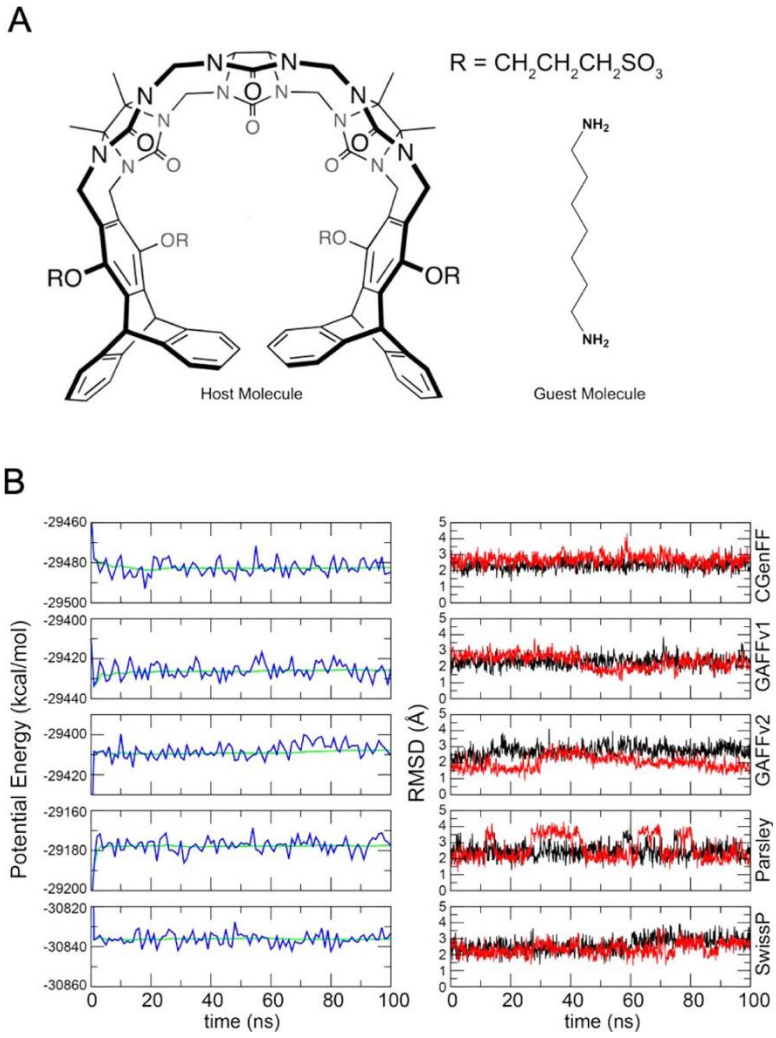

Figure 3.5: *(A) Structure of TrimerTrip and the guest molecule. (B) Convergence (green) and change (blue) of the potential energy for the complex systems and RMSD of heavy atoms in the unbound (black) and bound (red) states of the host molecule, TrimerTrip, through the simulations solvated with TIP3P for all force fields.*

There are two other tools for topology and parameter generation that I additionally considered in this study. One of them is LigParGen[250], designed to provide OPLS-AA parameters for use directly in popular MD machines. I obtained all topologies and parameters for the molecules in my benchmark using a locally implemented version of LigParGen software since the maximum ligand size allowed is 200 atoms on the server. Unfortunately, all host–guest pairs did not stay together during the simulations when these topologies and parameters were used. Further investigation revealed that during the host parameterization process, the tool assigns a regular amide carbon (atom-type CAM) to the carbon atoms bonded to the oxygen in the host molecule. This oversimplification may be responsible for the behaviour in these simulations, but confirmation of that would require a more detailed investigation and is beyond the scope of this work. Thus, I could not report

any results from this tool for the OPLS-AA force field at this stage. Another tool is the Automated Force Field Topology Builder (ATB)[251], which provides topologies and parameters for various molecules compatible with the GROMOS force fields. This server performs a multistep process containing a series of quantum mechanical (QM) calculations combined with a knowledge-based approach. This server can also provide topologies and parameter files that the GROMACS MD engine can read. However, these simulations did not run to a successful conclusion, suggesting the out-of-box results requiring manual intervention, which I was keen to avoid in this assessment. Furthermore, the GROMOS force fields are no longer supported within GROMACS, suggesting other issues about this force field.[252]

3.3.3. Performance of the Water Models

Water has many unique and unusual properties and is thus probably the most extensively studied molecule.[207, 253] Our understanding of this tiny molecule is still incomplete because of its extraordinary properties.[254, 255] The complexity of its properties has led to the design of many theoretical and computational models.[173, 176] In biomolecular simulations, a simulator should carefully select the water model with the most appropriate force field[211] because it is not always obvious whether the water model is suitable under the simulation conditions. Assessment studies of the models to replicate experimental data are crucial to help select the best water model, but such studies are relatively few.

TIP3P and its derivative Bind3P give the most accurate results compared to other water models, although it must be remembered that different water models may be appropriate for different uses, and it is challenging for any one model to reproduce all of the unique properties of water.[226] A sensitivity analysis approach was used to optimize the Lennard–Jones parameters of the original TIP3P for the development of the Bind3P water model.[57] Bind3P gives the most accurate result (i.e., smallest error) when used in conjunction with the Parsley force field. However, perhaps surprisingly, the best correlation is found when it is used in conjunction with GAFFv1. TIP3P is the best water model when working with GAFFv1. The OPC water model fairs well when used with GAFFv1 (RMSE = 2.07 kcal/mol). TIP5P is generally a very poor performer regardless of the host force-field choice.

The TIP5P water model is comparatively less utilized in molecular dynamics simulations than the TIP4P or TIP3P water models due to its higher computational demand. A comprehensive evaluation of common water models used in molecular dynamics simulations revealed that four-point water models, OPC, TIP4P/ε, and TIP4P-FB, best agreed with experimental data of the structural and dynamic properties.[256] Nonetheless, TIP5P demonstrated the highest level of agreement among the five-point water models. While the TIP5P model accurately replicates structural properties better than other models, it is unable to reproduce the total energy and molar heat capacity compared to TIP3P.[257] Additionally, it is not successful in replicating isobaric heat capacity compared to OPC and TIP3P water models.[174, 258] Thus, researchers should carefully consider specific features of the water models when selecting an appropriate water model for their particular simulation needs.

3.3.4. Impact of Aromatic versus Aliphatic Guests on the Assessment

The guest molecules except CTE, FTE, and NTE can be subgrouped into aromatic and aliphatic, with both groups having 11 guest molecules. I reanalyzed the performance of all force fields and water models according to these groups. All water models tend to produce more accurate and correlated results in aliphatic guest molecules than aromatics ones. Bind3P produces the best results with an RMSE = 2.11 ± 0.41 kcal/mol and an R^2 = 0.82 ± 0.06 averaged across all force fields, while it gives an average RMSE = 3.91 ± 0.65 kcal/mol and an R^2 = 0.30 ± 0.10 for aromatic molecules. GAFFv1 is the best choice for both groups. GAFFv2 has a good ranking ability for aliphatic molecules even though it produces more accurate results for aromatic guests with an RMSE = 5.26 ± 0.67 kcal/mol. The Parsley–Bind3P is the best combination for aliphatic guest molecules with an RMSE of 1.06 kcal/mol, whereas the GAFFv1–OPC combination gives the most accurate result with an RMSE of 1.37 kcal/mol for aromatic molecules. TIP3P (R2 = 0.95) and SPCE (R2 = 0.72) with Parsley in aliphatic and aromatic molecules are the most correlated combinations.

Interestingly, CGenFF produces more accurate results in aliphatic guests than SwissParam. However, SwissParam's ranking ability prevails over CGenFF. Overall, this analysis suggests that there is still much room for improvement in the treatment of aromatic molecules concerning the enthalpy calculation.

3.4. Conclusions

In this chapter, the binding enthalpy of 25 guest molecules with CB7 is computed by performing MD simulations using 40 different force-field and water model combinations. The computed binding enthalpies using the TIP3P and its derivative, Bind3P, are in good agreement with the experimental values. GAFFv1, perhaps surprisingly, outperforms the other force fields, producing more accurate results than other force fields. Generally, all force fields produce better results for aliphatic than aromatic guests, suggesting that there is scope for improvement in treating aromatic molecules. In these and similar studies, I was not able to include commercially modified force fields. It would be extremely useful to find ways to assess these as well in my efforts to improve force-field performance. Nevertheless, the findings here should still be useful in ongoing attempts to improve force fields and how well they reproduce the thermodynamic properties of binding events.

4 Accurate Prediction of Ligand-Protein Binding Enthalpies via Consideration of Loop Dynamics

This chapter explores the enthalpic components of ligand-protein binding free energy and challenges in prediction. The performance of absolute enthalpy of binding calculations for ten inhibitors against a member of the bromodomain family, BRD4-1, is evaluated using molecular dynamics simulation and isothermal titration calorimetry data. The content of this chapter has been published in the following journal article[259]:

Çınaroğlu, S. S., & Biggin, P. C. (2023). The role of loop dynamics in the prediction of ligand–protein binding enthalpy. *Chemical Science*, 14(24), 6792-6805. doi: 10.1039/D2SC06471E

4.1. Introduction

In recent years, significant progress has been made in predicting the binding free energy of small ligands for protein receptors.[260-264] In contrast, there has been relatively little progress in the accurate computation of the underlying thermodynamic components, namely the enthalpy (ΔH) and entropy ($T\Delta S$). Accurate computation of enthalpy has historically been viewed as particularly challenging[34, 265-267] due to the large fluctuations in potential energy systems tend to undergo. Thus, any estimates of the mean value would likely require vast amounts of sampling. Nevertheless, having a reasonably reliable estimation of the enthalpy, and more importantly, the error estimate, would be extremely useful in understanding the role of the underlying contributions, especially in the context of drug design.[26, 268] From a medicinal chemistry point of view, enthalpic contributions are perhaps intuitively easier to understand and conceptualize than the entropic components. During a fragment or lead compound elaboration, where ΔG is often being optimized, it would be extremely valuable to know during those steps that the changes made to the compound were indeed giving the expected improvement to the ΔG via the designs suggested by the medicinal chemist. Often,

such designs focus on improving interactions between chemical moieties, expecting a gain in favourable enthalpy. Computing and confirming this as part of the optimization process would be extremely useful. Moreover, being able to compute enthalpy reliably should provide quantitative insight into the phenomenon of entropy-enthalpy compensation.[36]

The increase in computational power over the past decade has meant that the accurate calculation of binding enthalpies might soon be realized, and work in particular from the Gilson group on small host-guest and other systems[2, 31, 35, 53] has shown strong potential. However, despite the promising results obtained for small model systems, the calculation of enthalpy contributions for larger proteins has remained challenging, and it is not even known what level of performance could be obtained even for well-characterized protein-ligand systems, such as bromodomains.[76, 112]

Thus, to evaluate the performance of enthalpy calculations for protein-ligand binding, I assembled a data set based on bromodomains in a similar vein to that which Aldeghi et al. had previously done for absolute binding free energy (ABFE) calculations.[112] Bromodomains (BRDs) are protein-protein interaction modules that selectively recognize acetylated lysine (K_{ac}) residues as a key event in the epigenetic reading process. A total of 61 human BRDs has been identified in 46 proteins, consisting of eight protein families.[269] Despite these many different families, all BRDs have a conserved structure that contains a left-handed bundle of four α helices (α_Z, α_A, α_B, α_C), linked by loop regions (ZA and BC loops), which surround the K_{ac} binding site (**Figure 4.1a**). Among eight families present in human proteome, the bromodomain and extraterminal (BET) family is characterized by two tandem N-terminal BRDs and an extraterminal (ET) domain, and is composed of BRD2, BRD3, BRD4, and BRDT.[270] BRD4 is thus a representative member of the BET family and has roles in activating critical genes involved in cell growth and cell cycle progression.[271]

It is perhaps not surprising that BRD4 has also been implicated in inflammation and cancer progression resulting in many inhibitor and probe molecules being developed [269], and indeed, this is an ongoing activity. Given the wealth of existing biophysical and structural data and the desire to develop yet more probes

with improved selectivity, bromodomains represent an ideal test system for computational studies.[112-114, 272-280] In most complexes, the binding pocket does not show major differences in conformation, which probably contributed to the success of rigorous free energy methods. However, it has recently been suggested the bromodomain fold is quite dynamic, including the relatively recent suggestion of cryptic pockets.[275, 279, 281]

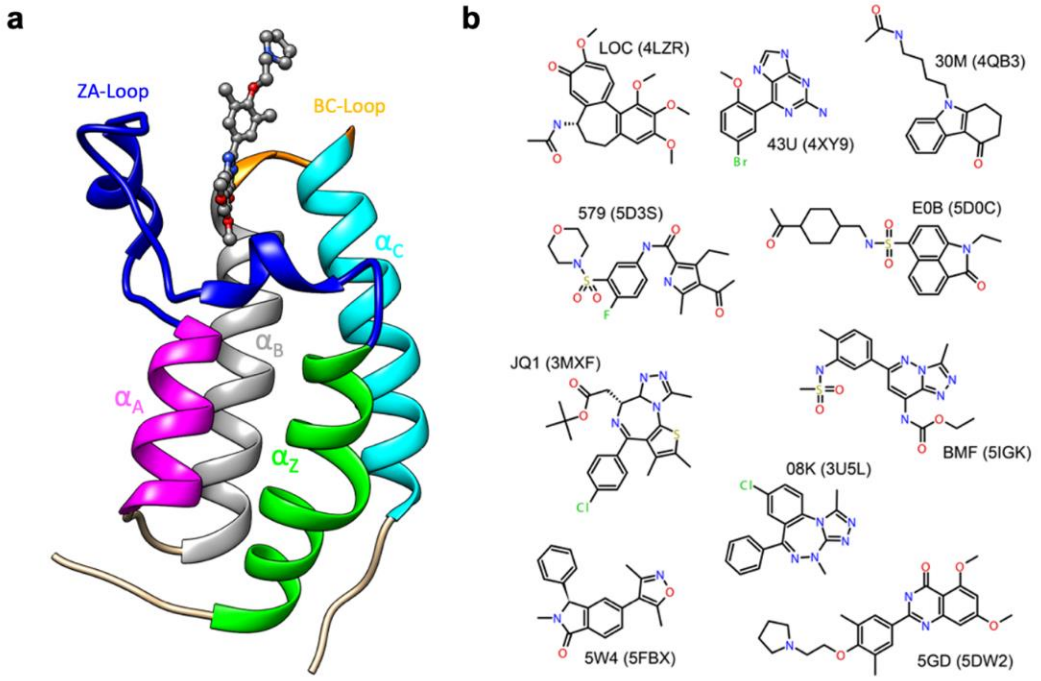

Figure 4.1: (a) *Cartoon of BRD4-1 with ligand complex (PDB:5DW2).* ***(b)*** *Chemical structures of compounds with their three-letter identifiers and corresponding PDB IDs for the complexes.*

The first bromodomain of BRD4, BRD4-1 has extensive biophysical and structural data (Table S4.1). I reasoned this would provide an ideal test case for investigating current ability to predict enthalpy and what additional insight into binding thermodynamics we might also gain. Here, I perform binding enthalpy calculations using the direct method for a non-redundant set of BRD4-1 and ligand complexes (**Figure 4.1b**)[31, 34, 35]. The results show that absolute enthalpy can be calculated for such systems to an error of about 1 kcal/mol. Furthermore, the source of errors in initial outliers can clearly be identified and provides new insight into the well-known problem of enthalpy-entropy compensation (**Figure S4.1**). In particular, the role of a key loop (the ZA loop) near the binding site is discussed.

4.2. Methods

4.2.1. Building the Benchmark

Initially, all ITC data was collected from the literature, and ITC entries with PDB structures were filtered for further steps (**Table S4.1**). Ligands in the PDB structures were clustered using binning clustering with default 0.4 similarity cut-off in the ChemMine Tools server.[282, 283] Representative PDB structures with the best resolution were selected from each cluster for the final benchmark. I use the PDB codes as representing names for all ligands throughout the manuscript for brevity and ease of referral to structures.

4.2.2. System Setup

The initial conformations were taken from crystal structures (3MXF, 3U5L, 4LZR, 4QB3, 4XY9, 5D0C, 5D3S, 5DW2, 5FBX, 5IGK, 2OSS). Missing atoms in the crystals were modeled with the DockPrep tool in UCSF Chimera [284] and all heteroatoms were removed from the system except the ligand of interest and all crystallographic waters. The N-terminal tail was deleted up to residue Asn54 to reduce computational cost and complexity of the simulations. Terminal residues were patched with acetylated N-terminus and amidated C-terminus using PDB Reader of CHARMM-GUI. [285, 286] Ligand molecules were parameterized with the general AMBER General Force Field for organic molecules (Version 2.11, May 2016) [287] and AM1-BCC[229, 230] charges using AmberTools19. I used the Amber ff14SB force field for the protein and the TIP3P water model for water molecules.[288] A periodic cubic water box was used for all systems with 20,000 water molecules for the complex and receptor-only simulations, while 2000 water molecules were used for ligand and solvent-only simulations.

4.2.3. Quantum-Optimized Parameters

Optimized GAFF2 parameters for 08K(3U5L), 30M(4QB3), BMF(5IGK) and three different ionization states of HEPES were obtained by using the Psi4 *ab initio* quantum engine at the HF/6-31G* level of theory.[289] Atomic charges were fitted to reproduce the electrostatic potential (ESP). All steps for getting optimized parameters and charges were performed using the parameterize parameterization tool, which attempts to improve the quality of parameters.[290] All parameters are available at doi: 10.5281/zenodo.7534582.

4.2.4. Absolute Binding Enthalpy Calculations

Computational calculation of the binding enthalpy requires 2 set of molecular dynamic simulations (the bound- and unbound state of the complex). The binding enthalpy (ΔH) is calculated by Equation 1, where $\langle \mathrm{E} \rangle_{complex}$, $\langle \mathrm{E} \rangle_{solvent}$, $\langle \mathrm{E} \rangle_{receptor}$, and $\langle \mathrm{E} \rangle_{ligand}$ are the averaged potential energies of the system from four separate simulations. In this method, the number of atoms between the bound and unbound state of the complex should exactly balance. Note that the pressure–volume contribution for the binding enthalpy is negligible.[31] A detailed description was written in section **2.4**.

$$\Delta H = \langle \mathrm{E} \rangle_{complex} + \langle \mathrm{E} \rangle_{water} - \langle \mathrm{E} \rangle_{receptor} - \langle \mathrm{E} \rangle_{ligand} \qquad (4.1)$$

All simulations were performed using the Gromacs v2020.3 software package.[30, 142, 291, 292] A 3-step steepest descent energy minimization with a maximum force of 10 kJ/mol/nm^2 was applied to all systems. In the first step, position restraints with a harmonic potential with a force constant of 1000 kJ/(mol/nm^2) were applied for all heavy atoms, then removed for solute heavy atoms, and the final step removed all restraints. NVT and NPT ensemble simulations for 1 ns were performed to equilibrate all systems with position restraints with the harmonic potential at a force constant of 1000 kJ/(mol/nm^2) on heavy protein and ligand atoms. Additionally, another NPT ensemble simulation for 1 ns was performed without restraints before the production run for data collection. The V-rescale and Parrinello-Rahman algorithms equilibrated the temperature at 300 K and the pressure at 1.0 bar, respectively. Unbonded interactions were calculated up to a cut-off of 1.0 nm with a potential-shift. A dispersion correction was applied to energy and pressure. All H-bond lengths were constrained with a LINear Constraint Solver (LINCS) algorithm. Coulomb interactions were evaluated with the Fast smooth Particle-Mesh Ewald (SPME) electrostatics method with an initial short-range cutoff of 1.0 nm. The leap-frog algorithm was used to run 20 independent 100 ns MD simulations with 2 fs time step.

The average of the potential energy and the standard error estimate were calculated by performing re-blocking analysis using the pyblock tool (https://pyblock.readthedocs.io)[193] for all individual calculations. A detailed description was written in section **2.6**.

4.2.5. Absolute Binding Free Energy Calculations

Absolute binding free energy calculations were performed using conformations obtained from the binding enthalpy simulations. For six complexes (3U5L, 4LZR, 4QB3, 4XY9, 5DW2 and 5IGK), I performed two sets of simulations reflecting the different conformations of the ZA loop. MDRestraintsGenerator, which is a framework for generating restraints for MD simulations, was used to provide the optimal Boresch restraints (1 distance, 2 angles and 3 diehderal harmonic restraints (10.5281/zenodo.6977294). The nonbounded ligand interactions were decoupled using a linear alchemical pathway for the van der Waals and the coulombic transformations with $\Delta\lambda = 0.05$ and 0.1, respectively. The ligand restraints transformation had 12 non-equally distributed λ values (0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.35, 0.5, 0.75, 1.0). Each calculation for absolute binding free energy comprised a total of 44 windows for the complex simulations and 32 windows for the ligand simulations. Each window was completed with 5-step runs. Firstly, energy minimization was carried out using the steepest descent algorithm with 10000 steps. Then, 10 ps NVT ensemble was performed using a leap-frog stochastic dynamics integrator with harmonic position restraints on the solute-heavy atoms with a force constant of 1000 kJ mol^{-1} nm^{-2}. After that, 100 ps isotropic ensemble using the Berendsen coupling algorithm was run with the same position restraints. Moreover, another NPT ensemble with the Parrinello–Rahman coupling algorithm was performed for 100 ps without position restraints. Finally, 10 ns production runs were performed for data collection. A detailed description about ABFE is found in section **2.5**.

4.2.6. Constructing the Unit Cell

A crystal unit cell for 2OSS was built to investigate the effect of crystal packing on the ZA-loop conformation. The Cell Unit tool in UCSF Chimera was used for constructing the unit cell for 2OSS with the $P2_12_12_1$ space group. Missing atoms in the crystals were modeled with the DockPrep tool in UCSF Chimera and 1,2-ethanediol molecules were removed from the system while crystallographic waters were kept. The cell unit contains 4 chains with lengths 37.418, 44.139, 78.413 on the xyz dimensions. I used three different force field including Amber FF14SB[288], CHARMM36[293], and OPLS-AA/M[294] force field. Simulations were performed with 3 replicates for the crystal lattice and 12 replicates for the single chains.

4.3. Results

4.3.1. Collating a non-redundant BRD4-1 benchmark data set

I first collected all binding data with thermodynamic components derived from ITC for BRD4-1 from the literature for this study. Then, only data sets that could be linked to high-resolution (better than 2.0 Å) were retained. Given the interest in generating probes and drug molecules against BRD4-1, many studies report lead-optimization attempts[295, 296]. This leads to many similar ligand molecules within the data set and a significant redundancy issue. To obtain a non-redundant dataset, ligand molecules were clustered using binning with a 0.4 similarity cut-off with the ChemMine Tools server. [282, 283] After clustering, I had 14 different clusters with cluster sizes ranging from 1 to 11 (**Table S4.1**). Four clusters, each with a single protein (4OGI, 4OGJ, 5EGU, 5BT4), were filtered out since the ligand bridged interactions between two protein chains in the crystal lattice and this was deemed artificial. Then, representative PDB structures with the best resolution were selected from each remaining clusters for the final benchmark (**Table 4.1 and Figure 4.1b**).

Table 4.1: *Non-redundant BRD4(1) benchmark*

PDB ID	Resolution (Å)	Space Group	Ligand ID	ΔH	TΔS	ΔG	Reference
3MXF	1.60	P 2_1 2_1 2_1	JQ1	-8.42	1.22	-9.64	Filippakopoulos *et al.*
3U5L	1.39	P 2_1 2_1 2_1	08K	-6.16 ±0.03	2.00	-8.16	Filippakopoulos *et al.*
4LZR	1.85	P 2_1 2_1 2_1	LOC	-9.00	-2.60	-6.40	Lucas *et al.*
4QB3	0.94	P 2_1 2_1 2_1	30M	-6.62 ±0.03	0.93	-7.55	Gacias *et al.*
4XY9	1.83	P 2_1 2_1 2_1	43U	-6.09 ±014	0.94	-7.03	Picaud *et al.*
5D0C	1.49	P 1 2_1 1	E0B	-10.20	-2.52	-7.68	Xue *et al.*
5D3S	1.75	P 2_1 2_1 2_1	579	-9.77	-1.73	-8.04	Hügle *et al.*
5DW2	1.12	P 2_1 2_1 2_1	5GD	-10.10	2.10	-8.00	Raux *et al.*
5FBX	1.85	P 2_1 2_1 2_1	5W4	-15.57 ±0.54	-4.66	-10.90	Montenegro *et al.*
5IGK	1.70	P 2_1 2_1 2_1	BMF	-11.09 ±0.4	-1.36	-9.73	Picaud *et al.*

All values are in kcal/mol.

4.3.2. Absolute Binding Enthalpy Calculations

Absolute calculation of binding enthalpy using the direct method requires a set of four simulations, including bound and unbound states of the system. In assessing binding enthalpies, sampling all conformational space is the key factor to achieve sufficient convergence of the potential energy. Relative binding enthalpies have

been reported for protein-ligand systems using the direct method.[34, 35] In previous studies, Roy *et al*. performed 40 independent 10 ns simulations to get sufficient sampling[34]; moreover, Li and Gilson reached over 250 μs simulations by seeding every 200 ns block with a new random number for the relative binding enthalpy calculation of a protein-ligand system.[35] Here, I performed 20 completely independent repeats of 100 ns simulations for each system. To assess convergence, I employed the blocking method[193], where the enthalpy is averaged over successively larger blocks and for each block size the standard error of the mean is computed. As discussed by Henriksen et al.[2], in an ideal case the SEM will display a plateau, but this is not always the case, and it is not easy to automate the detection of such a plateau either. Therefore, again following the work of Henriksen et al. [2], I took the maximum SEM value (**Figure S4.2 and Table S4.2**) to err on the side of caution. For most complexes, the maximum SEM is ~0.6 kcal/mol, but even the maximum (for 5FBX) is ~1.1 kcal/mol. As expected, the ligand and solvent-only profiles converge earlier than the complex.

From these initial simulations, the correlation with the experiment was moderately good with an $R^2 = 0.60$, and an average of Kendall's $\tau = 0.42$. The accuracy of the calculation is perhaps surprisingly good, with an average of root-mean-square error (RMSE) = 2.49 kcal/mol. Most binding enthalpies are within 2 kcal/mol absolute difference of experimental values (**Figure 4.2 & Table S4.2**). The best binding enthalpy predictions were obtained for 5D3S with the XD44 (4-acetyl-3-ethyl-N-[4-fluoro-3-(morpholin-4-ylsulfonyl)phenyl]-5-methyl-1H-pyrrole-2-carboxamide) ligand (579) and 4LZR bound to colchicine (LOC). Conversely, 3U5L in complex with a benzo-triazepine ligand (08K), 4QB3 with olinone (30M) and 5IGK with bromosporine (BMF) provided the worst binding enthalpy predictions as outliers (**Figure 4.2 & Table S4.2**).

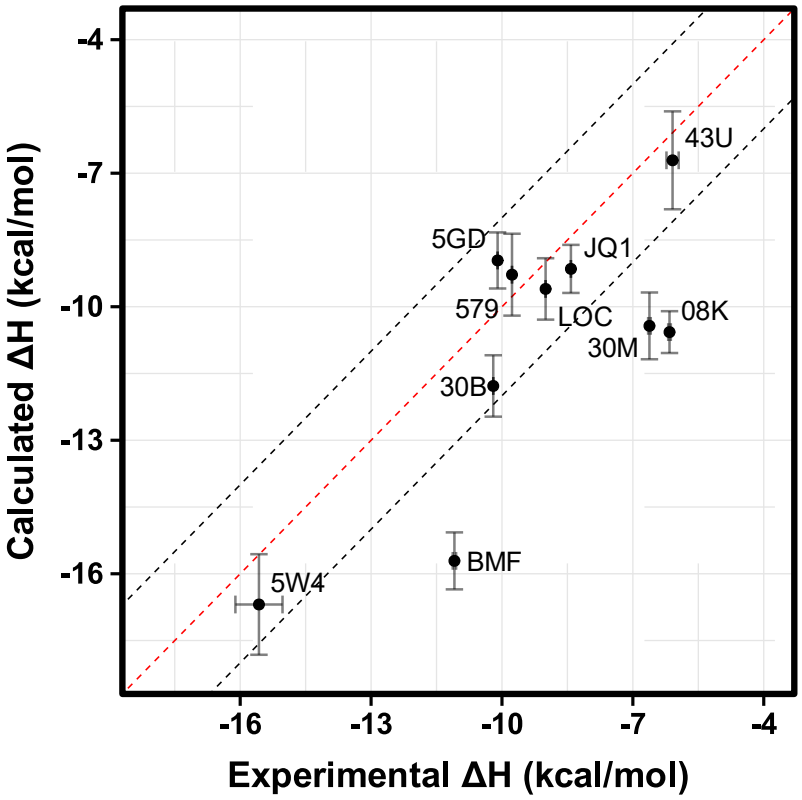

Figure 4.2: *Comparison of calculated binding enthalpies from experimental values. Error bars show SEM of the mean. The line of equivalence is shown in red and the black dashed lines indicate the 2 kcal/mol error limit. Some error bars are invisible due to missing or too small error values (**Table 4.1**). Three letter codes are the ligand codes as in **Table 4.1** and **Figure 4.1**.*

Although some of the computed enthalpy values are in excellent agreement with the experimental data, there are some clear outliers (08K, 30M and BMF), all of which are overestimates. I thus then sought to investigate these further and first considered improved parameterization. I optimized ligand charges for the outliers using the Psi4 *ab initio* quantum engine with the HF/6-31G* level of theory. After optimization, there was good agreement of the energy profile of GAFF2 parameters with the reference quantum mechanics (QM) calculations (**Figure S4.3**). Then, all simulations with 20 replicates, with a total of 120 simulations, were rerun using the optimized parameters for 3UL5, 4BQ3, and 5IGK. The binding enthalpy calculation for 4QB3 was improved at -5.13 ± 0.72 kcal/mol, which is closer to the experimental value and within 2 kcal/mol. However, both 3UL5 and 5IGK still remained as outliers. The predicted binding enthalpy for 5IGK with bromosporine was slightly improved, but for 3UL5, the values actually got worse and increased the absolute

difference to 7.69 kcal/mol from the experimental value (**Table S4.3**). Thus, the impact of re-parameterization was minimal.

I next considered the role of the buffer since I have performed all simulations with pure water to reduce complexity. In contrast, the ITC experiments were mostly performed in 50 mM HEPES and 150 mM NaCl solution (**Table S4.1**). However, binding thermodynamics may be sensitive to the solvent composition for both experimental[297, 298] and computational[53, 76] studies. To investigate the role of the buffer and to what extent replicating the conditions of the experiments influenced the calculations, I set up simulations with three different ionization states of HEPES with 3 molecules of species A, 6 molecules of species B, and 9 molecules of species C, as shown in **Figure S4.4** and NaCl for the apo-receptor (2OSS) and the 3U5L and 5IGK complexes. Force field parameters of HEPES were obtained using Psi4 with the HF/6-31G* level of theory, while Na^+ and Cl^- parameters were used as provided in the Amber ff14SB force field. I then performed further simulations with 20 replicates with a total of 60 simulations of the apo-receptor (2OSS), 3U5L and 5IGK in 50 mM HEPES and 150 mM NaCl solution. Although I obtained sufficient convergence of potential energy for these simulations (data not shown), the uncertainty increases, as expected, because HEPES and NaCl make the system more complex, requiring longer simulations or more replicas. However, the accuracy of enthalpy prediction itself remained poor. The enthalpy for 3U5L was slightly improved but 5IGK gave a worse result than previous calculations (**Table S4.3**). Thus, I concluded that explicit treatment of buffer in the calculations was not the main reason for large deviations from experimental data.

4.3.3. The ZA-loop adopts an alternative conformation that strongly affects binding enthalpy.

Simple observation of trajectories revealed a significant structural deviation in the ZA loop of some simulations, especially the apo structure, 2OSS (**Figure 4.3a-c**). Whilst some ligands appear to stabilize the ZA-loop in the crystal-like conformation (5IGK, 5FBX, 5D3S, 5D0C, and 3U5L simulations), it is clear that the others afford the ZA loop a greater level of dynamics as evidenced by simple root-mean squared deviation (RMSD – **Figure 4.3b**). Closer inspection revealed that in fact the ZA-loop can move to a distinct and alternative conformation, which in the case of the

apo (2OSS) state, exists for approximately 75% of the 2 µs simulation time (**Figure 4.3c**). In this alternative conformation, the ZA-loop moves outwards away from the acetyl-lysine binding pocket and induces a short helix (residue 88–91) within the ZA-loop. This outward movement makes the binding pocket open and more accessible. To investigate the role of this loop behaviour on the enthalpy I extended the number of repeats of the apo state to 100, to ensure that I obtained sufficient sampling of the crystal-like conformation of the ZA-loop. For the remainder of the discussion, I refer to the crystal-like conformation of the ZA-loop as **ZA1** while the alternative ZA-loop conformation as **ZA2.**

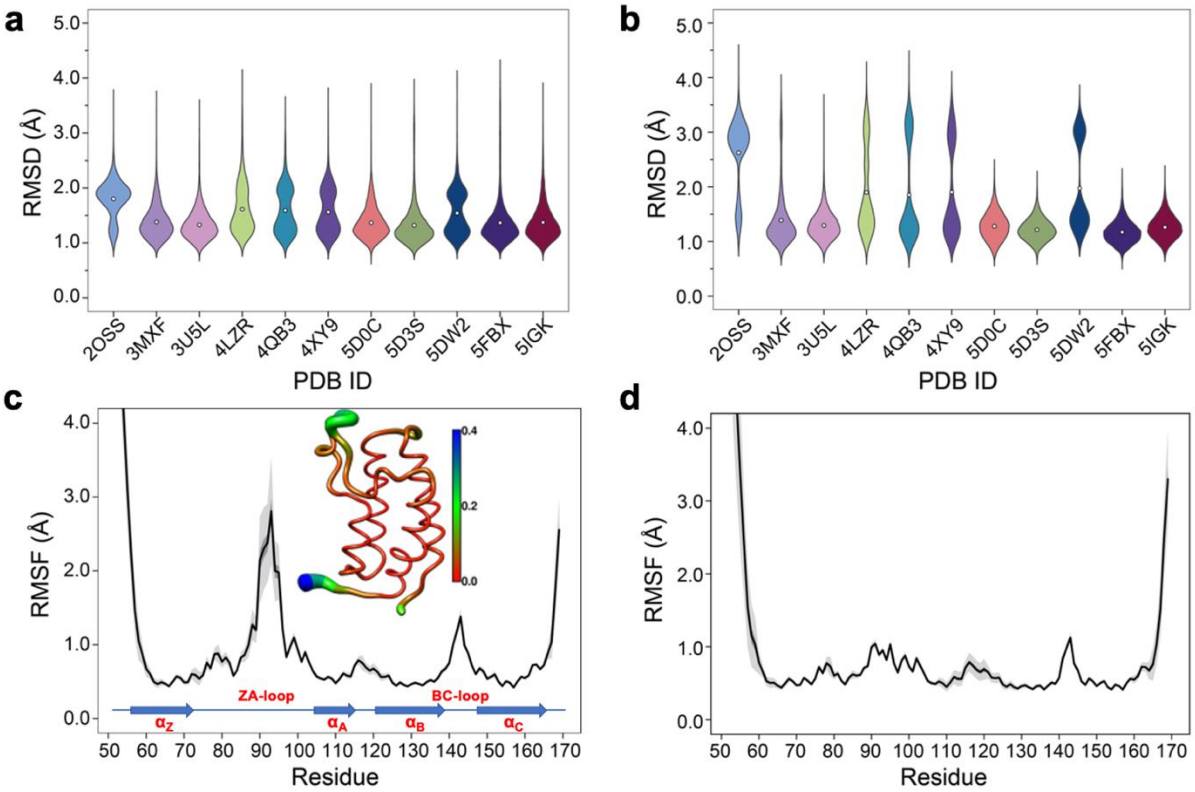

Figure 4.3: ***(a)*** *RMSD violin plots for backbone atoms of 2OSS apo-receptor and complex simulations, white circle represents the overall mean value of all 20 repeats.* ***(b)*** *RMSD violin plots for ZA-loop (76-106 residues -see Fig. 1.) backbone atoms of 2OSS apo-receptor and complex simulations after fitting whole protein backbone.* ***(c)*** *RMSF for backbone atoms of 2OSS apo-receptor. The inset figure shows regional fluctuations on the BRD4-1.* ***(d)*** *RMSF for backbone atoms of 5IGK complex simulations. Black line gives average RMSF of 20 repeats while grey shade represents the standard deviation.*

After completing 100 replicates of the apo-receptor (a total of 10 µs of simulation), I checked the difference in potential energy between the ZA1 and ZA2 conformational states, which was 0.84 ± 0.2 kcal/mol. Previous work has reported that the barrier between these states is of the order of ~2kcal/mol .[114] In addition

to the apo-receptor (2OSS) displaying this alternative ZA2 loop conformation, so did some ligand-bound simulations including 4LZR, 4QB3, 4XY9, and 5DW2 (**Figure 4.3b, Figure S4.5**). I also checked the potential energy difference between ZA1 and ZA2 (**Table S4.4**) for these complexes. The 4LZR complex gave the biggest potential energy difference with 5.77 ± 1.5 kcal/mol while 4XY9, at 1.83 ± 1.93 kcal/mol, had the lowest difference amongst these four complexes. Only 5DW2 exhibited lower potential energy for the ZA2 conformation of these four complexes.

The existence of these significant differences of potential energy thus raised the question of how these two conformations affect binding enthalpy calculations. To explore this, I calculated binding enthalpies using ZA1 or ZA2 conformations exclusively for the 4LZR, 4QB3, 4XY9, and 5DW2 complexes (**Figure 4.4 & Tables S4.4, S4.5**). Using ZA1 or ZA2 conformations for 4LZR and 4XY9 exclusively gave less accurate predictions than using both conformations. 4QB3 gives a more accurate enthalpy prediction with ZA1 than with ZA2 alone or use of all simulation data combined. This also explains why the simulations with QM-refined parameters (**Table S4.3** and **Figure S4.3**) and gave more precise results than the initial runs, since the occupation of the ZA1 conformation in the second simulation set was more than the first one (**Table S4.5**). Conversely, 5DW2 interestingly gave more accurate binding enthalpy estimates when only the ZA2 conformation was used.

As the remaining outliers, 5IGK and 3U5L, were comprised solely of ZA1 conformations in all 60 replicates and given the above indication of the importance of the ZA loop behaviour, I decided to initiate replicates for these two complexes starting from the ZA2 conformation. For this purpose, I extracted a snapshot having the ZA2 conformation from a random apo-receptor simulation, then I manually docked BzT-7 (08K) of 3U5L and bromosporine (BMF) of 5IGK to the receptor via superimposition and performed 20 repeats of 100 ns. The ZA-loop stayed as the ZA2 conformation in all simulations for 3U5L (**Figure S4.9**) and almost all simulations for 5IGK (there was a transition from ZA2 to ZA1 in the last 20 ns of only one simulation). I obtained sufficient convergence of potential energy for these simulations and then calculated binding enthalpies. Surprisingly, 3U5L gave a

highly accurate prediction with -6.52 ± 0.64 kcal/mol, which is 0.36 kcal/mol away from the experimental value (-6.16). However, I did not obtain an improvement in accuracy of enthalpy for 5IGK with ZA2, indeed the experimental value is nearly in the middle of the predicted binding enthalpies for ZA1 and ZA2, suggesting that both conformational states contribute to the enthalpy. To confirm this, I combined simulation data for both ZA1 and ZA2 from 40 simulations in total and recomputed the binding enthalpy. The combined simulation data gave a prediction of -11.29 ± 3.07 kcal/mol, which is 0.20 kcal/mol away from the experimental value (-11.09 kcal/mol) (**Figure 4.4a**). Together, these computations reveal a hidden complexity of binding thermodynamics in what might be considered a relatively simple system.

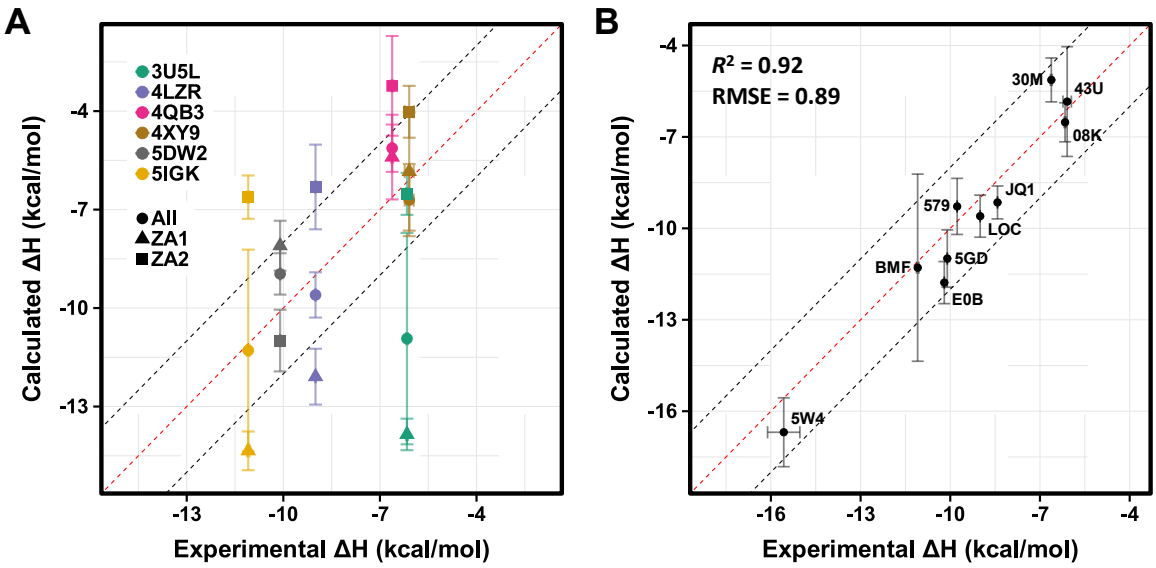

Figure 4.4. *Comparison of calculated binding enthalpies to experimental values.* ***(a)*** *Values obtained by considering only ZA1 (triangles) or ZA2 (diamonds) or all (circles) ZA loop conformations for outliers and complexes observed to adopt alternative conformations of the ZA loop* ***(b)*** *The best calculated binding enthalpy values that can be obtained for the whole dataset. Error bars show the maximum standard error the mean estimate. Some error bars are invisible due to missing or small error values (**Table 4.1**).*

4.3.4. The Relationship to the Absolute Binding Free Energy

Given the influence of the ZA loop conformation on the enthalpy predictions, I was interested to see to what extent the ZA loop conformation also affected the absolute binding free energy, ΔG. Thus, for the four complexes that displayed alternative ZA loop conformations (see **Figure 4.3b**); 4LZR, 4QB3, 4XY9 and 5DW2, along with the two outliers 5IGK and 3U5L (where loop stability is greater for ZA1 **Figure S4.7**), I computed the binding free energies (**Table 4.2**). For four of the complexes (4LZR, 4QB3, 4ZY9 and 5DW2) the loop conformation that favours the lower ΔG

is mirrored by the enthalpy results (**Table S4.4**). The 3U5L complex, however, is a more complicated result. Whereas the enthalpy calculations suggest that the ZA2 conformation leads to better agreement with the experiment (**Figure 4.4a**), the free energy calculations give a lower binding free energy for the ZA1 conformation (albeit heavily overestimated compared with the experiment). Similar trends for 3U5L have been reported by Heinzelmann *et al*.[114] using a different approach, the Attach-Pull-Release method, which produced an over-estimated result (–10.61 kcal/mol) for ZA1 but a closer-to-experiment result (–8.07 kcal/mol) for ZA2. Furthermore, an almost similar result (−9.9 ± 0.8) for ZA1 has been observed by Aldeghi et al[112] and also by Bertin[299] (−9.1 ± 0.3) (https://thesis.unipd.it/handle/20.500.12608/21280) using the alchemical decoupling free energy method. Together, these results suggest that the 08K ligand in 3U5L tends to give an overestimated binding affinity when the crystallographic conformation (ZA1) is used. Compared with other ligands that bind BRD4(1), the ligand has modest enthalpic contributions to the binding free energy, but one of the most favourable entropic contributions (**Table 4.1**). Analysis of the energetic components of the enthalpy (**Table S4.6**) shows a large coulombic contribution in the ZA1 conformation that is almost completely absent in ZA2. In the case of 5IGK (bromosporine complex), the ΔG values are higher than experiment for the ZA1 loop conformation but lower for the ZA2 conformation(**Table 4.2**). The value obtained for the ZA1 conformation in this work is completely consistent with the value Aldeghi et al. obtained in previous work[113] (and was initiated from a dock to the apo state, thus using the ZA1 conformation). Calculated ΔG and ΔH values for both ZA loop conformations are nearly equidistant to the experimental values for 5IGK.

Table 4.2: *ABFE Results for ZA Loop Conformations*

PDB ID	ΔG_{Exp}	ΔG_{ZA1}	ΔG_{ZA2}
3U5L	-8.16	-11.36 ± 0.28	-6.85 ± 0.28
4LZR	-6.40	-6.52 ± 0.53	-2.50 ± 0.21
4QB3	-7.55	-7.55 ± 0.37	-4.94 ± 0.64
4XY9	-7.03	-4.88 ± 0.48	-2.87 ± 0.95
5DW2	-8.00	-7.69 ± 0.69	-8.10 ± 0.32
5IGK	-9.73	-11.79 ± 0.37	-4.84 ± 0.40

ΔG values were obtained via running 3 independent ABFE calculation using different starting structures.

4.3.5. The Dynamics of the ZA Loop

Given the clear role of the ZA-loop conformations in enthalpy prediction accuracy, I analysed the 100 apo-receptor simulations in terms of the transition between ZA1 and ZA2. The transition from ZA1 to ZA2 occurs in all simulations with a mean transition time of 22.59 ± 1.84 ns. Interestingly, the transition was irreversible in most cases, and the reverse transition from ZA2 to ZA1 happened in only two simulations. Moreover, the ZA-loop quickly transitioned back again to ZA2 whenever a reverse transition happened. Heinzelmann et al. reported ZA2 to be more favourable than the ZA1 by −2.54 kcal/mol.[114] **Figure 4.5a** shows an example of the reverse transition from ZA2 to ZA1 around 80-90 ns. A pairwise RMSD analysis also confirmed the reverse transition from ZA2 to ZA1 in the simulation (**Figure S4.6**) and is clearly obvious by simple observation.

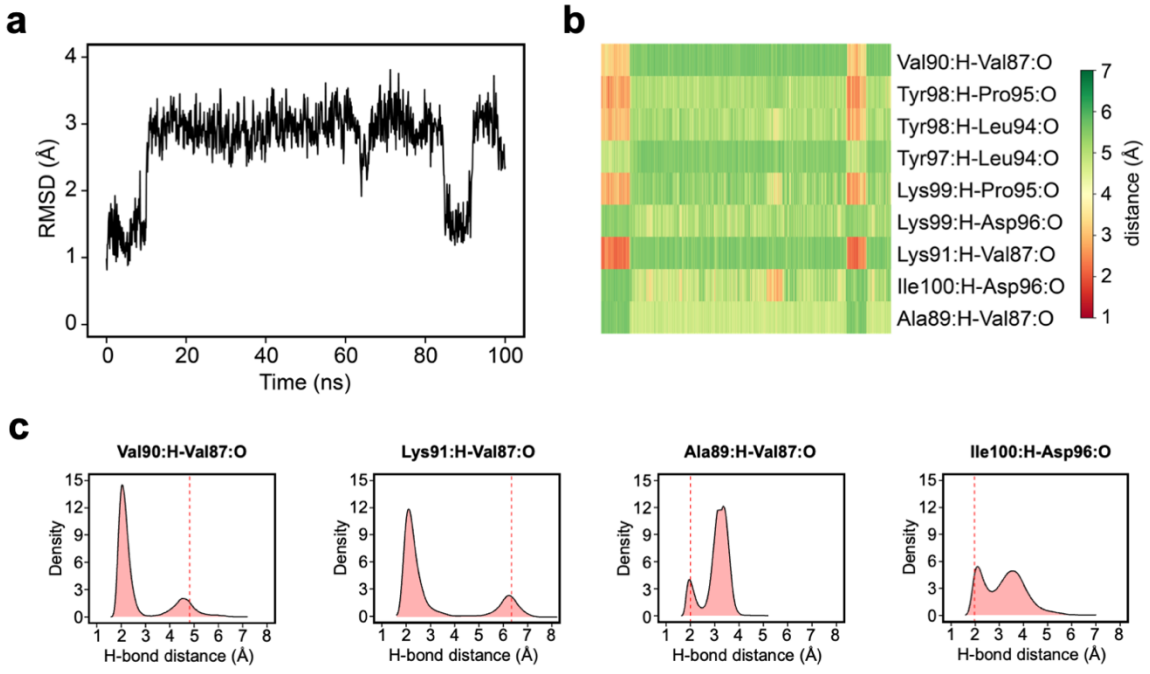

Figure 4.5: ***(a)*** *RMSD plot for ZA-loop (76-106 residues) backbone atoms from a representative case for the transition between ZA1 and ZA2 from one apo-receptor simulation of 100 replicates.* ***(b)*** *Hydrogen bond distance profile of some important backbone hydrogen bonds in the ZA-loop from the representative simulation.* ***(c)*** *Density plots of distance distribution for key hydrogen bonds across 100 replicates of apo-receptor simulations. The red dashed line shows the average distance of starting minimized structures from ZA1 conformations.*

During the transition, the hydrogen bond profiles of key backbone residues change dramatically (**Figure 4.5b, 4.5c**) as well as sidechains. The transition of ZA1 to ZA2 is associated with the backbone torsion angles of ψ (N-CA-C-N) Asp88 and φ (C-

N-CA-C) Asp96 (**Figure 4.6**). These two dihedrals behave like a hinge allowing the ZA-loop to transit from ZA1 and ZA2. The ψ Asp88 shuttles from 50 to -40 degrees while the φ Asp96 moves from -150 to -60 degrees. The distributions of the other backbone torsions in the ZA-loop do not show such clear-cut modal distributions, except ω (CA-C-N-CA) of Gln84, but is possibly not related to the transition from ZA1 and ZA2 (**Figure S4.8**). Potentially, ω GLN84 is related to a recently explained hidden transient state of the ZA-loop, where the event includes breaking two backbone hydrogen bonds between the ZA-loop and the α_A helix.[279]

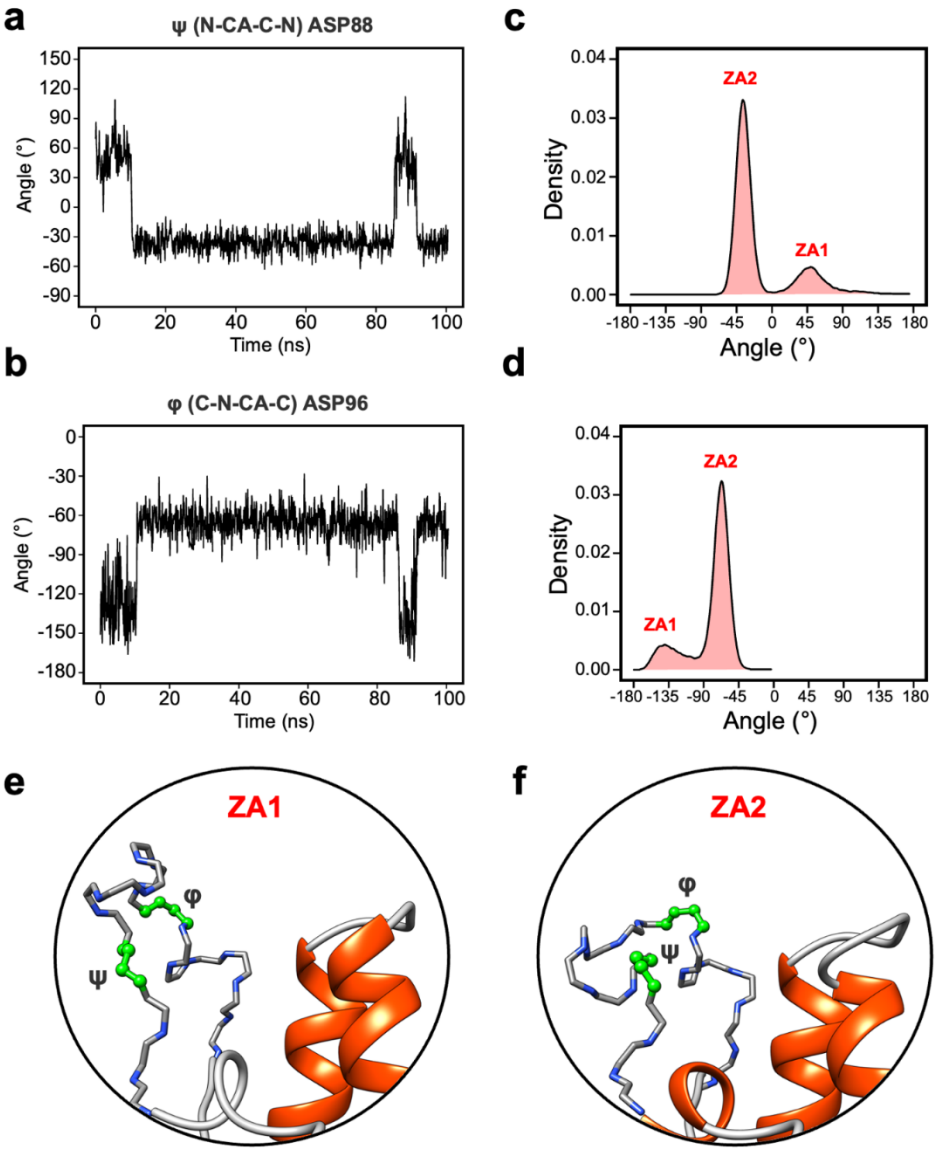

Figure 4.6: (a) *The angle change of the ψ (N-CA-C-N) Asp88 from the simulation in Figure 6a.* ***(b)*** *The angle change of the φ (C-N-CA-C) Asp96 from the simulation in Figure 6A.* ***(c)*** *The angle distribution of the ψ (N-CA-C-N) Asp88 across 100 apo-receptor simulations.* ***(d)*** *The angle distribution of the φ (C-N-CA-C) Asp96 across 100 apo-receptor simulations.* ***(e,f)*** *Main chain representation of ZA1 and ZA2 conformations. Green ball & stick regions show φ and ψ dihedral angles.*

4.3.6. Crystal-Packing of Apo-BRD4-1 Explains ZA-Loop Conformations

Given the ubiquity of the ZA2 conformation in my simulation data but the lack of observation in crystallographic data, I investigated the role of crystal lattice packing. I built a crystal unit cell for the apo BRD4-1 receptor (2OSS) with the $P2_12_12_1$ space group (**Figure 4.7a**). The cell unit contains 4 chains with lengths 37.418, 44.139, 78.413 Å on the *x*, *y* and *z* dimensions. I performed 3 replicates (100 ns) for the crystal unit and 12 replicates (100 ns) for simulations with a single chain (monomer).

I also explored potential force-field influence by examining three force fields: Amber FF14SB, CHARMM36 and OPLS-AA/M. As a result, in total, I performed a total of 9 simulations for a complete unit cell and 36 simulations for a single chain. I was first interested in investigating ZA-loop dynamics in both simulation setups. As expected, all chains retained their crystal-like conformation in the unit cell simulations (**Figure 4.7b**). In contrast, in all single-chain simulations in all three different force-fields, the ZA-loop exhibited much higher flexibility (**Figure 4.7b**), thus supporting the notion that crystal-packing artefacts likely constrain the ZA conformation in the apo state. Crystal-packing effects are likely to be present in the complexes as well – simulations of similar unit-cell simulations of the complexes (**Figure S4.9**) reveal the ZA-loop does not move away from its lattice conformation. Almost all complex PDBs except 5D0C share same space group with the apo BRD4-1 receptor (2OSS) (see **Table 4.1**). Interestingly, simulations of the 5D0C lattice appear to allow more flexibility of the ZA-loop.

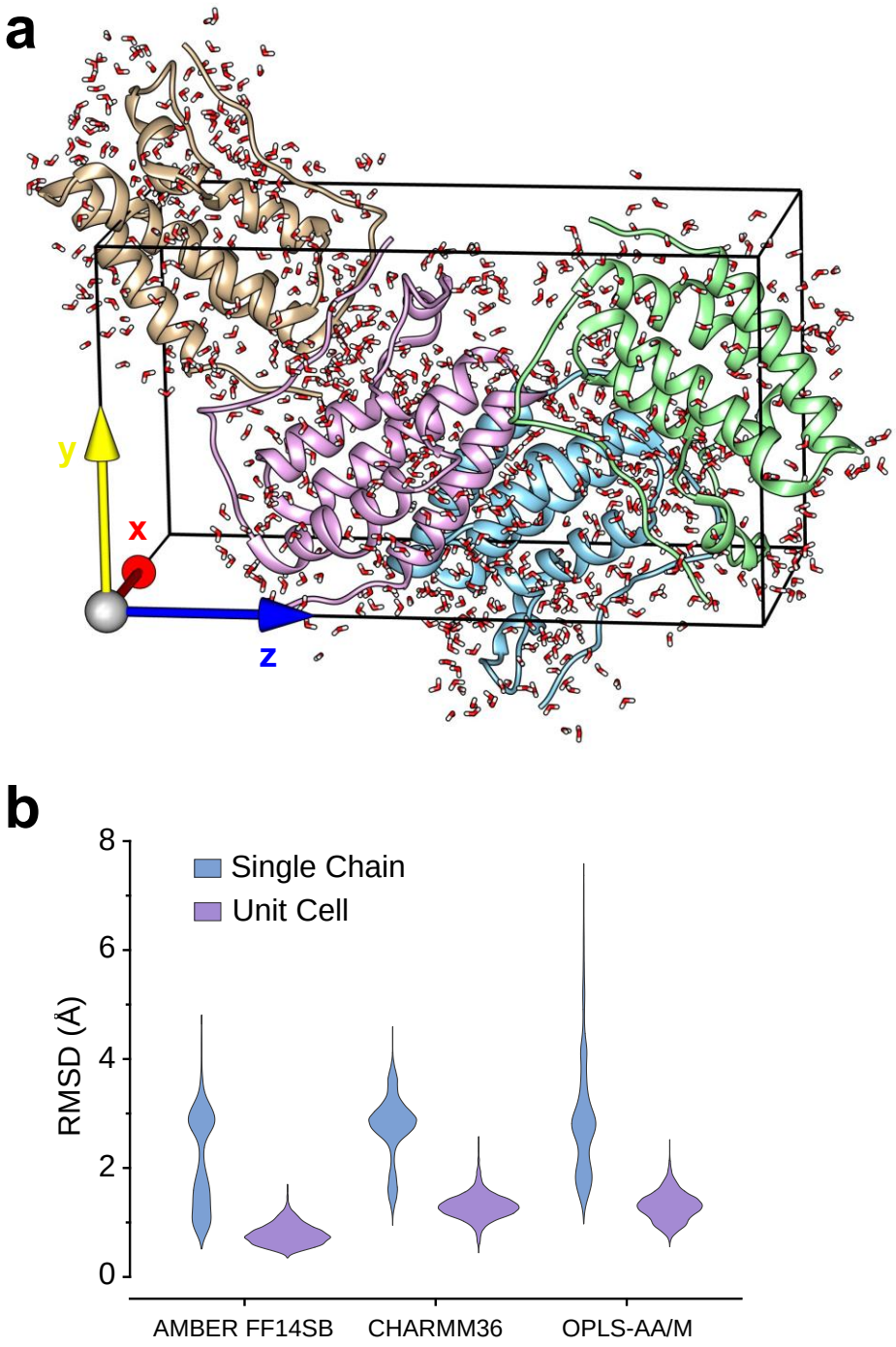

Figure 4.7. (a) *2OSS crystal cell unit containing 4 identical chains.* ***(b)*** *RMSD violin plots for ZA-loop (76-106 residues) backbone atoms. Blue violin shows ZA-loop RMSD from 12 independent single-chain simulations while purple violins show for the unit cell simulations.*

4.4. Discussion

Although much work has been done on model systems[27, 53, 58] the increase in the number of degrees of freedom associated with protein-ligand systems that are the size of typical drug-targets has led to an apparent reticence to explore enthalpy predictions and their potential utility. Given the growth in available computational power, I reasoned it would be useful to investigate the level of accuracy (and precision) one might obtain for a well-characterized protein-ligand system. To that end, I focussed on bromodomains, a system that previously showed that accurate ΔG predictions could be obtained via absolute binding free energies for different compounds[112] at the same protein and for the same compound at different proteins.[113]

I focussed on the BRD4-1 system as it is well characterized both in terms of structure and biophysical characterizations. I first asked how accurate the calculation of enthalpy for this system would be assuming standard procedures reported in the literature, similar to the previous approach for ΔG predictions.[112] The results were surprisingly accurate and precise (**Figure 4.2**) with only three obvious outliers from a 2 kcal/mol error boundary, thus suggesting that it is indeed possible to make reasonably accurate predictions of enthalpy. Outliers provide an opportunity to gain more insight into what factors are important in the prediction of enthalpy. Parameterization is certainly one aspect that can offer improvement as indeed was found for some systems here when charges were optimized. However, in some instances this was clearly not the case, and this led us ultimately to the identification of the influence of the ZA loop on the enthalpy prediction. We should note that although I investigated ligand parametrization for outliers in this work, I did not explore the protein force-field's role. It certainly would be useful to investigate that in future work.

The dynamics and flexibility of the ZA loop has been reported across many different bromodomains, including for example ATAD2[300] , BRD2-2 [301], BRD4-1[302], BRPF1[303], BAZ2B [304, 305], BRD9[306], BRG1[307] and CBP[303]. Thus, it has been proposed that the dynamic nature of the ZA-loop plays an important role in selectivity due to its flexibility and sequence variation between bromodomains.[269, 304, 308] However, experimental 3D structures in the PDB

databank mostly show no significant changes in the secondary structure despite the apparent plasticity. A recent work reported an investigation of 297 crystal structures of BRD4-1 and concluded that there is a high level of similarity in the binding pocket region, regardless of the bound ligand.[309] Nevertheless, there are many studies highlighting flexibility. Eron *et al*., using hydrogen–deuterium exchange mass spectrometry (HDX-MS), reported significant deuterium uptake on the ZA-loop of the apo state, but in contrast, the solvent shielding data pointed out a high degree of stabilization of the ZA-loop with the CFT-1297 degrader ligand.[310] Yu *et al*. reported significant chemical shift around the ZA-loop for BRD4-1 upon ligand binding even though they showed no significant differences between the solution structure of BRD4-1 and its crystal structure in the backbone resonance assignment data.[102] Furthermore, ATAD2 displays "open" or "closed" conformational states of the ZA-loop.[300, 311-313] Further examples from BRG1[307], BRD7[314] and BRD2-2[301], show large structural deviations in the ZA-loop when compared to the other regions according to NMR experiments.

Moreover, computational studies have provided detailed analysis of the dynamic nature of the ZA-loop.[114, 315-318] Tumdam, R., *et al*. showed that the ZA-loop in the apo-BRD4-1 can adopt a similar conformation to that which I observe here.[317] The nature of the transitions of the ZA loop we observe appear very similar to that reported by Heinzelmann *et al*[114] who also computed that the free energy difference was 2.54 kcal/mol (with TIP3P) more favourable for their "open state" (equivalent to our ZA2 state here). Additionally, Cheng *et al*. showed an "in/out" transition of the ZA-loop in BRD4-1 using QM/MM, explaining the differential binding of RVX-208 & 297.[302] More recently, a hidden state across all bromodomain families was proposed via used MD simulations and Markov state modelling in which important backbone hydrogen bonds are broken and the ZA-loop displaces away from the α_A helix.[279]

Seven of the ten complexes here gave excellent predictions of enthalpy of binding. Most of the simulation time for six of those complexes is spent in the ZA1 (close to crystal) conformation. The 5DW2 complex gives an excellent prediction of enthalpy, but rather interesting, readily transitions to the ZA2 loop conformation and indeed using only these conformations gives much more accurate predictions.

Analysis of the 5DW2 simulations with the ZA2 conformations reveals that the ligand makes more close contacts (< 0.4 A) with Asp88 and Tyr139 in ZA2 than ZA1 (**Figure S4.10**).

Of the three initial outliers, simulations of the olinone (30M) complex (4QB3) gave some improvement when charges were refined with QM calculations. However, in these simulations, the complex occupied more time in the ZA1 conformation, and it appears this conformation gives a more accurate estimate of free energy. Interestingly, olinone appears to make interactions with the BC loop in both conformations.

Accurate enthalpy predictions for bromosporine (BMF) in complex with BRD4-1 (5IGK) were only possible when combining predictions from both ZA1 and ZA2 conformations. Whilst simulations of the BzT-7 (08K) complex (3U5L) demonstrate that more accurate predictions can be obtained with just the ZA2 conformation, even though this conformation was never transitioned to in the initial set of 20 simulations. It may be the BzT-7 creates an energy barrier for the loop transition, though this is likely an indirect effect as BzT-7 is one of the smaller ligands and does not interact directly with the ZA loop. Further work would be necessary to explore this in more detail.

Of course, enthalpy is only one component of the binding free energy, ΔG. The results above demonstrate that the interaction and behaviour of a small, but crucial loop, near the binding site plays a key role in shaping that component. The calculations performed here were all retrospective, but such calculations in the future will only be useful if they can be useful prospectively (ie where we do not know/have the ITC measurements). What do these results mean in that context? Firstly, the trends in the initial data are reasonably good. Secondly, it is interesting to note that all the outliers were over-predictions. In one case (4QB3) that could be indirectly ascribed to parameterization, but in the other two cases (3U5L and 5IGK) clearly reflects contributions from conformational states that were not originally sampled, and thus likely reflects a simple sampling problem. If one has knowledge of important (for ligand-binding) dynamics up-front, then strategies can be incorporated to mitigate this. At the very least this would enable one to approach the predicted values with the necessary caution for sensible interpretations. Thus,

in a prospective scenario, if one observes conformational transitions (in for example loops as here) and obtains enthalpy values that differ by several kcal/mol, that should at least suggest that great care should be taken over any future interpretation of the thermodynamics.

4.5. Conclusion

I have demonstrated that absolute ligand-binding enthalpy calculations for a well-characterized drug-target, BRD4-1, can give reasonably accurate results. My results clearly show a strong dependence on the behaviour of the ZA loop to the predicted enthalpy. I have also demonstrated that this alternative loop conformation is likely readily accessible, if not dominant, in the apo state in solution and that crystal lattice packing likely constrains the conformation. Indeed, it may well be the case that the ZA1 conformation of the loop is particularly amenable to lattice formation and thus the reason why many complexes exhibit this conformation. This observation highlights the need to take particular care when using apo state for docking studies and in the subsequent processes of rational drug design, like FEP calculations.

A key question that remains very open at this stage is just how generalizable this approach in is giving accurate predictions of ΔH. Can we expect this approach to become prospective? Prediction of ΔH may be useful in trying to optimize enthalpic contributions during a drug-discovery campaign. However, to do that with confidence in a prospective fashion will depend on more studies showing the approach can deliver across various systems. The work here should be taken as evidence and encouragement that it is feasible, at least for some systems. Studies on additional systems would also allow us to begin to understand how strong entropic contributions (by inference) might be linked to particular moieties or water molecules.

However, an alternative way to use this approach might be to draw researchers' attention to hidden conformational states that may not be immediately apparent from the initial structural biology work. Outliers could be a way to identify such behaviour. The predicted contribution of different states to the enthalpic signature may provide a useful metric to gauge the importance of different states and how valuable they might be in terms of targeting. The ability to make accurate enthalpy predictions alongside accurate ΔG free energy predictions moves us considerably closer to being able to design ligands with desired thermodynamic properties, something that has long been sought after.[319] This will be significantly easier if one has a good understanding of the dynamics of the protein before commencing such studies.

5 Correct Conformational Sampling Ensures Accurate Binding Enthalpies for Protein-Protein Systems

In this chapter, the binding enthalpy of protein-protein binding events is explored. The current methods have not been evaluated in terms of predicting the enthalpy of binding for protein-protein complexes. To address this, a set of eleven protein-peptide complexes with available structural and isothermal titration calorimetry data is examined. The content of this chapter has been submitted in the following journal article:

Çınaroğlu, S. S., & Biggin, P. C. (2023). Computed protein-protein enthalpy signatures as a tool for identifying conformation sampling problems. *Journal of Chemical Information and Modeling*.

5.1. Introduction

Protein-protein interactions play a vital role in many biological processes and underpin cellular signal transduction events within (and beyond) the cell.[320-322] A deeper understanding of the interactome will be necessary to fully appreciate the role of protein-protein interactions in pathology[323, 324] and will aid in identifying novel targets for therapeutic intervention. The need for safer, more effective medicines is more urgent than ever.[325, 326] Protein-protein interactions are a promising and rapidly growing area of drug discovery, with important implications for treating many diseases.[327-332]

Understanding the principles of protein-protein associations requires a comprehensive description of the thermodynamic and kinetic characteristics of the binding.[333, 334] Identifying the driving forces that stabilize the interaction has been actively pursued for decades.[113, 335-338] Protein-protein binding is a complicated process including hydrophobic, van der Waals, and electrostatic interactions. Three-dimensional structures of protein-protein complexes can be obtained using X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. These provide valuable information on critical interactions between binding site residues for an atomic-level understanding of binding mechanisms.[339] Alongside the structural efforts, calorimetric approaches that

provide thermodynamic signatures of binding give complementary insight into the forces that stabilize protein-protein complexes.[340, 341]

Here, I consider the calculation of absolute binding enthalpy from molecular dynamics simulation of a diverse set of protein-protein complexes using the direct method with the multi-box approach. Given the progress in computational power alongside high-quality biophysical data, I reasoned that it would be useful to ascertain what kind of performance could be achieved for protein-protein systems and help make it clear to the community where improvement is likely to be needed. I show that absolute enthalpy can be calculated for diverse protein-protein systems within an error of ~2 kcal/mol. Furthermore, the data presented here could serve as a benchmark for future studies and improving force-field parameters.

5.2. Methods

5.2.1. Building a nonredundant benchmark

First, protein-protein complexes were obtained from the PDBbind v2020.[342, 343] Next, the chain with the longest sequence length was defined as a receptor, while the short chain was defined as the peptide for each PDB entry. PDB entries having receptors with a sequence longer than 150 residues were eliminated to reduce the computational cost in molecular dynamics simulations. The remaining PDB entries (736) were investigated as to whether they have binding enthalpy data reported by ITC (Isothermal titration calorimetry) in the literature.

Then I employed two different clustering approaches to reduce the redundancy in the data. First, I extracted receptor amino acid sequences from RCSB PDB[344], then the amino acid sequences were aligned using Clustal omega[345] and clusters were defined by the result of their phylogenetic relationships of the sequences. Secondly, a sequence-independent clustering based on the local backbone similarity matching were performed using MaxCluster tool for receptors.[346, 347] The final benchmark was selected by choosing the lowest number of total residues of the protein-peptide system from each clusters. I also paid attention to the existence of complete ITC details and the same temperature (25 °C) while choosing the representatives.

5.2.2. Molecular Dynamics Simulation Setup

The PDB structures were obtained from the RCSB PDB database for the benchmark (**Table 5.1**). Missing atoms and loops were modeled with the Modeller implemented in UCSF Chimera[284], and all heteroatoms were removed from the system except all crystallographic waters. I used the Amber ff14SB force field for the protein and the TIP3P water model for water molecules.[226, 288] Zinc AMBER force field (ZAFF) was additionally used for Zn^{2+} atoms parametrization in PHD zinc finger of BAZ2A (PDB: 4Q6F)[348, 349] while default parameters in ff14SB were used for Ca^{2+} atoms in Calmodulin (PDB: 2LQC). All input files are available at doi: 10.5281/zenodo.7635945

All simulations were run using the Gromacs v2020 software package.[30, 142, 233, 291] A 3-step steepest descent energy minimization with a maximum force of 10 kJ/mol/nm^2 was applied to all systems.[350] In the first step, position restraints with an harmonic potential with a force constant of 1000 kJ/(mol/nm^2) were applied for all heavy atoms, then

there remove for solute heavy atoms, and the final step removed all restraints. NVT and NPT ensemble simulations for 1 ns were performed to equilibrate all systems with position restraints with the harmonic potential at a force constant of 1000 kJ/(mol/nm^2) on heavy protein and ligand atoms. Additionally, another NPT ensemble simulation for 1 ns was performed without restraints before the production run for data collection. The V-rescale[235] and Parrinello-Rahman[236] algorithms equilibrated the temperature at 300 K and the pressure at 1.0 bar, respectively. Unbonded interactions were calculated up to a cut-off of 1.0 nm with a potential-shift. A dispersion correction was applied to energy and pressure. All H-bond lengths were constrained with a LINear Constraint Solver (LINCS) algorithm[154, 237]. Coulomb interactions were evaluated with the Fast smooth Particle-Mesh Ewald (SPME) electrostatics method with an initial short-range cutoff of 1.0 nm[148]. The leap-frog algorithm was used to run 20 independent 100 ns MD simulations with 2 fs time step. Snapshots from the production runs were taken every 100 ps for 3D coordinates, while 100 fs for energy data.

5.2.3. Absolute Binding Enthalpy Calculations

Computational calculation of the binding enthalpy requires two sets of molecular dynamic simulations (the complex's bound- and unbound state). The binding enthalpy (ΔH) is calculated by Equation 1, where $\langle \mathrm{E} \rangle_{complex}$, $\langle \mathrm{E} \rangle_{solvent}$, $\langle \mathrm{E} \rangle_{receptor}$, and $\langle \mathrm{E} \rangle_{ligand}$ are the averaged potential energies of the system from four separate simulations. In this method, the number of atoms between the bound and unbound state of the complex should exactly balance. Note that the pressure-volume contribution for the binding enthalpy is negligible.[31] A detailed description is found in section **2.4**.

$$\Delta H = \langle \mathrm{E} \rangle_{complex} + \langle \mathrm{E} \rangle_{water} - \langle \mathrm{E} \rangle_{receptor} - \langle \mathrm{E} \rangle_{ligand} \qquad (5.1)$$

where $\langle \cdots \rangle$ is the time-averaged value obtained from multiple independent MD trajectories. The use of multiple simulations can provide a broader sampling of conformational space than a single, long simulation. Thus, trajectories allow the determination of mean values distributed across the potential energy surface.

Convergence of the potential energy was evaluated by plotting the cumulative average of the simulation data while the uncertainty (standard error of the mean (SEM)) in the energy values obtained from K independent trajectories was estimated using re-blocking

analysis[193], as implemented in the pyblock tool (https://pyblock.readthedocs.io). SEM of ΔH is given by:

$$\sigma_{\Delta H} = \sqrt{\sum_{i=1}^{4} \sigma_i^2} \tag{5.2}$$

where σ_i obtained from the reblocking analysis is the SEM of the overall potential energy of each system (complex, water, receptor, peptide) because ΔH is an additive combination of mean energies. A detailed description about the blocking method is written in section **2.6**.

5.3. Results

5.3.1. Getting a nonredundant benchmark

I extracted 76 PDB entries where the larger protein/peptides in the complex was smaller than 150 residues and where binding enthalpy data was available in the literature (**Figure 5.1a**). The data set contains 25 NMR and 51 X-ray crystal structures with a resolution ranging from 1.11 to 2.6 Å. Receptor proteins come from 49 different genes of 12 organisms, including two virus genomes. Receptor size ranges from 54 to 148 residues, while the peptides range from 4 to 49 residues (**Figure 5.1b**).

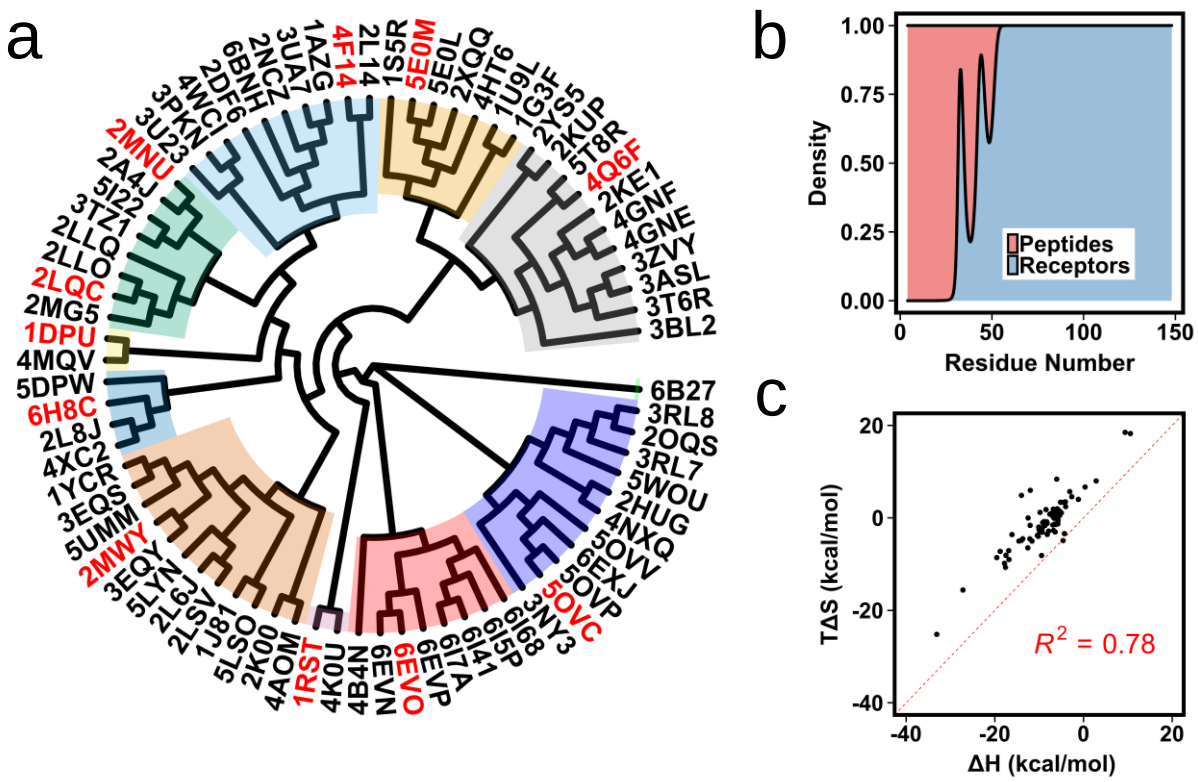

Figure 5.1: a) *Phylogenetic tree of receptor sequences from the seventy-six PDBs. Red PDB IDs are the ones used for the final benchmark.* ***b)*** *Density distribution for sequence lengths of peptides and receptors.* ***c)*** *Entropy–enthalpy scatter plot for the 76 protein-protein complexes.*

After getting PDB entries having binding enthalpies, I performed clustering based on the sequence and structure of the receptors. First, I ran Clustal Omega, a multiple-sequence alignment program, for receptor sequences and obtained the phylogenetic tree of the receptor sequences. The tree-based clustering gave 11 different clusters with cluster sizes ranging from 1 to 11 (**Figure 5.1a**). 6B27 was the only member of cluster 1; however, it

is a tandem repeat of SH3 domains, which mostly exist in cluster 9. After sequence-based clustering, I also employed structure-based clustering using the MaxCluster tool (http://www.sbg.bio.ic.ac.uk/maxcluster/) using the local backbone similarity matching method. As I found some domains (e.g., SH3 and EF-Hand) in different clusters in the sequence-based clustering process, I also performed a structure-based clustering. Structure-based clustering gave us eight different clusters with cluster sizes ranging from 5 to 12 members. I also had an additional cluster for all other PDBs having unrelated structures that are not similar. Finally, I picked representative PDBs with the smallest number of total residues and complete ITC details at the same temperature (25 °C), giving 11 different complexes in the benchmark (**Table 5.1**).

Table 5.1: *A nonredundant benchmark for testing binding enthalpies of the protein-peptide systems.*

PDB	Clusters[a]	Receptor[b]	Peptide[b]	ΔH (kcal/mol)[c]	Ref.
1DPU	7(1)	$P15927_{(201–272)}$	$P13051_{(66–91)}$	−16.80 ± 0.28	[351]
1RST	4(1)	$P22629_{(39–163)}$	AWRHPQFGG	−12.56 ± 0.09	[352]
2LQC	8(5)	$P0DP23_{(1–77)}$	$Q13936_{(45–68)}$	−6.91 ± 0.07	[353]
2MNU	8(1)	$P02751_{(5–83)}$	$APT_{(201–226)}$	−4.60 ± 0.10	[354]
2MWY	5(6)	$O15151_{(23–111)}$	$P04637_{(15–29)}$	−17.50 ± 0.30	[355]
4F14	9(3)	$O76041_{(955–1014)}$	$A4UGR9_{(2245–2258)}$	–8.46 ± 0.48	[356]
4Q6F	11(4)	$Q9UIF9_{(1673–1728)}$	$P68431_{(1–9)}$	−9.81 ± 0.04	[357]
5E0M	10(8)	$Q24117_{(410–478)}$	$Q9H4H8_{(434–446)}$	–8.20 ± 0.10	[358]
5OVC	2(2)	$Q9JLU4_{(570–664)}$	$P97836_{(986–992)}$	−6.80 ± 0.04	[359]
6EVO	3(9)	$O15460_{(142–236)}$	PPGPRGPPG	−8.70 ± 0.80	[360]
6H8C	6(7)	$P60520_{(3-117)}$	$Q9GZZ9_{(337-350)}$	−5.91± 0.09	[361]

[a] *Numbers show sequence-based clusters while the numbers in the parenthesis show structure-based clusters.*
[b] *UniProt IDs for receptor and peptides, some peptides are shown as their amino acid residues.*
[c] *Experimental binding enthalpy values.*

5.3.2. Overall Results for Absolute Binding Enthalpy Calculations

To my knowledge, there has been no previous attempt to compute binding enthalpies for protein-protein systems using the direct method with molecular dynamics simulations. Thus, in the first instance, I wanted to evaluate just how well a standard approach would perform. Here, the absolute binding enthalpy for eleven protein-peptide complexes was estimated by the direct method (eq 2.44) using ensemble averaging (eq 2.45). In getting binding enthalpies, sufficient sampling to cover all conformational space is crucial to obtaining a converged pattern of potential energy. In previous studies, Roy *et al.* performed multiple independent simulations[34], while Li and Gilson ran long simulations to get sufficient sampling for the relative binding enthalpy calculation of a protein-ligand system[35]. Here, I preferred to perform multiple independent simulations to tackle the convergence issue since I obtained good results for bromodomain-ligand complexes

(Chapter 4). Thus, sixty trajectories, totalling six µs of simulation data, were generated from molecular dynamics simulations of the bound and unbound states for each protein-peptide complex. I produced one million snapshots per simulation for the potential energy, which means sixty million data points used for each *ΔH* calculation. Before performing enthalpy calculations, I first checked the overall convergence pattern for each complex. The cumulative convergence in increasing blocks of all simulation data shows a reasonably well-converged pattern of the calculated *ΔH,* as exemplified here for 4F14 (**Figure 5.2a**). This pattern was almost identical to most protein-peptide complexes (**Figure S5.1**). However, 2LQC and 6H8C exhibited a rather uneven profile, indicating that these complexes require more simulation data for convergence. Nevertheless, 2LQC gave a highly accurate result, but 6H8C was the worst (**Figure 5.2b**). Overall, the correlation with the experiment was poor, with an R^2 = 0.17, and the accuracy of the calculations was weak with an average of *RMSE* = 5.74 kcal/mol. The results for only 5 PDBs complexes are within a 2 kcal/mol error limit.

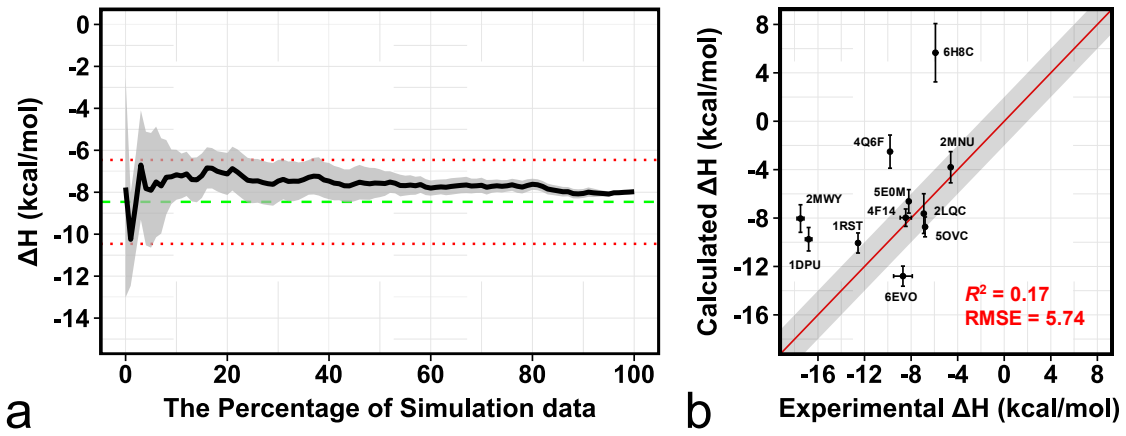

Figure 5.2: a) *Convergence pattern of the calculated ΔH by using 60 trajectories for protein-peptide complex in PDB:**4F14** giving good agreement with the experimental ΔH value. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.* ***b)*** *Comparison of calculated binding enthalpies from experimental values. Error bars were drawn by using the highest SEM values from reblocking analysis. The equivalence line is shown in red, and the black shadow indicates the 2 kcal/mol error limit.*

5.3.3. Tail Conformations Affect the Accuracy of Predictions

Although the initial results were quite poor in correlation with the experiment, about half of the predictions were close to the experimental value, which encouraged me to identify possible causes of error in the inaccurate results. Firstly, 1DPU having a 32 kDa subunit of Replication protein A (RPA32) as the receptor gave an inaccurate *ΔH* with a 7.12

kcal/mol difference between calculated and experimental values (**Figure 5.3a**, blue line). RPA32 interacts with DNA damage response proteins, including SMARCAL1, Tipin, UNG2 and XPA[351]. In the 1DPU structure, $RPA32_{(201–272)}$ is in complex with a 73-88 residue-long peptide from UNG2. To identify potential problems with the ΔH calculation, I first performed clustering on the simulations of the complex using the single-linkage method with a 1 Å cutoff. I extracted the most populated conformations from each simulation. After superposition, the conformations were inspected, and I observed highly flexible termini. Colouring conformations by the difference between calculated and experimental ΔH, revealed two distinct conformations for the C-terminal, highly correlative with the experimental ΔH (**Figure 5.3b**). In one conformation (which I term **1DPU$_{tail2}$**) good agreement is found with the experiment, and here, the terminal residue E270 interacts with K265 and K78 of UNG2. On the other hand, in the other conformation (**1DPU$_{tail1}$**) (**Figure 5.3e-f**), E270 interacts with K217 and R88 of UNG2. This apparent correlation between conformations and binding enthalpies led me to perform additional simulations to increase the sampling and check the stability of the interaction between the terminal residue and K265. Then I performed further simulations with 20 replicates using the **1DPU$_{tail2}$** conformation as a starting structure. I analysed the distance between the C-terminal C atom and the side-chain N atom of residue K265 to check the stability of the interaction (**Figure 5.3d**). The second set of simulations revealed that the **1DPU$_{tail2}$** conformation was highly stable across 20 independent simulations. The most striking result from the binding enthalpy calculation is that the second set of simulations provided a noticeably more accurate ΔH (**Figure 5.3a,** orange line) than the first simulation set. This result indicated that the second simulation set produced correct conformational samples for the complex simulations.

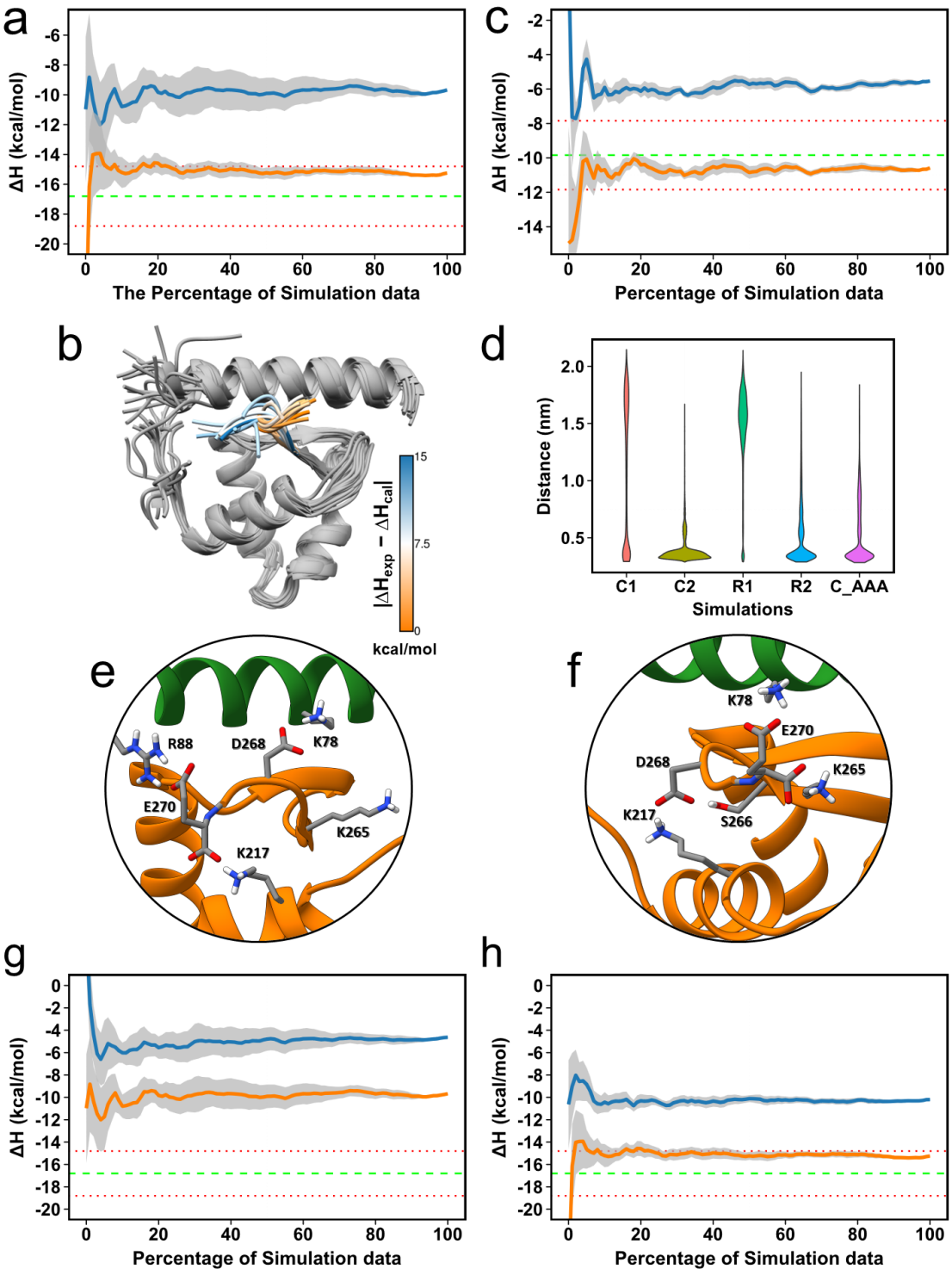
a
ΔH (kcal/mol)
The Percentage of Simulation data
c
ΔH (kcal/mol)
Percentage of Simulation data
b
|ΔH_exp − ΔH_cal|
kcal/mol
d
Distance (nm)
C1
C2
R1
R2
C_AAA
Simulations
e
R88
D268
K78
E270
K265
K217
f
K78
E270
D268
K265
K217
S266
g
ΔH (kcal/mol)
Percentage of Simulation data
h
ΔH (kcal/mol)
Percentage of Simulation data

Figure 5.3: a) *Convergence pattern of the calculated ΔH by using 60 trajectories for two simulation setups for RPA32-UNG2 complex. The blue line shows calculated ΔH for the first simulation set while the orange line shows for the second simulation set. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.* ***b)*** *The most populated conformations from each simulation for 1DPU complex. Blue-Orange color scale represents the absolute difference between calculated and experimental ΔH.* ***c)*** *Convergence pattern of the calculated ΔH by using C_AAA with R1 (orange) and R2 (blue) simulations. Peptide has a mutation (RNK/AAA) in C_AAA simulations.* ***d)*** *The violin plots for the distance between C-terminal C atom and side-chain N atom of the residue K265. C1 and R1 are initial simulations for complex and apo-receptor, respectively. C2 and R2 are additional simulations having more sampling of the second tail conformation in b.* ***e-f)*** *Residues around the tail and their interactions in two conformations extracted from simulations.* ***g)*** *Convergence pattern of the calculated ΔH by using C1 with R1 (orange) and R2 (blue) simulations.* ***h)*** *Convergence pattern of the calculated ΔH by using C2 with R1 (orange) and R2 (blue) simulations.*

To further corroborate this finding, I considered a modified system, UNG2$^{RNK/AAA}$, where the peptide has three alanine mutations. ITC results have also been reported for this system[351]. Importantly, this complex contains the K78A mutation, thus disrupting the interaction between E270 and K78 observed above. Nonetheless, the tail was quite stable in the **1DPU$_{tail2}$** conformation, as shown in **Figure 5.3d,** despite the mutation. Surprisingly, the UNG2$^{RNK/AAA}$ peptide also provided an accurate ΔH with the **1DPU$_{tail2}$** conformation (**Figure 5.3c**). In these calculations, the apo-receptor simulations mostly had **1DPU$_{tail1}$** conformation and provided good results when used with complex simulations with the **1DPU$_{tail2}$** conformation (**Figure 5.3g-h**). I explored this relationship further and performed additional simulations using only **1DPU$_{tail2}$** conformations for the apo-receptor to examine the effects on the ΔH prediction. The result was always worse when the receptor comprised the **1DPU$_{tail2}$** conformation in the apo-receptor simulations (**Figure 5.3g-h**). These findings suggest that complex and apo-receptor should have different tail conformations for accurate ΔH predictions.

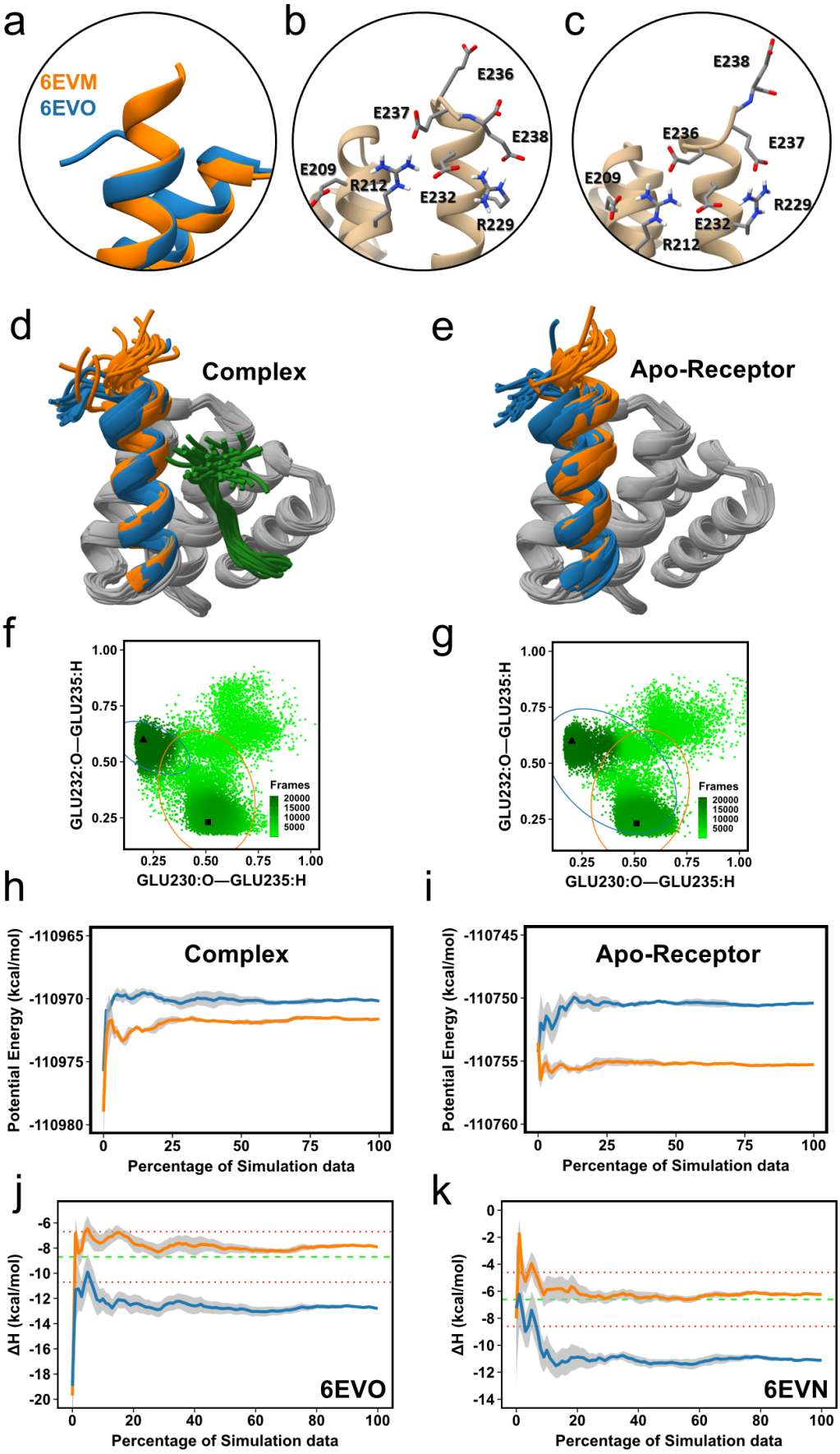
a
6EVM
6EVO
b
E236
E237
E238
E209
R212
E232
R229
c
E238
E236
E237
E209
R229
E232
R212
d
Complex
e
Apo-Receptor
f
GLU232:O—GLU235:H
GLU230:O—GLU235:H
Frames
20000
15000
10000
5000
g
GLU232:O—GLU235:H
GLU230:O—GLU235:H
Frames
20000
15000
10000
5000
h
Complex
Potential Energy (kcal/mol)
Percentage of Simulation data
i
Apo-Receptor
Potential Energy (kcal/mol)
Percentage of Simulation data
j
ΔH (kcal/mol)
6EVO
Percentage of Simulation data
k
ΔH (kcal/mol)
6EVN
Percentage of Simulation data

Figure 5.4: a) *The tail conformations in crystal structures.* ***b-c)*** *Residues around the tail and their interactions in two conformations extracted from simulations.* ***d-e)*** *The most populated conformations from each simulation set.* ***f-g)*** *Distance (nm) distribution of two hydrogen bonds from apo-receptor and complex simulations with two different tail setups. Elliptic circles show 75% probability distribution for tail1 (blue) and tail2 (orange) simulations. Triangle shows hydrogen bond distances in 6EVO crystal structure while square shows distances in 6EVM.* ***h-i)*** *Convergence plots of the potential energy for complex and apo-receptor simulations having two different tail conformations.* ***j-k)*** *Convergence pattern of the calculated ΔH for 6EVO and 6EVN. The blue line shows calculated ΔH for the first simulation set (tail1) while the orange line shows for the second simulation set (tail2). Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.*

The 6EVO complex also exists in two distinct tail conformations at the C-terminus. 6EVO contains the peptide-substrate-binding (PSB) domain of human type II collagen prolyl-4-hydroxylase (C-P4H-II) complexed with a proline-rich procollagen peptide (PPGPRGPPG)[360]. I found two different tail conformations at the C-terminal in both complex and apo-receptor simulations and also in two different crystal structures: 6EVO (**tail1**) and 6EVM (**tail2**) (**Figure 5.4a**). Like 1DPU (**Figure 5.3e-f**), the tail contains negatively charged residues, interacting with positively charged residues around the tail (**Figure 5.4b-c**). In the original paper for 6EVO, there were five ITC data reports with four crystal structures, but I picked 6EVO for our benchmark since the other peptides had relatively weak binding affinities[360]. Thus, I used 6EVO in our simulations for the binding enthalpy. However, in the initial calculations, the difference between calculated and experimental enthalpy was 4.10 kcal/mol (**Figure 5.4j**, blue line). In the initial simulations, the tail mostly stayed in the **tail1** conformation. However, it was unclear how the **tail2** conformation would affect enthalpy predictions. Thus, I run an additional 20 independent simulations using the **tail2** conformation for both complex and apo-receptor. Interestingly, the tail mostly remained stable with whatever conformation was used as the starting conformation (**Figures 5.4d-e**). I used two important backbone hydrogen bonds as a proxy for the tail's conformational state. E235:H makes a hydrogen bond with E230:O in 6EVO, while it makes a different hydrogen bond with E232:O in 6EVM. This analysis revealed that the **tail1** conformation was more stable in the complex form than the apo-receptor (**Figures 5.4f-g**). Then I checked the potential energies for each simulation setup. Complex simulations for protein-peptide did not give a significant difference with 1.41 kcal/mol for both simulation setups; on the other hand, apo-receptor simulations had very different potential energy profiles with 4.88 kcal/mol difference between **tail1** and **tail2** setups (**Figure 5.4h-i**). As a result, simulations of apo-receptor with **tail2** conformations improved the binding enthalpy prediction to a difference of 0.83 kcal/mol between the calculated and experiment (**Figure 5.4j**, orange line). To confirm this finding further, I

performed an additional ΔH calculation for 6EVN having the same receptor complex with PPGPAGPPG peptide. This peptide has R5A mutation and −6.6 ± 0.1 kcal/mol binding enthalpy as determined by ITC. Apo-receptor simulations with the **tail2** conformation provided more accurate ΔH than simulations mostly having **tail1** conformation (**Figure 5.4k**). Thus, overall, the results show that the receptor should have the **tail2** conformation in the apo state but the **tail1** conformation in the bound state to obtain an accurate binding enthalpy calculation for the PSB domain of human C-P4H-II.

5.3.4. Accurate Binding Enthalpy Behind Unknown Helix Formation

In two other outliers, I observed that a small α-helix tended to formed at the N-termini of the peptides. Firstly, 6H8C (**Figure 5.5**a) is an NMR structure of the human GABARAPL2 protein in complex with the LIR motif that is found within UBA5 (ubiquitin-like modifier activating enzyme 5)[361]. Starting from the PDB coordinates the complex gave us the worst initial ΔH prediction with the highest absolute difference of 11.78 kcal/mol between calculated and experimental values (**Figure 5.5b**). However, using simulation data where the helix was mostly present in the complex resulted in a greatly improved enthalpy prediction (**Figure 5.5c**). This finding prompted me to conduct further simulations to increase the sampling of the helical conformation and test the stability of the helix. Thus, I ran additional twenty independent simulations using the complex with the helix present from the start. The secondary structure was computed for each simulation, and the analysis revealed that the helix was relatively stable across almost all simulations (**Figure 5.5a**). Consequently, the second simulation set with the helix present dramatically improved the ΔH prediction (**Figure 5.5b**). However, I should be cautious in the interpretation here because the peptide used for the ITC measurements excluded the first 4 residues used in the NMR experiments (and also included 3 extra residues at the C-terminus) [361] (**Figure 5.5d**). Thus, it is difficult to make true direct comparison here.

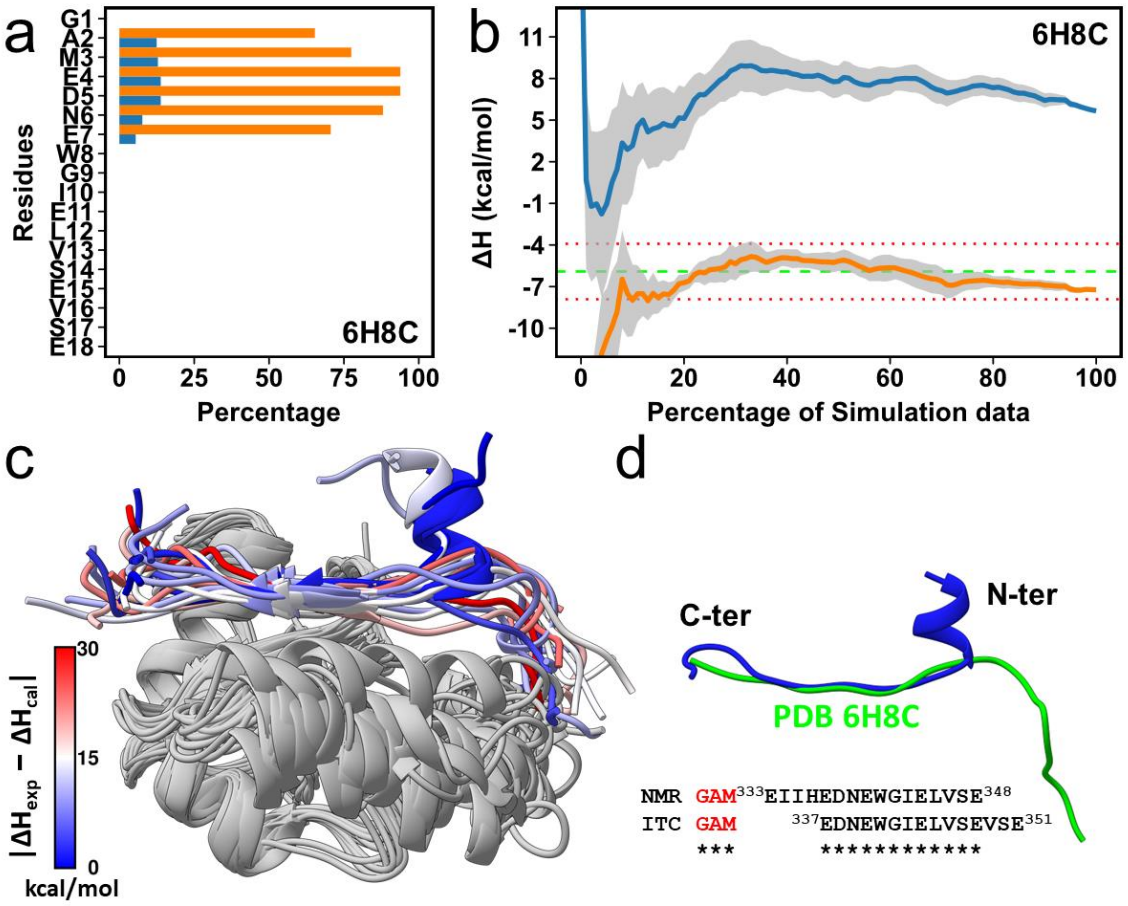

Figure 5.5: a) *The α-helix percentage per residue of the peptide.* ***b)*** *Convergence pattern of the calculated ΔH. The blue line shows calculated ΔH for the first simulation set while the orange line shows for the second simulation set with the helix formation. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.* ***c)*** *The most populated conformations from each simulation for complex simulations. Blue-Red color scale represents the absolute difference between calculated and experimental ΔH.* ***d)*** *Superimposition of the peptide from the NMR structure and the simulation. Sequence alignment of the peptide used in NMR and ITC experiment. The first three residues (GAM) are due to a cloning artefact.[361]*

The other complex where helix formation was observed was the MDMX-p53 complex from 2MWY. In this complex, the N-terminal domain of $MDMX_{(23–111)}$ binds the first transactivation domain (TAD1) of $p53_{(15-29)}$[362]. When I looked at the experimental structures (**Figure 5.6a**), the TAD1 peptide has a small helical structure, but an increase in the amount of helix of the TAD1 region was noticeable in our simulations. Interestingly, this same helix formation at the same region is also found in the AlphaFold model (https://alphafold.ebi.ac.uk/entry/P04637). This helix increase was clearly visible in more than 50% of the simulations (**Figure 5.6b**). I identified ten simulations with the highest helix formation and examined their contribution to the binding enthalpy. While the simulations with the highest amount of helix improved the enthalpy result (-9.68 kcal/mol),

the simulations with the least amount made the result even worse (-6.40 kcal/mol). Overall, the ΔH for 2MWY was estimated at -8.04 ± 0.08 kcal/mol for all simulations (**Figures 5.2b, 5.6c**). This suggested to me that, once again, the presence of additional helix in the peptide would improve the enthalpy prediction. Thus, I ran an additional twenty independent simulations for MDMX-p53 with the additional helical content. The secondary structure analysis revealed that the helix was highly stable across the simulations (**Figure 5.6b**), and the second simulation set with the helix formation significantly improved the ΔH prediction (**Figure 5.6c**).

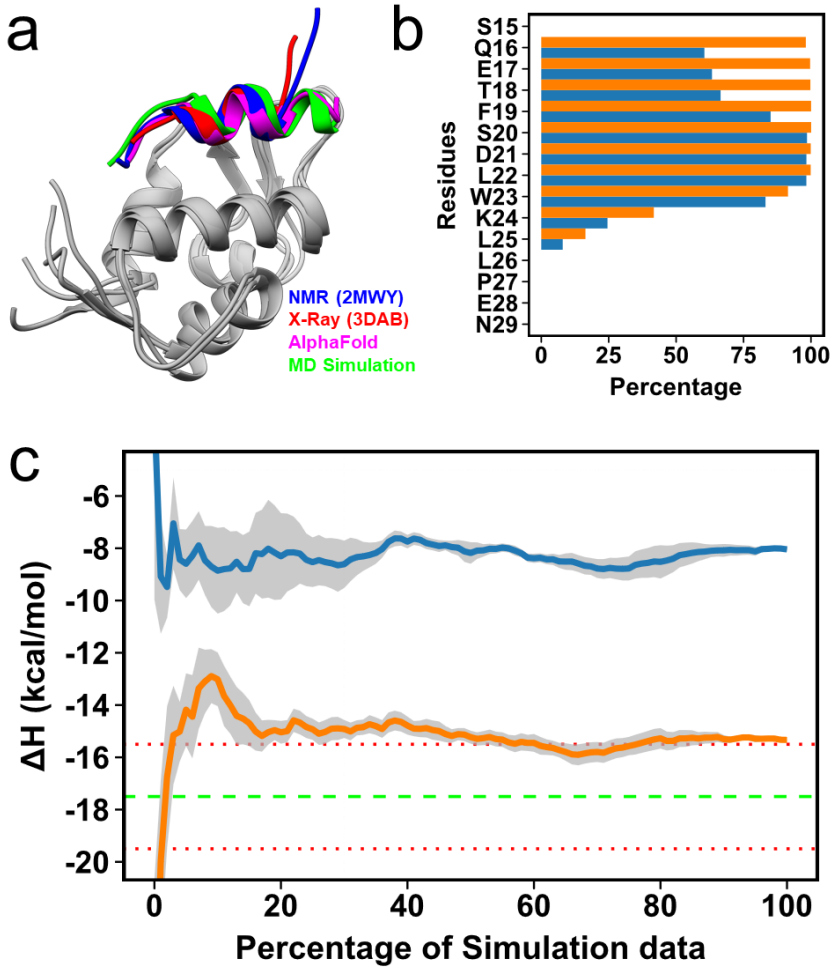

Figure 5.6: a) *Superimposition of three different MDMX-P53 complexes. Blue is NMR structure (PDB: 2MWY) which we used in the study. Red is X-Ray structure (PDB: 3DAB). Green was observed in the simulations. Magenta: Alphafold model obtained from https://alphafold.ebi.ac.uk/entry/P04637* ***b)*** *The α-helix percentage per residue of the peptide.* ***c)*** *Convergence pattern of the calculated ΔH. The blue line shows calculated ΔH for the first simulation set while the orange line shows for the second simulation set with the helix formation. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.*

5.3.5. Using Proper Parameters for Metal Cations in Metalloproteins

In this benchmark, I had two metalloproteins; Calmodulin having two Ca^{2+} ions (PDB: 2LQC) and the PHD zinc finger domain of BAZ2A that has two Zn^{2+} ions (PDB: 4Q6F) (**Figure 5.7**). Calmodulin makes a complex with the "NSCaTE" peptide from the N-terminal cytoplasmic domain of the L-type voltage-gated calcium channel α1C subunit[353]. This complex gave us an accurate ΔH with the absolute difference of 0.72 kcal/mol between calculated and experimental values (**Figure 5.7c and Table S5.1**) when I used the default parameters for Ca^{2+} within AMBER ff14SB. On the other hand, the BAZ2A PHD zinc finger in complex with an unmodified H3K4 histone peptide in the 4Q6F structure provided an incorrect ΔH (**Table S5.1**) when using the default parameters for Zn^{2+} in AMBER ff14SB. Therefore, I investigated other parameters for Zn^{2+} and ran an additional forty simulations using the Zinc AMBER Force Field (ZAFF)[348, 349] for the complex and apo-receptor simulations. The new setup improved the *ΔH* prediction with 5.16 kcal/mol improvement for 4Q6F (**Figure 5.7c**) and placed just around the 2 kcal/mol error level.

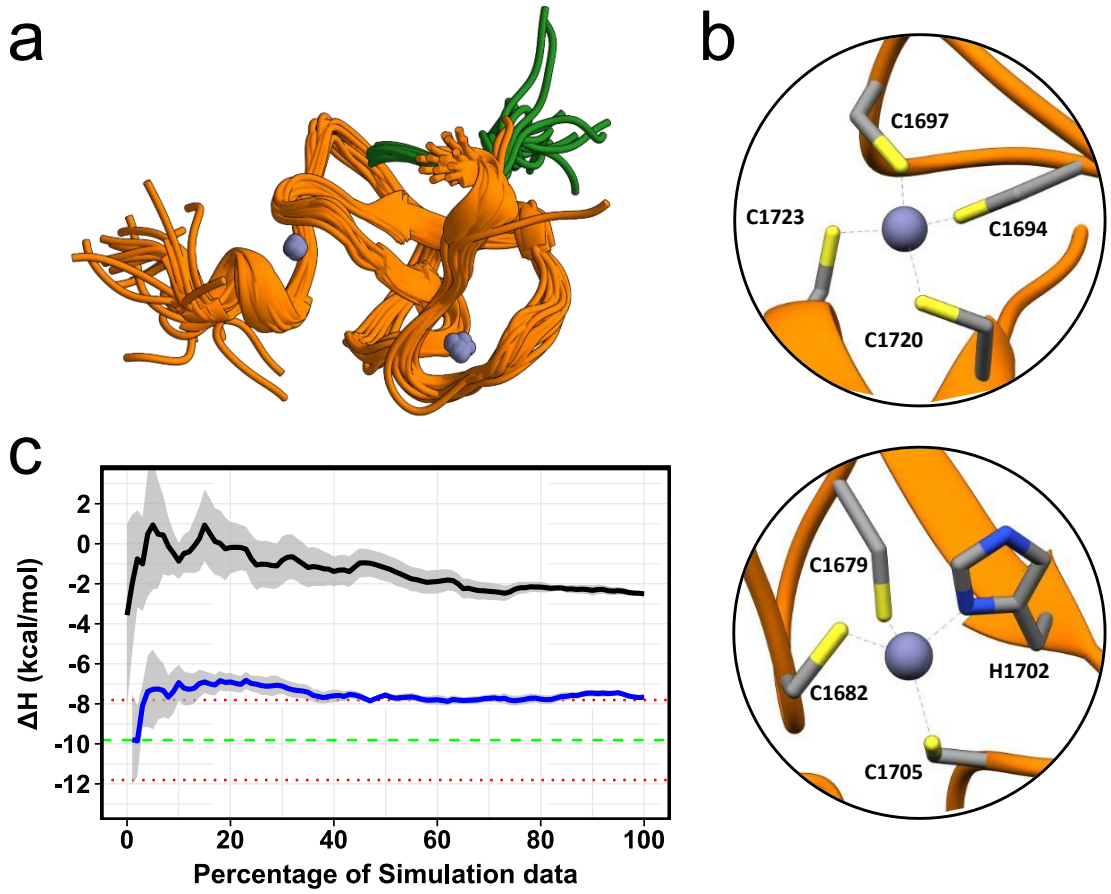

Figure 5.7: a) *The most populated conformations from each simulation for complex simulations.* ***b)*** *Zn metal centers and surrounding residues in the receptor. different MDMX-P53 complexes.* ***c)*** *Convergence pattern of the calculated ΔH by using default Zn^{2+} parameters in Amber FF14SB (black) and Zinc AMBER force field (ZAFF) (blue)*

5.3.6. Considering Experimental Conditions

I have performed all simulations with pure water to reduce complexity since I obtained good results for bromodomain-ligand complexes in pure water (Chapter 4). The ITC experiments were mostly conducted in diverse buffer conditions (**Table S5.2**). Nevertheless, I wanted to see how accurate this approach is given an explicit buffer composition typical of that used in ITC experiments. I picked the PSB domain of human C-P4H-II (6EVN and 6EVO) since it required two different sampling approaches for in complex and apo-receptor simulations. Thus, I setup additional simulations having 20 mM TRIS (tris(hydroxymethyl)aminomethane), 50 mM NaCl, and 50 mM glycine for the receptor-peptide complex and the apo-receptor. I used the **tail1** conformation in 6EVO for complex simulations as a starting conformation while the **tail2** conformation found in 6EVM was used for the apo-receptor simulations (**Figure S5.4a**) since this setup provided the most accurate results (**Figure 5.4c-d**). The same hydrogen bonds were used as a proxy to check the stability of the tail conformation (**Figure 5.8a-b**). The simulations show that we can get an accurate ΔH prediction whether simulating in pure water or a complete buffer condition for the PSB domain of human C-P4H-II (**Figure 5.8b**). Perhaps surprisingly, 6EVO came up with a more accurate ΔH prediction in pure water than in buffer condition (**Figures 5.4c and 5.8b**).

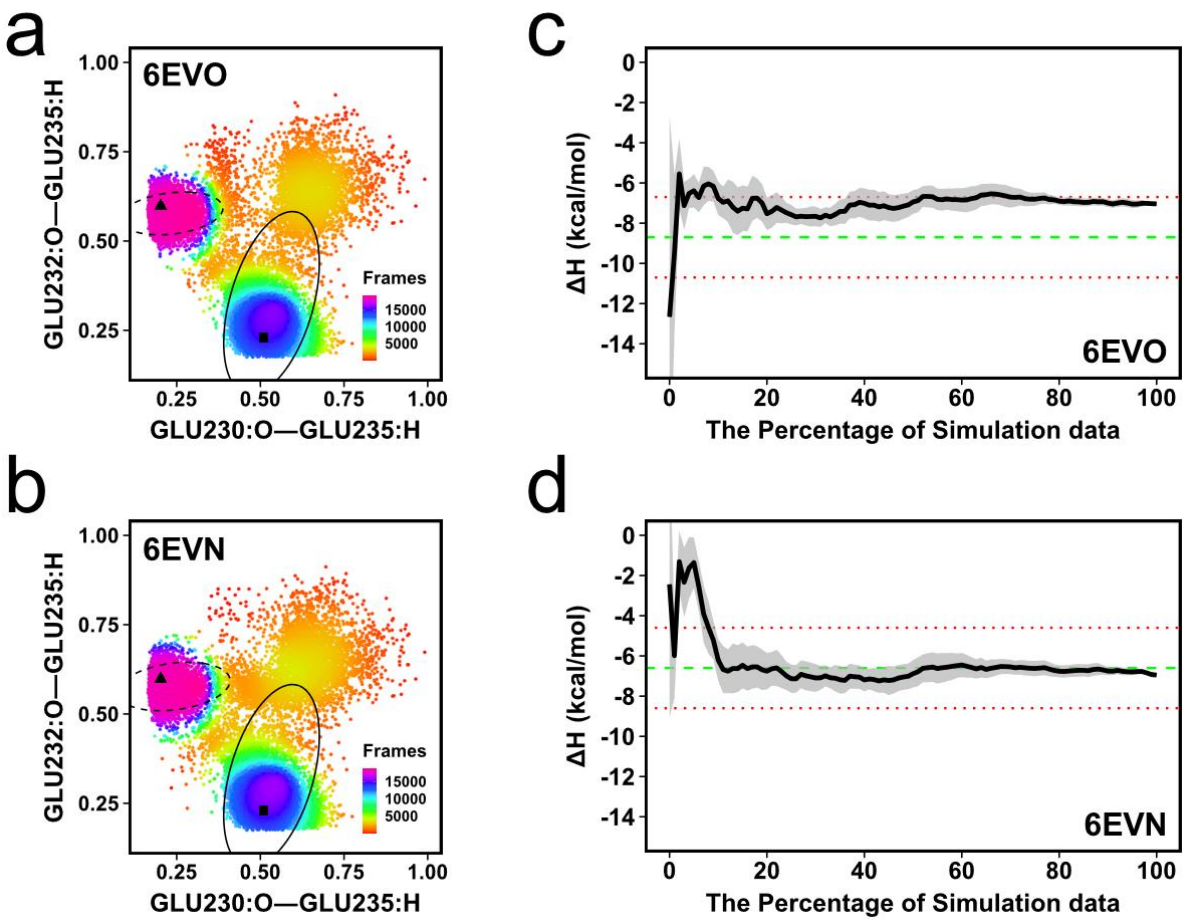

Figure 5.8: a-b) *Distance distribution of two hydrogen bonds from apo-receptor and complex simulations. Elliptic circles show 75% probability distribution for complex (dashed) and apo-receptor simulations. Triangle shows hydrogen bond distances in 6EVO crystal structure while square shows distances in 6EVM.* ***c-d)*** *Convergence pattern of the calculated ΔH for 6EVO and 6EVN in experimental condition. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.*

5.3.7. Decompositions of Binding Enthalpies into Physical Components

Analysis of binding enthalpies using molecular dynamics simulations allows for analyzing subcomponents of the binding enthalpy[31, 53]. I investigated the determinants of the binding enthalpies from simulations giving accurate results. **Table 5.2** includes subcomponents of binding enthalpies for eleven protein-peptide systems into three physical determinants: changes in the van der Waals forces (**LJ**); changes in electrostatic interactions (**Coul**); changes in valence terms, which include bond-stretch, bond-angle, and dihedral terms (**Val**). The Lennard-Jones term was consistently negative for all complexes, making the major contribution to the binding enthalpy. Unlike the Lennard-Jones contribution, the electrostatic contribution was mostly unfavourable, while valence terms varied in sign and magnitude. Interestingly, despite unfavourable contributions,

there were moderate correlations between electrostatic contributions and experimental enthalpy/entropy values with 0.52-58 Pearson's r. Hydrogen bonds (in the force-field model employed here) are primarily electrostatic attractions between molecules. Thus, I checked relationships with hydrogen bonds. Similarly, I found modest correlations for the total number of hydrogen bonds between receptor and peptide with 0.48-52 Pearson's *r* for experimental enthalpy/entropy values (**Table S5.4**). Perhaps surprisingly, there was no relationship between experimental enthalpy/entropy values and absolute hydrogen bond difference (ΔHBond) between bound and unbound states, but ΔHBond highly correlated with **Val** (*r* = 73). Moreover, I analyzed hydrogen bonds between solute and solvent and observed higher correlations with Pearson's *r* of 0.70-75 for the average number of hydrogen bonds per SASA (solvent-accessible surface area) with experimental enthalpy/entropy values, respectively. I also observed moderate correlations for the hydrophobic proportion of the SASA; **Coul** was negatively while **LJ** was positively correlated with the hydrophobic percentage of total SASA in bound and unbound states (**Table S5.4**).

Table 5.2: *The subcomponents of the binding enthalpies.*

PDB	Val	Coul	LJ
1DPU	-1.80 ± 0.01	3.83 ± 0.14	-17.27 ± 0.09
1RST	0.01 ± 0.01	-1.45 ± 0.13	-8.62 ± 0.08
2LQC	2.92 ± 0.01	-0.66 ± 0.14	-9.89 ± 0.09
2MNU	-2.13 ± 0.01	7.33 ± 0.16	-9.00 ± 0.10
2MWY	3.95 ± 0.01	-2.39 ± 0.11	-16.89 ± 0.07
4F14	-0.49 ± 0.01	3.69 ± 0.14	-11.18 ± 0.09
4Q6F	-5.86 ± 0.01	12.24 ± 0.12	-14.05 ± 0.07
5E0M	-3.18 ± 0.01	10.75 ± 0.14	-14.20 ± 0.09
5OVC	-1.16 ± 0.01	8.90 ± 0.14	-16.46 ± 0.08
6EVO	2.34 ± 0.01	3.59 ± 0.12	-13.90 ± 0.07
6H8C	3.09 ± 0.01	8.83 ± 0.14	-19.15 ± 0.09

Coul*: Coulombic electrostatic contribution.* ***Val****: contribution from changes in bond-stretch, angle-bend, and dihedral terms.* ***LJ****: Lennard-Jones contribution. All values are in kcal/mol.*

5.4. Discussion

The combination of molecular dynamics simulations, protein's 3D structures, and experimental calorimetric data can provide a greater understanding of the driving forces for protein-protein binding[112, 113, 363, 364]. The present chapter shows that atomistic simulations can be used to estimate binding enthalpies for protein-protein systems in aqueous solution with high accuracy (**Figure 5.9**) and can be used to investigate calorimetric characterizations of the binding. This will be very helpful for interpreting calorimetric calculations with atomistic details because calorimetric data often apparently contradict accepted scientific theories[26, 33, 35, 365]. For example, it was often thought that the enthalpic contribution of the binding is dominated by polar interactions in protein-ligand bindings[26, 28]; however, the calculations here emphasize the importance of the van der Waals interactions for protein-protein systems (**Table 5.2**). Moreover, a gain in the number of hydrogen bonds upon the binding was usually thought to be related to more favourable enthalpies. Still, I found no such correlation between the overall change in the number of hydrogen bonds and the binding enthalpy, at least for these systems.

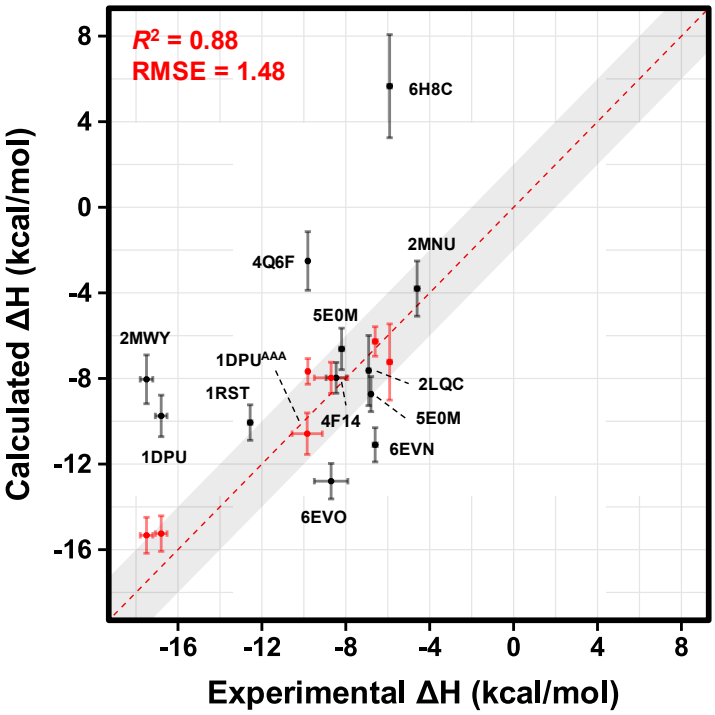

Figure 5.9: *Comparison of calculated binding enthalpies from experimental values for 13 different protein-peptide systems (the 11 benchmarks plus 2 mutations). Red ones show the result after an artificially guided conformational sampling of the outliars.* R^2 *and RMSD were given by using the best enthalpy values. The line of equivalence is shown in red, and the grey shadow indicates the 2 kcal/mol error limit. Error bars were drawn by using the highest SEM values from reblocking analysis.*

Any calculations related to the free energy landscape of the binding between proteins will most likely be computationally very expensive since proteins can adopt many alternative conformations[366, 367]. Thus, insufficient conformational sampling is expected to hamper the accurate calculation of calorimetric components. Moreover, the free energy is a combination of many terms, which also increases the risks of errors and, thus, lowers precision and accuracy. Yet, a current meta-analysis of 853 studies from 34 different research groups showed that free energy methods were close to chemical accuracy (error = 1.58 kcal/mol).[368] On the other hand, binding enthalpy/entropy calculations have generally not been given much consideration with molecular dynamics simulations, and there are limited studies, especially for complex systems[24, 31, 34, 35, 369]. It has been shown that an accurate binding enthalpy calculation is possible with a simplistic model, host-guest systems[2, 53, 57, 58]. Notable efforts have also been made for complex systems to obtain relative binding enthalpies[34, 35]. All of these studies have demonstrated that binding enthalpy calculations are sensitive to the choice of force field and sufficient sampling.

In Chapter 3, I reported a comprehensive assessment of force fields and water models using a set of 25 cucurbit[7]uril–guest (CB7) pairs and shown the TIP3P water model combination with General Amber force field (GAFF) provided good agreement with the experimental values[58]. From the sampling perspective, it is unclear how much simulation data is needed to get accurate binding enthalpies of complex systems. Roy *et al*. ran forty independent ten ns simulations to get sufficient sampling with 1 *ps* writing frequency for energy data[34]. Furthermore, Li and Gilson reached over 250 μs worth of simulation time by seeding every 200 ns and recorded energies every 2 *ps[35]*. So far, these were the only examples from the literature using the direct method for complex systems. Here, I performed 20 completely independent repeats of 100 ns simulations and recorded energy data every 100 *fs* for each system. This approach provided highly accurate and well-converged binding enthalpies for relatively small protein-protein systems (**Figure 5.9**), but inaccurate enthalpies mostly came from incorrect global sampling (**Figure 5.2b**) exemplified by conformation states not seen experimentally.

As reported here, it is well known that receptors can adopt alternative conformations upon peptide binding[282, 374-378]. Here, I observed the C-terminal loop region having two major conformations in the RPA32 (**Figure 5.3b**) and the PSB domain of human C-P4H-II proteins (**Figure 5.4a**). In the NMR structure (1DPU), RPA32 exhibits conformational heterogeneity in the C-terminal, but in the simulations, two distinct conformations were populated (**Figure 5.3d**), and RPA32 provided accurate enthalpies for $UNG2^{WT}$ and $UNG2^{RNK/AAA}$ peptides only when it had two different conformations in bound and unbound states. Likewise, the PSB domain of human C-P4H-II also provided accurate enthalpies for two different peptides (6EVO and 6EVN) in pure water and also experimental buffer condition (20 mM TRIS, 50 mM NaCl and 50 mM glycine) when having two different conformations in bound and unbound states. Conversely, the PSB domain has these conformations in X-ray crystal structures (**Figure 5.4a**): **tail1** conformation in 6EVN, 6EVO and 6EVP, **tail2** conformation in 6EVL (apo) and 6EVM, which contains a weak binder $(P)_9$ peptide. These suggest that the C-terminal of the PSB domain tend to have **tail1** conformation with good binders but the **tail2** conformation with weak binders or when in the apo form. This observation of conformational difference is mirrored in our simulations and enthalpy calculations. Interestingly, there are charged residues around the C-terminal in both cases of the RPA32 (**Figure 5.3e-f**) and PSB domain of human C-P4H-II (**Figure 5.4b-c**), and they possibly drive the conformational change via electrostatic interactions.

In contrast to large folded domains, peptides are usually in the form of unstructured/disordered molecules in the unbound state[370]. They sometimes undergo folding upon binding, which results in better structural adaptation with a much lower level of conformational freedom. For example, the Bak peptide gets an α-helix when complexed to Bcl-x_L although it is an unstructured random coil in solution[371, 372]. Importantly, the free p53 TAD, which I used here in the benchmark as a peptide (2MWY), is intrinsically disordered. Furthermore, NMR studies showed that the p53 TAD did not tend to create a helical structure in the free form[373-375]. However, the p53 TAD gets more helical upon binding to MDM2, MDMX[376, 377] and the Nuclear Receptor Coactivator Binding Domain of CREB Binding Protein[378]. These structural stabilizations are concomitant with a large enthalpic gain[379]. In the same way, I observed large enthalpic gains in two different protein-peptide systems upon helix formation in MDM2-p53 (2MWY) and GABARAPL2-UBA5 (6H8C) (**Figure 5.5**). The p53-MDM2/MDMX

complexes have been well-characterized as a target for cancer therapy[362, 377, 380-382]. The UBA5 peptide (^{337}EDNEWGIELVSEVSE^{351}) is an unstructured region in crystal structures and the AlphaFold model[383, 384], but it formed a small helix at the N-terminal in our simulations. Consequently, complex simulations having more helix formation in the peptides provided more accurate results for the binding enthalpies of MDM2-p53 and GABARAPL2-UBA5 complexes. Interestingly, there is an artificial cloning artefact (GAM) in the ITC constact of the UBA5 peptide (**Figure 5.5d**). Thus, this artifial extension in the N-terminal possibly drives the helix formation in UBA5 peptide upon binding. On the other hand, the p53 TAD was already a helix in the PDB structure, but it was shorter than I observed in the simulations and the AlphaFold model (**Figure 5.6a**). Similarly, more helix formation ensured more accurate binding enthalpy for the MDM2-p53 complex as well. At this point, it is also worth remembering that the underlying tendency of the force-field to adopt helical conformations is an aspect that has received attention in recent years[385] and that force-field development is an ongoing exercise.[386]

Considering complete experimental conditions in simulations might increase the complications during the setup and the computational time. The choice of the buffer can affect the thermodynamic signatures if there are proton movements between solute and buffer during binding[387]. Moreover, it has been suggested that buffer salts might increase the hydration of the ligands on the binding of agonists and antagonists at the histamine H3 receptor [388]. Nonetheless, the buffer conditions did not appear to dramatically affect the enthalpy calculations in these calculations since I got highly accurate calculations with simulations in pure water (**Figure 5.9**). In addition, I reported accurate binding enthalpies for bromodomain-ligand complexes in pure water after considering ZA-loop dynamics (Chapter 4). On the contrary, Gao *et el*. reported the strong sensitivity of the computed binding enthalpies to the concentration of NaCl (0–500 mM) for CB7-B2 system[53]. Although strong sensitivities to salt concentrations were observed in calculations with different water models, the experimental buffer condition had no significant effect, according to our observations. Here, I reported similar results for binding enthalpies of the PSB domain of human C-P4H-II with two different peptides (6EVO and 6EVN) in pure water and also experimental buffer condition (20 mM TRIS, 50 mM NaCl, and 50 mM glycine). Although I observe no real effect of buffer in this

study, this is an area that requires a more systematic and complete approach beyond the scope of this paper.

More than 50% of proteins include metal ions, making them prevalent in biological systems.[389, 390] Metalloproteins can perform vital tasks like scavenging free radicals and catalysing respiration reactions in the cell.[391, 392] Consequently, two metalloproteins were included in the benchmark; Calmodulin having two Ca^{2+} (PDB: 2LQC) and the PHD zinc finger domain of BAZ2A having two Zn^{2+} (PDB: 4Q6F). 2LQC gave us an accurate ΔH, while 4Q6F provided an incorrect ΔH when I used the default parameters in AMBER ff14SB. Here, the nonbonded model, the simplest model among nonpolarizable models, was used for metals[393]. It only consists of the electrostatic and van der Waals terms. However, it was not reproducible for the hydration-free energy and ion–oxygen distance of the first solvation shell for the zinc ion when using the nonbonded model[348, 394]. Later, Li and Merz proposed bonded models for different zinc-coordinated systems (Zinc AMBER Force Field, ZAFF), which reproduced the experimental hydration-free energy, coordination number, and ion–oxygen distance simultaneously [348, 349, 395]. Obtaining better results after using ZAFF pointed out that further improvement in force field accuracy for Zn^{2+} was needed.

5.5. Conclusion

In conclusion, I have demonstrated that accurate calculations of the absolute enthalpy for the protein-peptide system can be obtained. In addition, as part of this study, I have assembled a benchmark of diverse protein-peptide complexes complete with binding enthalpy/entropy data. Such benchmarks are beneficial for drug discovery studies by providing opportunities for comparing algorithms [35, 36, 120-122]. I initially retrieved 2852 protein-protein complexes from the PDBbind database. Subsequently, I obtained a manually curated dataset consisting of protein-protein structures having binding thermodynamics based on ITC experiments. The dataset is available at doi: 10.5281/zenodo.7635945, including the PDB IDs and ITC details. It should be an invaluable resource for computational structural biologists trying to predict thermodynamic components from the structure and test new methods for protein-protein interactions.

GROMACS reminds me*: "I didn't know what MD was. I think I've managed to catch up."*
~Berk Hess

6 Conclusion and Future Directions

6.1. Conclusion

In this thesis, I conducted several computational studies on the binding enthalpy of different molecular systems. The main focus of the research was to assess the performance of absolute enthalpy predictions within a drug-design context. The technique was specifically tested and applied to systems involving cucurbit[7]uril-guest, BRD4-ligand, and protein-protein interactions.

As molecular modellers study increasingly complex biomolecules and their dynamics over longer periods, the computational requirements continue to increase. MD simulations of macromolecules are highly demanding computationally, making them well-suited for implementation on graphics processing units (GPUs).[16, 396] Modern GPUs greatly surpass CPUs in terms of computational power.[15, 29] Using a single GPU, it has been demonstrated that the method used in this thesis can quickly compute binding enthalpy for a small host-guest system in just a day. The ability to quickly calculate thermodynamic estimates would greatly aid the *in silico* drug discovery process, such as virtual screening and lead optimization. Advanced hardware architecture has made thermodynamic simulations highly accurate and comparable to direct experimental measurements.[2, 368, 397] To be a useful alternative to experimental methods, the method outlined in this thesis, which significantly reduces the time required to obtain results to a few days, will soon be widely available for complex systems.

In Chapter 3, I evaluated the accuracy of water models in conjunction with the host-guest force field for calculating the absolute binding enthalpy for 25 host-guest pairs. It would be ideal to use actual protein-ligand or protein-protein data to evaluate and improve force fields; however, the accurate thermodynamics calculation for protein-ligand or protein-protein systems through simulations is still too computationally intensive to be used in force-field evaluation or optimization. On the other hand, host-guest systems offer a more

straightforward but informative alternative for testing and improving force fields for binding calculations.[57, 398] Host-guest systems are well-suited for study using MD simulations due to their simplicity and non-covalent interactions.[2, 31, 54] It's even possible to conduct these simulations on personal laptop computers.[146] These simulations provide insight into the behaviour of host-guest systems at the atomic level, including the strength of binding, specificity, and stability of the complex under varying conditions. Since host molecules are more rigid and have fewer degrees of freedom than proteins, errors due to inadequate conformational sampling can be reduced, enabling a more precise focus on other sources of error. For these reasons, I aimed to evaluate host-guest binding enthalpies to select force field parameters and water models in Chapter 3 before going to complex systems.[58] Previous studies have shown how the choice of force field and water parameters affects binding enthalpies.[2, 53] It appears that enthalpies are more affected by these parameters than binding free energies.[54] Additionally, Chapter 3 has shown that binding enthalpies of host-guest molecules are particularly sensitive to the type of force field and water models.[58] If this trend holds, it suggests a purely *in silico* case of entropy-enthalpy compensation.[26, 35] Selecting the appropriate force field or water model for biomolecular systems in MD simulations can be challenging due to the continuous emergence of new options.[399-401]

In Chapter 4, I tested the ability of absolute binding enthalpy calculations based on the direct method to predict the binding enthalpies of 10 diverse ligands binding to BRD4-1. Protein-ligand binding plays a crucial role in many cellular processes. The main goal in drug design is to identify a ligand that can effectively change the target protein's activity through high-affinity binding.[5] However, a recent meta-analysis of 853 studies from 34 different research groups found that free energy methods using MD simulations were almost as accurate as chemical methods (with an error of 1.58 kcal/mol).[368] In contrast, binding enthalpy/entropy calculations using MD simulations have not received much attention. Previous studies[34, 35] and Chapter 4, using the BRD4-1 system, demonstrate that it is possible to reach chemical accuracy for the binding enthalpy of protein-ligand systems through MD simulations. However, it is obvious that more comprehensive studies having diverse protein-ligand systems are needed to reach a widely acceptable conclusion. The results in Chapter 4 indicate a significant correlation between the behaviour of the ZA loop and the predicted enthalpy. The ZA loop affects not only enthalpy calculations but also binding energy calculations.[114] The ZA1 (crystallographic) conformation is found

to mostly provide accurate enthalpy results, as shown in Chapter 4 and also binding free energy calculations in previous studies.[112-115] This explains why previous binding energy calculation studies using a crystallographic structure, where short MD windows prevented ZA loop transitions, produced accurate results.

In Chapter 5, the assessment of the method was extended to multiple protein-protein systems. Protein-protein interactions are critical in all cellular activities, from signal transduction to enzymatic activities. Understanding these interactions is crucial for understanding the underlying molecular mechanisms of various diseases and discovering potential drug targets. Combining MD simulations with 3D structures of proteins makes it possible to gain a deeper understanding of the forces driving protein-protein binding.[322, 402] Chapter 5 demonstrates that atomistic simulations can be utilized to accurately calculate binding enthalpies for protein-protein systems in an aqueous solution and study the binding process's calorimetric characteristics. The method was able to produce highly precise and well-converged binding enthalpies for small protein-protein systems. However, inaccurate enthalpies resulted from incorrect conformational sampling, manifesting as conformational states not observed in experimental structures. It is probable that the most effective case for estimating binding enthalpy values will be in complexes that are conformationally well-ordered and do not have prolonged conformational fluctuations that significantly impact the overall potential energy of the system. Thus, the simulator should be careful when interpreting the outcomes of MD simulations, as they may not always precisely give the actual behaviour of the molecular system in solution.

We desire to accurately predict the thermodynamic properties (changes in free energy, enthalpy and entropy) of the binding event between any protein and its binding partner. While recent progress has allowed relatively good prediction of binding free energies, predicting binding enthalpies remains very difficult, and improvement in enthalpy predictions has been limited. Reliably calculating absolute binding enthalpy values remains a significant challenge in biomolecular interactions. Predicting quantitative enthalpy changes upon binding has remained an important unsolved problem. Better enthalpy prediction would be highly desirable to understand the thermodynamics of the binding fully.

One potential reason for the limited progress in enthalpy predictions is the belief that the amount of sampling required to get accurate results is too large and impractical. The initial

results confirmed this on complex systems (**Figure 5.2**), as they showed inferior agreement with experimental values. In some cases, the simulations could not transition between two conformational states, implying a significant energy barrier between them. Thus, getting an accurate result might be difficult if simulations cannot sufficiently sample the relevant conformations. However, the calculations still provide valuable insights into the energetics and structural dynamics despite the limitations.

Despite the initially poor enthalpy predictions, there are a few things to take away. First, the lack of agreement indicates that the simulations are not correctly capturing the system's dynamic behaviour. So, the protein/molecule might get stuck in one conformational state that does not match the actual conformational space giving accurate binding enthalpy. So even though the predicted enthalpy is wrong, it signals that there are likely other relevant conformations that the simulations are not sampling. Second, suppose one is interested in the dynamics of a complex. A wrong enthalpy prediction can indicate that the simulation needs more sampling to explore other conformations. Thus, the calculations offer helpful information about the binding enthalpy along the conformational landscape of the simulated system.

Although an artificial guiding conformational sampling, taking the best predictions across all the simulations, gives excellent agreement with the experiment (**Figure 4.4 and 5.9**), the enthalpy predictions can be quite accurate when the relevant conformations are sufficiently sampled. A commonly used force field, AMBER ff14SB, can potentially predict enthalpies to a high level of accuracy if adequate conformational sampling is done. However, it's important to remember that a force field should ideally predict both binding free energies and enthalpies correctly - good performance on binding enthalpies does not mean those binding free energies will be correct as well. With enough sampling to cover the essential conformational landscape, the force field can yield quantitative enthalpy predictions.

The method showed impressive performance in predicting binding enthalpies for various systems, but only after artificially guiding the conformational sampling based on experimental data. This may be overstating the true capabilities of the method since actual blind predictions wouldn't have access to experimental data to guide the sampling. One should also be cautious about the accuracy of experimental enthalpy values from calorimetry.[26] The inaccurate enthalpy predictions stemmed from the inadequate

sampling of relevant conformations and missing conformational states not seen in experimental structures. The subsequent analyses and discovery of alternative conformations and their potential role in the discrepancies are exciting findings. One takeaway is that future work could start simulations from conformations other than experimental ones.

The most tempting test cases for reliably estimating binding enthalpies without experimental data may be conformationally ordered complexes without significant conformational fluctuations that dramatically impact the potential energy. One needs a good knowledge of the conformational landscape from simulations or literature data to properly guide the enthalpy calculations. Otherwise, blind predictions without experimental data could yield inaccurate results if the system's conformational space is insufficiently characterized. A prior understanding of the system seems essential for robust computational prediction of binding enthalpies without experimental data.

Overall, in this thesis, I discussed the validation and application of absolute binding enthalpy calculations using the direct method based on statistical mechanics principles and uses physics-based computer simulations of molecular systems. The capability to perform precise enthalpy calculation along with accurate free energy predictions brings us closer to being able to design ligands with the desired thermodynamic properties. This goal has been pursued for a long time. The information presented can be helpful for computational and theoretical chemists to improve and apply these methods, as well as for medicinal chemists and structural biologists specifically interested in molecular recognition.

6.2. Future Directions

I found host-guest systems an informative alternative for testing and improving force fields for thermodynamics calculations. Although I used only one host molecule (CB7) in Chapter 3, an assessment study involving much more diverse host molecules such as cucurbiturils, cavitands, calixarenes, pillararenes and cyclodextrins would be much more informative. In addition, future studies using a diverse set of host-guest systems will be more informative with the addition of newly added or untested water models (e.g. OPC3[403], TIP5P/2018[399], TIP7P[400]) or force fields (e.g. OPLS4[401], AMOEBA[404], Sage[224]).

In Chapter 4, I only considered one target protein, but a more informative study would involve a wider range of drug targets, including MCL-1[189, 264], EZH2[405], MDM2[406], and others.[261] Forthcoming studies having diverse protein-ligand systems will provide us with much more insight into accurate binding enthalpy calculations.

Although I use small protein-protein systems to reduce computational demand here, enthalpy calculations of larger systems potentially having more conformational fluctuations will become possible in the future. For example, the binding enthalpies of non-covalent interactions between an antibody and antigen association can be examined by using available structural and thermodynamic data.[407] These investigations can provide insights into the high affinity and specificity of the antigen-antibody interaction by examining the structural and thermodynamic interactions.

The study of interactions between molecules also includes interactions involving proteins or ligands that bind to DNA or RNA. It is known that the interactions between functional nucleic acids and specific proteins are responsible for many cellular processes involving DNA or RNA. These interactions are essential for many processes, including replication, transcription, repair, transport, translation, splicing, and silencing of RNA.[408] Furthermore, small molecules binding nucleic acids can activate molecular functions for the treatment of diseases.[409, 410] Thus, it is important to understand how proteins or ligands interact with nucleic acids in order to comprehend the significance of these interactions in various biological processes and medical applications.[411] We can also conduct absolute binding enthalpy calculations, providing insight into the molecular driving forces of the interactions between protein/ligand and DNA/RNA.

The limited use of absolute binding enthalpy is due to the significant amount of human effort required to perform. The process involves assigning force field parameters, creating initial system configurations, establishing bound and unbound states, equilibrating separated simulations, conducting multiple production runs, and analysing the results. An open-source package written in a flexible and widely used programming language would greatly increase the adoption of absolute binding enthalpy by making it easier to use, replicate, customize, and extend these calculations.

Appendices

Supplementary material for Chapter 3

Table S3.1: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for OPC Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-14.61 ±0.56	-17.71 ±0.54	-21.42 ±0.53	-16.57 ±0.54	-22.00 ±0.55
A02	-19.30 ±0.23	-15.49 ±0.55	-19.16 ±0.54	-23.31 ±0.53	-19.23 ±0.53	-21.89 ±0.56
A03	-21.90 ±0.23	-16.90 ±0.55	-21.75 ±0.54	-24.15 ±0.54	-19.92 ±0.54	-24.06 ±0.56
A04	-20.10 ±0.23	-17.75 ±0.55	-22.33 ±0.54	-25.38 ±0.55	-21.70 ±0.54	-24.22 ±0.56
A05	-19.50 ±0.23	-16.15 ±0.56	-19.62 ±0.54	-23.52 ±0.54	-19.92 ±0.53	-23.73 ±0.56
B02	-15.80 ±0.12	-14.64 ±0.55	-16.68 ±0.53	-17.57 ±0.53	-16.36 ±0.52	-21.50 ±0.55
B05	-15.60 ±0.23	-17.68 ±0.55	-17.07 ±0.54	-21.85 ±0.53	-19.10 ±0.54	-24.32 ±0.56
B11	-16.30 ±0.23	-18.65 ±0.55	-18.05 ±0.55	-22.58 ±0.56	-21.37 ±0.54	-26.31 ±0.57
BER	-9.08 ±0.29	-9.88 ±0.55	-6.59 ±0.54	0.22 ±0.54	-18.20 ±0.54	-11.04 ±0.57
CA0	-9.91 ±0.12	-12.08 ±0.54	-8.95 ±0.54	-15.13 ±0.53	-11.39 ±0.53	-16.47 ±0.56
CA1	-10.82 ±0.06	-13.19 ±0.54	-8.25 ±0.54	-12.96 ±0.53	-9.91 ±0.54	-16.59 ±0.55
CA2	-11.58 ±0.06	-15.97 ±0.56	-9.02 ±0.55	-12.37 ±0.54	-13.27 ±0.53	-19.24 ±0.55
CTE	-13.20 ±2.31	-7.10 ±0.54	-7.72 ±0.54	-0.98 ±0.56	-11.42 ±0.54	-5.31 ±0.56
DHC	-8.12 ±0.14	-12.06 ±0.55	-6.56 ±0.55	8.11 ±0.55	-15.07 ±0.55	-11.81 ±0.56
FTE	-11.10 ±0.23	-11.64 ±0.55	-11.13 ±0.54	-10.02 ±0.54	-16.99 ±0.53	-11.83 ±0.56
NTE	-15.60 ±1.27	-12.59 ±0.56	-10.15 ±0.54	-8.20 ±0.54	-13.78 ±0.53	-13.75 ±0.56
P01	-3.98 ±0.01	-12.47 ±0.55	-5.55 ±0.54	-7.85 ±0.54	-8.52 ±0.53	-10.96 ±0.56
P02	-6.45 ±0.06	-13.64 ±0.56	-6.14 ±0.54	-6.63 ±0.54	-11.36 ±0.54	-21.41 ±0.56
P03	-5.16 ±0.02	-15.15 ±0.56	-3.84 ±0.54	-5.44 ±0.53	-10.88 ±0.53	-13.47 ±0.56
P04	-4.55 ±0.09	-17.64 ±0.56	-6.03 ±0.55	-6.54 ±0.53	-13.03 ±0.53	-18.20 ±0.56
P1A	-6.98 ±0.02	-16.93 ±0.55	-7.35 ±0.53	-6.89 ±0.55	-10.73 ±0.53	-13.62 ±0.55
P2A	-6.58 ±0.01	-18.58 ±0.55	-7.31 ±0.54	-8.92 ±0.54	-13.86 ±0.53	-16.17 ±0.56
P3A	-7.08 ±0.05	-17.85 ±0.55	-6.88 ±0.55	-6.73 ±0.55	-10.55 ±0.53	-18.05 ±0.57
PAL	-8.84 ±0.37	-14.73 ±0.56	-10.79 ±0.54	-2.64 ±0.54	-19.17 ±0.54	-14.96 ±0.57
THP	-7.31 ±0.24	-12.82 ±0.56	-6.36 ±0.54	-2.41 ±0.54	-14.08 ±0.54	-10.17 ±0.55
	Slope	0.09	0.99	1.30	0.53	0.62
	Intercept	-13.56	0.42	3.86	-8.86	-9.96
	R^2	0.03	0.88	0.61	0.53	0.38
	Kendall's τ	0.10	0.80	0.55	0.56	0.43
	RMSE	6.32	2.07	5.75	4.94	7.25
	MSE	39.92	4.28	33.10	24.41	52.50
	MAE	5.23	1.52	4.27	4.02	6.27

Table S3.2: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for SPC Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-18.34 ±0.48	-22.91 ±0.47	-24.85 ±0.48	-20.80 ±0.48	-23.62 ±0.47
A02	-19.30 ±0.23	-16.97 ±0.48	-22.01 ±0.47	-24.67 ±0.48	-21.39 ±0.48	-23.29 ±0.48
A03	-21.90 ±0.23	-18.70 ±0.48	-23.83 ±0.47	-26.54 ±0.49	-21.35 ±0.48	-24.59 ±0.48
A04	-20.10 ±0.23	-18.07 ±0.48	-23.40 ±0.47	-25.95 ±0.48	-22.71 ±0.48	-24.89 ±0.48
A05	-19.50 ±0.23	-17.25 ±0.48	-22.09 ±0.47	-25.91 ±0.48	-22.04 ±0.48	-24.73 ±0.48
B02	-15.80 ±0.12	-17.05 ±0.47	-20.04 ±0.47	-20.08 ±0.47	-18.69 ±0.47	-22.89 ±0.48
B05	-15.60 ±0.23	-17.36 ±0.47	-17.95 ±0.47	-21.78 ±0.48	-18.92 ±0.47	-22.38 ±0.47
B11	-16.30 ±0.23	-19.74 ±0.49	-19.26 ±0.48	-23.81 ±0.49	-21.73 ±0.49	-25.83 ±0.48
BER	-9.08 ±0.29	-11.45 ±0.48	-7.71 ±0.47	-1.15 ±0.49	-19.04 ±0.48	-10.89 ±0.48
CA0	-9.91 ±0.12	-12.92 ±0.48	-11.81 ±0.46	-17.70 ±0.48	-13.55 ±0.48	-18.05 ±0.47
CA1	-10.82 ±0.06	-14.32 ±0.48	-10.32 ±0.48	-15.31 ±0.49	-14.19 ±0.48	-16.65 ±0.47
CA2	-11.58 ±0.06	-15.89 ±0.48	-10.90 ±0.47	-13.04 ±0.48	-14.11 ±0.47	-19.21 ±0.48
CTE	-13.20 ±2.31	-9.14 ±0.47	-10.53 ±0.47	-8.73 ±0.48	-14.45 ±0.47	-6.40 ±0.47
DHC	-8.12 ±0.14	-13.67 ±0.48	-8.49 ±0.47	-5.14 ±0.48	-17.49 ±0.47	-13.14 ±0.47
FTE	-11.10 ±0.23	-10.57 ±0.48	-12.78 ±0.47	-10.48 ±0.48	-17.28 ±0.48	-12.30 ±0.48
NTE	-15.60 ±1.27	-12.15 ±0.48	-13.07 ±0.48	-8.04 ±0.48	-16.91 ±0.48	-12.00 ±0.47
P01	-3.98 ±0.01	-10.01 ±0.48	-6.12 ±0.47	-6.87 ±0.48	-6.17 ±0.48	-10.91 ±0.48
P02	-6.45 ±0.06	-10.73 ±0.49	-5.01 ±0.47	-5.70 ±0.48	-10.21 ±0.47	-17.87 ±0.47
P03	-5.16 ±0.02	-12.70 ±0.48	-3.13 ±0.47	-3.56 ±0.48	-9.37 ±0.47	-11.49 ±0.47
P04	-4.55 ±0.09	-14.63 ±0.48	-5.14 ±0.47	-5.78 ±0.47	-11.09 ±0.48	-15.19 ±0.47
P1A	-6.98 ±0.02	-13.59 ±0.47	-6.82 ±0.47	-5.46 ±0.48	-8.76 ±0.47	-11.44 ±0.48
P2A	-6.58 ±0.01	-15.16 ±0.47	-6.38 ±0.48	-6.44 ±0.48	-11.19 ±0.48	-14.78 ±0.47
P3A	-7.08 ±0.05	-15.11 ±0.48	-6.50 ±0.47	-6.11 ±0.48	-8.92 ±0.48	-15.89 ±0.47
PAL	-8.84 ±0.37	-15.48 ±0.48	-12.24 ±0.47	-3.97 ±0.49	-18.92 ±0.48	-15.52 ±0.47
THP	-7.31 ±0.24	-12.52 ±0.49	-8.32 ±0.48	-3.56 ±0.49	-15.44 ±0.48	-10.94 ±0.48
	Slope	0.35	1.18	1.43	0.76	0.76
	Intercept	-10.37	1.22	3.96	-6.85	-8.07
	R^2	0.43	0.93	0.79	0.73	0.54
	Kendall's τ	0.45	0.85	0.61	0.68	0.51
	RMSE	4.95	2.22	4.73	4.92	6.58
	MSE	24.52	4.92	22.36	24.17	43.35
	MAE	4.27	1.89	4.04	4.08	6.07

Table S3.3: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for SPC/E Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-18.04 ±0.55	-22.03 ±0.76	-24.00 ±0.54	-21.62 ±0.54	-23.72 ±0.54
A02	-19.30 ±0.23	-17.69 ±0.56	-21.10 ±0.77	-25.41 ±0.54	-22.84 ±0.54	-24.11 ±0.55
A03	-21.90 ±0.23	-20.03 ±0.55	-23.35 ±0.76	-26.36 ±0.54	-24.29 ±0.54	-25.14 ±0.55
A04	-20.10 ±0.23	-19.57 ±0.55	-24.46 ±0.79	-26.41 ±0.54	-24.06 ±0.54	-26.22 ±0.55
A05	-19.50 ±0.23	-18.77 ±0.55	-21.95 ±0.76	-26.97 ±0.54	-22.93 ±0.55	-25.44 ±0.54
B02	-15.80 ±0.12	-18.37 ±0.55	-18.38 ±0.79	-20.22 ±0.54	-19.71 ±0.55	-22.09 ±0.54
B05	-15.60 ±0.23	-19.68 ±0.55	-19.68 ±0.76	-23.31 ±0.54	-22.84 ±0.54	-25.07 ±0.55
B11	-16.30 ±0.23	-21.51 ±0.55	-21.10 ±0.77	-26.04 ±0.55	-24.76 ±0.55	-28.36 ±0.55
BER	-9.08 ±0.29	-14.27 ±0.55	-8.84 ±0.78	-2.57 ±0.54	-20.90 ±0.54	-13.10 ±0.55
CA0	-9.91 ±0.12	-13.32 ±0.55	-11.27 ±0.76	-17.33 ±0.54	-15.34 ±0.55	-18.13 ±0.53
CA1	-10.82 ±0.06	-15.21 ±0.55	-10.79 ±0.77	-16.23 ±0.53	-15.16 ±0.54	-18.28 ±0.54
CA2	-11.58 ±0.06	-16.77 ±0.54	-12.01 ±0.78	-14.80 ±0.54	-15.91 ±0.54	-20.00 ±0.54
CTE	-13.20 ±2.31	-9.95 ±0.55	-11.63 ±0.76	-9.83 ±0.54	-15.79 ±0.54	-7.09 ±0.54
DHC	-8.12 ±0.14	-15.88 ±0.56	-4.55 ±0.77	-6.09 ±0.55	-18.97 ±0.54	-14.29 ±0.54
FTE	-11.10 ±0.23	-11.54 ±0.55	-14.21 ±0.77	-11.92 ±0.55	-20.80 ±0.54	-14.39 ±0.54
NTE	-15.60 ±1.27	-13.92 ±0.55	-13.74 ±0.77	-10.75 ±0.54	-18.61 ±0.54	-15.30 ±0.53
P01	-3.98 ±0.01	-12.37 ±0.55	-7.02 ±0.80	-7.70 ±0.54	-8.80 ±0.54	-12.22 ±0.54
P02	-6.45 ±0.06	-13.37 ±0.55	-4.97 ±0.77	-6.08 ±0.54	-12.54 ±0.54	-21.38 ±0.54
P03	-5.16 ±0.02	-15.12 ±0.54	-4.53 ±0.78	-5.01 ±0.53	-11.70 ±0.55	-13.63 ±0.54
P04	-4.55 ±0.09	-17.45 ±0.55	-6.57 ±0.77	-7.15 ±0.54	-13.44 ±0.55	-17.70 ±0.54
P1A	-6.98 ±0.02	-15.72 ±0.54	-9.27 ±0.77	-7.24 ±0.53	-11.96 ±0.53	-13.54 ±0.54
P2A	-6.58 ±0.01	-17.73 ±0.54	-8.49 ±0.79	-9.53 ±0.53	-14.86 ±0.55	-17.16 ±0.54
P3A	-7.08 ±0.05	-17.92 ±0.55	-7.86 ±0.77	-8.31 ±0.54	-11.94 ±0.54	-17.03 ±0.54
PAL	-8.84 ±0.37	-17.91 ±0.56	-13.31 ±0.77	-5.32 ±0.54	-21.99 ±0.54	-16.64 ±0.55
THP	-7.31 ±0.24	-14.97 ±0.55	-10.11 ±0.77	-4.67 ±0.55	-18.01 ±0.55	-12.70 ±0.53
	Slope	0.26	1.11	1.37	0.70	0.69
	Intercept	-13.26	-0.24	2.14	-9.79	-10.43
	R^2	0.24	0.89	0.79	0.67	0.48
	Kendall's τ	0.35	0.80	0.62	0.68	0.50
	RMSE	6.51	2.60	4.81	6.96	7.95
	MSE	42.38	6.78	23.16	48.51	63.19
	MAE	5.38	2.25	4.09	6.24	7.27

Table S3.4: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for TIP3P Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-17.00 ±0.42	-22.67 ±0.45	-24.81 ±0.45	-22.30 ±0.44	-20.47 ±0.42
A02	-19.30 ±0.23	-14.43 ±0.42	-20.27 ±0.45	-25.41 ±0.45	-21.90 ±0.45	-18.45 ±0.43
A03	-21.90 ±0.23	-16.78 ±0.43	-23.02 ±0.45	-25.95 ±0.45	-22.89 ±0.45	-19.49 ±0.43
A04	-20.10 ±0.23	-16.50 ±0.43	-21.20 ±0.45	-25.46 ±0.46	-23.14 ±0.45	-20.51 ±0.43
A05	-19.50 ±0.23	-16.29 ±0.43	-21.65 ±0.45	-24.85 ±0.45	-22.35 ±0.46	-19.56 ±0.42
B02	-15.80 ±0.12	-16.19 ±0.42	-19.10 ±0.44	-20.15 ±0.44	-20.50 ±0.45	-19.51 ±0.43
B05	-15.60 ±0.23	-14.52 ±0.42	-16.84 ±0.44	-20.57 ±0.45	-19.12 ±0.45	-16.93 ±0.43
B11	-16.30 ±0.23	-16.20 ±0.44	-17.37 ±0.46	-21.85 ±0.46	-20.44 ±0.46	-18.84 ±0.43
BER	-9.08 ±0.29	-8.67 ±0.42	-5.02 ±0.46	-0.13 ±0.45	-18.47 ±0.46	-6.18 ±0.42
CA0	-9.91 ±0.12	-11.92 ±0.43	-10.82 ±0.44	-17.05 ±0.45	-14.61 ±0.45	-13.84 ±0.42
CA1	-10.82 ±0.06	-12.60 ±0.42	-9.98 ±0.44	-14.85 ±0.45	-14.90 ±0.45	-12.45 ±0.43
CA2	-11.58 ±0.06	-14.09 ±0.42	-9.57 ±0.45	-13.49 ±0.45	-14.56 ±0.46	-13.83 ±0.42
CTE	-13.20 ±2.31	-8.05 ±0.42	-9.06 ±0.45	-8.59 ±0.45	-13.75 ±0.45	-3.57 ±0.43
DHC	-8.12 ±0.14	-11.84 ±0.43	-6.84 ±0.45	-3.98 ±0.46	-16.49 ±0.45	-8.23 ±0.43
FTE	-11.10 ±0.23	-7.33 ±0.43	-11.21 ±0.45	-9.36 ±0.45	-16.95 ±0.46	-7.81 ±0.43
NTE	-15.60 ±1.27	-8.62 ±0.42	-12.61 ±0.45	-7.91 ±0.45	-16.86 ±0.46	-8.46 ±0.43
P01	-3.98 ±0.01	-7.46 ±0.43	-4.38 ±0.44	-5.33 ±0.44	-6.98 ±0.45	-6.11 ±0.42
P02	-6.45 ±0.06	-8.52 ±0.43	-3.43 ±0.45	-4.15 ±0.46	-10.35 ±0.45	-13.45 ±0.43
P03	-5.16 ±0.02	-9.73 ±0.42	-2.35 ±0.45	-3.71 ±0.45	-10.02 ±0.44	-7.38 ±0.43
P04	-4.55 ±0.09	-11.92 ±0.42	-2.74 ±0.44	-4.26 ±0.45	-10.41 ±0.45	-10.28 ±0.43
P1A	-6.98 ±0.02	-10.32 ±0.43	-4.92 ±0.44	-4.37 ±0.45	-9.16 ±0.45	-7.20 ±0.43
P2A	-6.58 ±0.01	-11.68 ±0.43	-4.92 ±0.45	-6.16 ±0.45	-10.96 ±0.45	-9.46 ±0.43
P3A	-7.08 ±0.05	-12.41 ±0.42	-4.13 ±0.45	-5.46 ±0.45	-8.98 ±0.45	-8.77 ±0.44
PAL	-8.84 ±0.37	-13.27 ±0.42	-10.27 ±0.45	-3.82 ±0.45	-19.30 ±0.45	-10.26 ±0.43
THP	-7.31 ±0.24	-11.16 ±0.42	-7.00 ±0.44	-2.99 ±0.46	-14.89 ±0.46	-7.23 ±0.43
	Slope	0.40	1.23	1.46	0.81	0.77
	Intercept	-7.57	3.20	5.01	-6.52	-3.24
	R^2	0.49	0.93	0.80	0.79	0.61
	Kendall's τ	0.54	0.82	0.64	0.71	0.51
	RMSE	3.94	2.22	4.62	4.92	3.59
	MSE	15.49	4.93	21.34	24.25	12.86
	MAE	3.45	1.90	4.05	4.26	2.68

Table S3.5: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for TIP4P Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-19.85 ±0.53	-23.46 ±0.51	-26.36 ±0.54	-23.41 ±0.51	-25.49 ±0.53
A02	-19.30 ±0.23	-18.81 ±0.54	-23.54 ±0.52	-27.89 ±0.53	-23.80 ±0.53	-26.79 ±0.53
A03	-21.90 ±0.23	-21.87 ±0.54	-25.73 ±0.52	-28.72 ±0.53	-25.90 ±0.52	-27.63 ±0.52
A04	-20.10 ±0.23	-20.94 ±0.54	-26.55 ±0.51	-32.16 ±0.53	-27.36 ±0.52	-28.21 ±0.53
A05	-19.50 ±0.23	-19.92 ±0.54	-24.23 ±0.51	-28.47 ±0.53	-25.40 ±0.53	-28.63 ±0.53
B02	-15.80 ±0.12	-19.35 ±0.53	-22.24 ±0.52	-23.28 ±0.52	-21.82 ±0.51	-25.33 ±0.52
B05	-15.60 ±0.23	-20.47 ±0.53	-20.97 ±0.50	-26.23 ±0.53	-22.97 ±0.52	-26.81 ±0.53
B11	-16.30 ±0.23	-22.01 ±0.55	-21.76 ±0.53	-27.90 ±0.53	-25.26 ±0.53	-28.49 ±0.54
BER	-9.08 ±0.29	-14.16 ±0.53	-10.76 ±0.51	-4.11 ±0.54	-22.43 ±0.52	-14.43 ±0.53
CA0	-9.91 ±0.12	-14.93 ±0.53	-13.48 ±0.52	-19.97 ±0.53	-16.46 ±0.52	-20.57 ±0.52
CA1	-10.82 ±0.06	-15.84 ±0.52	-12.50 ±0.52	-17.68 ±0.53	-14.85 ±0.52	-19.62 ±0.51
CA2	-11.58 ±0.06	-19.13 ±0.53	-12.12 ±0.52	-17.43 ±0.53	-17.78 ±0.52	-22.20 ±0.52
CTE	-13.20 ±2.31	-9.50 ±0.54	-11.66 ±0.44	-11.24 ±0.53	-16.67 ±0.52	-7.20 ±0.53
DHC	-8.12 ±0.14	-15.35 ±0.53	-11.17 ±0.44	2.13 ±0.54	-19.79 ±0.52	-15.36 ±0.54
FTE	-11.10 ±0.23	-14.77 ±0.54	-14.15 ±0.45	-12.76 ±0.52	-20.37 ±0.51	-16.91 ±0.52
NTE	-15.60 ±1.27	-13.04 ±0.52	-11.58 ±0.45	-12.10 ±0.53	-16.78 ±0.52	-16.67 ±0.53
P01	-3.98 ±0.01	-11.37 ±0.52	-9.39 ±0.45	-11.11 ±0.53	-10.52 ±0.53	-13.53 ±0.53
P02	-6.45 ±0.06	-13.47 ±0.54	-8.58 ±0.45	-10.83 ±0.52	-13.79 ±0.53	-22.21 ±0.52
P03	-5.16 ±0.02	-15.42 ±0.54	-6.79 ±0.45	-8.71 ±0.52	-12.08 ±0.53	-14.94 ±0.53
P04	-4.55 ±0.09	-17.84 ±0.52	-8.97 ±0.44	-10.13 ±0.53	-15.71 ±0.54	-19.18 ±0.53
P1A	-6.98 ±0.02	-17.82 ±0.53	-10.57 ±0.44	-9.84 ±0.53	-13.77 ±0.52	-14.34 ±0.53
P2A	-6.58 ±0.01	-19.14 ±0.54	-10.58 ±0.45	-12.24 ±0.53	-16.17 ±0.52	-17.64 ±0.52
P3A	-7.08 ±0.05	-18.60 ±0.54	-10.30 ±0.44	-10.74 ±0.54	-12.93 ±0.53	-19.21 ±0.53
PAL	-8.84 ±0.37	-17.40 ±0.54	-13.93 ±0.45	-7.69 ±0.52	-23.21 ±0.53	-17.14 ±0.53
THP	-7.31 ±0.24	-15.06 ±0.54	-10.51 ±0.46	-6.45 ±0.54	-18.19 ±0.54	-14.59 ±0.53
	Slope	0.32	1.06	1.41	0.72	0.76
	Intercept	-13.28	-2.55	0.47	-10.62	-11.22
	R^2	0.28	0.86	0.70	0.67	0.51
	Kendall's τ	0.37	0.78	0.59	0.68	0.50
	RMSE	6.98	4.02	6.94	7.98	9.37
	MSE	48.68	16.15	48.17	63.65	87.85
	MAE	5.83	3.71	6.14	7.34	8.85

Table S3.6: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for TIP4P-Ew Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-17.74 ±0.58	-22.21 ±0.56	-25.59 ±0.58	-20.23 ±0.57	-24.77 ±0.59
A02	-19.30 ±0.23	-18.69 ±0.59	-22.96 ±0.57	-26.72 ±0.57	-23.22 ±0.58	-26.52 ±0.59
A03	-21.90 ±0.23	-20.40 ±0.60	-24.48 ±0.58	-28.74 ±0.58	-22.88 ±0.58	-28.06 ±0.59
A04	-20.10 ±0.23	-21.02 ±0.59	-26.70 ±0.58	-30.07 ±0.58	-25.88 ±0.58	-28.36 ±0.59
A05	-19.50 ±0.23	-18.79 ±0.59	-24.33 ±0.56	-27.30 ±0.58	-23.50 ±0.57	-26.48 ±0.59
B02	-15.80 ±0.12	-17.88 ±0.58	-21.52 ±0.56	-21.71 ±0.57	-20.19 ±0.58	-24.86 ±0.59
B05	-15.60 ±0.23	-21.06 ±0.59	-21.99 ±0.57	-25.59 ±0.59	-22.26 ±0.57	-27.34 ±0.58
B11	-16.30 ±0.23	-22.83 ±0.61	-23.18 ±0.57	-28.94 ±0.59	-24.52 ±0.59	-29.82 ±0.59
BER	-9.08 ±0.29	-14.97 ±0.60	-11.25 ±0.57	-4.10 ±0.58	-21.84 ±0.58	-14.65 ±0.59
CA0	-9.91 ±0.12	-14.15 ±0.59	-12.56 ±0.56	-18.61 ±0.57	-14.92 ±0.57	-20.07 ±0.59
CA1	-10.82 ±0.06	-15.82 ±0.59	-12.19 ±0.56	-16.82 ±0.58	-15.48 ±0.58	-19.38 ±0.59
CA2	-11.58 ±0.06	-17.98 ±0.59	-12.77 ±0.57	-16.47 ±0.58	-16.88 ±0.58	-22.60 ±0.58
CTE	-13.20 ±2.31	-10.68 ±0.58	-12.10 ±0.56	-9.40 ±0.58	-14.30 ±0.58	-12.89 ±0.59
DHC	-8.12 ±0.14	-8.48 ±0.60	-10.88 ±0.57	2.15 ±0.58	-19.60 ±0.57	-15.08 ±0.59
FTE	-11.10 ±0.23	-14.85 ±0.59	-16.25 ±0.56	-13.42 ±0.59	-20.96 ±0.57	-20.82 ±0.60
NTE	-15.60 ±1.27	-14.56 ±0.60	-15.36 ±0.56	-12.38 ±0.58	-17.05 ±0.57	-17.99 ±0.58
P01	-3.98 ±0.01	-14.07 ±0.59	-9.68 ±0.57	-10.53 ±0.58	-10.36 ±0.57	-13.54 ±0.59
P02	-6.45 ±0.06	-15.09 ±0.58	-8.90 ±0.57	-9.13 ±0.57	-13.66 ±0.58	-23.20 ±0.59
P03	-5.16 ±0.02	-17.47 ±0.58	-7.50 ±0.56	-8.36 ±0.58	-13.16 ±0.57	-15.64 ±0.59
P04	-4.55 ±0.09	-19.08 ±0.59	-9.90 ±0.56	-10.44 ±0.57	-15.48 ±0.58	-21.09 ±0.59
P1A	-6.98 ±0.02	-18.91 ±0.59	-11.88 ±0.56	-10.09 ±0.58	-14.13 ±0.57	-15.67 ±0.58
P2A	-6.58 ±0.01	-19.57 ±0.60	-12.84 ±0.57	-12.95 ±0.58	-17.62 ±0.58	-19.44 ±0.59
P3A	-7.08 ±0.05	-19.85 ±0.60	-10.39 ±0.57	-11.66 ±0.58	-12.76 ±0.58	-22.40 ±0.58
PAL	-8.84 ±0.37	-18.72 ±0.60	-15.10 ±0.57	-7.87 ±0.59	-22.18 ±0.59	-18.04 ±0.59
THP	-7.31 ±0.24	-15.42 ±0.61	-10.80 ±0.57	-6.26 ±0.58	-17.99 ±0.58	-14.27 ±0.59
	Slope	0.21	1.01	1.35	0.58	0.68
	Intercept	-14.63	-3.69	0.28	-11.67	-12.92
	R^2	0.12	0.86	0.69	0.55	0.52
	Kendall's τ	0.20	0.78	0.56	0.60	0.47
	RMSE	7.48	4.32	6.52	7.59	9.97
	MSE	55.97	18.67	42.57	57.68	99.37
	MAE	5.98	3.86	5.83	6.69	9.19

Table S3.7: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for TIP5P Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-19.00 ±0.23	47.49 ±0.88	-17.00 ±0.86	-21.68 ±0.84	-19.51 ±0.85
A02	-19.30 ±0.23	-19.30 ±0.23	45.24 ±0.87	-21.63 ±0.84	-26.65 ±0.85	-23.64 ±0.87
A03	-21.90 ±0.23	-21.90 ±0.23	40.56 ±0.87	-24.99 ±0.85	-29.54 ±0.83	-26.74 ±0.85
A04	-20.10 ±0.23	-20.10 ±0.23	43.72 ±0.87	-24.86 ±0.87	-29.05 ±0.86	-28.29 ±0.83
A05	-19.50 ±0.23	-19.50 ±0.23	44.37 ±0.88	-21.10 ±0.86	-27.38 ±0.83	-24.43 ±0.86
B02	-15.80 ±0.12	-15.80 ±0.12	47.08 ±0.86	-15.04 ±0.85	-15.51 ±0.84	-18.22 ±0.85
B05	-15.60 ±0.23	-15.60 ±0.23	38.07 ±0.88	-22.02 ±0.83	-27.10 ±0.84	-27.90 ±0.85
B11	-16.30 ±0.23	-16.30 ±0.23	38.46 ±0.89	-26.30 ±0.84	-30.45 ±0.86	-31.78 ±0.85
BER	-9.08 ±0.29	-9.08 ±0.29	48.00 ±0.88	-15.23 ±0.85	-9.17 ±0.85	-30.31 ±0.87
CA0	-9.91 ±0.12	-9.91 ±0.12	48.11 ±0.86	-9.07 ±0.84	-14.72 ±0.83	-15.07 ±0.85
CA1	-10.82 ±0.06	-10.82 ±0.06	44.82 ±0.86	-8.58 ±0.84	-14.88 ±0.86	-18.49 ±0.84
CA2	-11.58 ±0.06	-11.58 ±0.06	42.75 ±0.86	-13.28 ±0.87	-16.02 ±0.85	-17.42 ±0.84
CTE	-13.20 ±2.31	-13.20 ±2.31	51.64 ±0.86	-9.30 ±0.84	-6.67 ±0.85	-17.41 ±0.84
DHC	-8.12 ±0.14	-8.12 ±0.14	48.59 ±0.87	-13.76 ±0.84	-10.50 ±0.88	-24.68 ±0.86
FTE	-11.10 ±0.23	-11.10 ±0.23	52.24 ±0.86	-12.83 ±0.85	-12.95 ±0.86	-24.39 ±0.84
NTE	-15.60 ±1.27	-15.60 ±1.27	46.38 ±0.88	-13.52 ±0.86	-13.43 ±0.85	-14.54 ±0.84
P01	-3.98 ±0.01	-3.98 ±0.01	45.40 ±0.89	-11.46 ±0.84	-14.60 ±0.83	-18.65 ±0.84
P02	-6.45 ±0.06	-6.45 ±0.06	43.85 ±0.86	-13.00 ±0.84	-15.35 ±0.84	-22.12 ±0.86
P03	-5.16 ±0.02	-5.16 ±0.02	40.73 ±0.88	-12.94 ±0.84	-14.55 ±0.85	-22.02 ±0.85
P04	-4.55 ±0.09	-4.55 ±0.09	38.93 ±0.88	-15.45 ±0.85	-18.06 ±0.84	-24.91 ±0.86
P1A	-6.98 ±0.02	-6.98 ±0.02	38.61 ±0.88	-16.45 ±0.85	-18.68 ±0.84	-22.81 ±0.84
P2A	-6.58 ±0.01	-6.58 ±0.01	39.56 ±0.89	-14.28 ±0.85	-16.80 ±0.86	-23.07 ±0.86
P3A	-7.08 ±0.05	-7.08 ±0.05	38.50 ±0.88	-15.86 ±0.85	-16.74 ±0.84	-21.78 ±0.85
PAL	-8.84 ±0.37	-8.84 ±0.37	42.40 ±0.87	-21.65 ±0.85	-14.58 ±0.87	-31.63 ±0.87
THP	-7.31 ±0.24	-7.31 ±0.24	45.66 ±0.88	-13.43 ±0.86	-9.65 ±0.86	-22.25 ±0.87
	Slope	-0.09	0.57	0.84	0.11	0.39
	Intercept	43.01	-9.40	-7.93	-21.57	-13.23
	R^2	0.01	0.38	0.45	0.02	0.12
	Kendall's τ	-0.06	0.36	0.33	0.10	0.22
	RMSE	56.25	6.30	7.86	12.98	8.95
	MSE	3163.95	39.70	61.83	168.44	80.03
	MAE	55.80	5.31	6.75	11.21	7.04

Table S3.8: *Computed binding enthalpies and statistics between the computed and experimental binding enthalpies for Bind3P Water Model*

	ΔH_{exp}	CGenFF	GAFFv1	GAFFv2	Parsley	SwissParam
A01	-19.00 ±0.23	-16.35 ±0.41	-20.74 ±0.43	-21.99 ±0.42	-19.63 ±0.42	-18.38 ±0.41
A02	-19.30 ±0.23	-13.49 ±0.41	-18.61 ±0.43	-21.30 ±0.42	-19.11 ±0.42	-16.42 ±0.41
A03	-21.90 ±0.23	-15.83 ±0.42	-20.33 ±0.43	-22.09 ±0.42	-19.72 ±0.43	-17.38 ±0.41
A04	-20.10 ±0.23	-14.88 ±0.41	-20.34 ±0.43	-21.82 ±0.42	-20.14 ±0.42	-17.17 ±0.41
A05	-19.50 ±0.23	-15.11 ±0.41	-20.02 ±0.42	-21.59 ±0.42	-18.98 ±0.43	-17.35 ±0.40
B02	-15.80 ±0.12	-15.96 ±0.41	-17.42 ±0.43	-17.08 ±0.41	-17.68 ±0.41	-17.77 ±0.41
B05	-15.60 ±0.23	-13.39 ±0.41	-14.49 ±0.43	-16.11 ±0.41	-15.37 ±0.42	-14.20 ±0.41
B11	-16.30 ±0.23	-14.33 ±0.42	-14.53 ±0.44	-16.82 ±0.43	-16.78 ±0.43	-15.85 ±0.41
BER	-9.08 ±0.29	-6.98 ±0.42	-2.42 ±0.43	2.54 ±0.42	-14.04 ±0.42	-3.38 ±0.41
CA0	-9.91 ±0.12	-11.01 ±0.41	-9.99 ±0.43	-14.69 ±0.41	-11.12 ±0.43	-11.07 ±0.41
CA1	-10.82 ±0.06	-11.92 ±0.42	-8.38 ±0.42	-12.06 ±0.42	-12.07 ±0.43	-10.18 ±0.41
CA2	-11.58 ±0.06	-13.71 ±0.41	-8.27 ±0.43	-9.59 ±0.41	-11.83 ±0.42	-11.56 ±0.40
CTE	-13.20 ±2.31	-6.70 ±0.41	-8.32 ±0.42	-4.49 ±0.42	-11.99 ±0.42	-1.24 ±0.41
DHC	-8.12 ±0.14	-10.01 ±0.42	-5.34 ±0.43	-0.86 ±0.41	-13.42 ±0.42	-5.38 ±0.40
FTE	-11.10 ±0.23	-6.41 ±0.42	-7.21 ±0.43	-4.81 ±0.42	-13.01 ±0.43	-5.46 ±0.41
NTE	-15.60 ±1.27	-7.80 ±0.42	-10.56 ±0.43	-4.32 ±0.42	-14.46 ±0.42	-5.71 ±0.41
P01	-3.98 ±0.01	-5.81 ±0.42	-2.92 ±0.42	-2.18 ±0.43	-3.83 ±0.42	-3.67 ±0.41
P02	-6.45 ±0.06	-7.09 ±0.42	-2.90 ±0.43	-1.90 ±0.41	-7.79 ±0.42	-10.74 ±0.40
P03	-5.16 ±0.02	-8.14 ±0.42	-0.86 ±0.44	-0.60 ±0.42	-6.42 ±0.42	-5.03 ±0.40
P04	-4.55 ±0.09	-9.67 ±0.42	-1.47 ±0.43	-1.03 ±0.42	-7.89 ±0.42	-8.26 ±0.41
P1A	-6.98 ±0.02	-8.87 ±0.42	-3.31 ±0.43	-0.46 ±0.41	-4.95 ±0.42	-4.42 ±0.41
P2A	-6.58 ±0.01	-10.22 ±0.43	-2.61 ±0.43	-2.68 ±0.41	-7.35 ±0.42	-6.65 ±0.41
P3A	-7.08 ±0.05	-10.84 ±0.42	-2.26 ±0.43	-1.84 ±0.42	-5.32 ±0.43	-5.55 ±0.41
PAL	-8.84 ±0.37	-11.53 ±0.41	-7.85 ±0.43	-0.17 ±0.42	-15.61 ±0.42	-7.81 ±0.41
THP	-7.31 ±0.24	-9.39 ±0.41	-4.96 ±0.42	-0.45 ±0.42	-12.23 ±0.42	-4.02 ±0.41
	Slope	0.45	1.21	1.42	0.83	0.79
	Intercept	-5.74	4.76	7.91	-3.09	-0.49
	R^2	0.54	0.93	0.79	0.81	0.61
	Kendall's τ	0.54	0.79	0.59	0.71	0.51
	RMSE	3.76	3.15	5.46	2.56	4.08
	MSE	14.17	9.89	29.84	6.56	16.63
	MAE	3.22	2.65	4.40	1.83	2.86

Table S3.9: *Statistics between the computed and experimental binding enthalpies for aliphatic molecules*

CgenFF	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	0.28	0.38	0.41	0.36	0.42	0.38	0.17	0.32
Intercept	-11.10	-10.70	-11.34	-9.20	-12.55	-12.56	46.45	-8.95
R2	0.33	0.63	0.53	0.72	0.58	0.39	0.04	0.60
Kendall's τ	0.45	0.56	0.53	0.75	0.56	0.45	0.24	0.53
RMSE	3.25	2.72	3.25	2.89	4.03	3.89	60.22	3.57
MSE	10.57	7.42	10.54	8.36	16.20	15.11	3625.94	12.72
MAE	3.04	2.52	2.78	2.42	3.12	3.15	60.04	2.98
GAFFv1	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	1.21	1.24	1.18	1.21	1.25	1.21	1.28	1.15
Intercept	3.56	1.62	0.58	2.33	-0.11	-0.64	2.38	3.09
R2	0.93	0.94	0.92	0.93	0.92	0.87	0.68	0.91
Kendall's τ	0.85	0.82	0.78	0.71	0.82	0.82	0.60	0.67
RMSE	1.56	2.71	2.83	1.92	4.60	4.56	4.21	1.66
MSE	2.43	7.32	7.99	3.70	21.13	20.78	17.69	2.74
MAE	1.28	2.46	2.40	1.67	4.25	4.10	3.25	1.37
GAFFv2	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	1.05	1.06	1.01	1.06	1.08	1.09	1.28	0.99
Intercept	-2.87	-4.44	-5.95	-4.05	-7.45	-6.34	-2.13	-1.52
R2	0.83	0.86	0.83	0.91	0.82	0.80	0.64	0.85
Kendall's τ	0.75	0.82	0.67	0.82	0.75	0.64	0.67	0.75
RMSE	4.10	5.69	6.37	5.14	8.98	8.18	7.71	2.15
MSE	16.84	32.34	40.59	26.38	80.56	66.87	59.40	4.62
MAE	3.68	5.44	6.12	4.97	8.75	7.89	6.70	1.75
Parsley	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	0.84	0.80	0.81	0.82	0.94	0.78	0.84	0.82
Intercept	-3.37	-5.92	-7.58	-6.30	-6.92	-8.14	-9.20	-3.12
R2	0.72	0.88	0.79	0.95	0.86	0.73	0.39	0.94
Kendall's τ	0.71	0.71	0.67	0.82	0.82	0.67	0.53	0.78
RMSE	2.27	3.03	4.85	3.50	6.12	4.98	7.69	1.06
MSE	5.15	9.18	23.53	12.28	37.42	24.84	59.11	1.13
MAE	1.79	2.80	4.51	3.35	5.93	4.56	6.52	0.81
SwissParam	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	0.61	0.66	0.65	0.65	0.69	0.66	0.90	0.66
Intercept	-11.82	-11.62	-12.78	-7.03	-14.16	-14.43	-5.69	-4.46
R2	0.60	0.77	0.63	0.82	0.76	0.64	0.53	0.81
Kendall's τ	0.53	0.67	0.64	0.60	0.60	0.53	0.56	0.60
RMSE	6.02	6.33	7.37	2.22	9.29	9.25	5.21	2.13
MSE	36.25	40.04	54.32	4.92	86.33	85.52	27.14	4.56
MAE	5.50	6.03	6.98	1.87	9.09	8.95	4.05	1.70

Table S3.10: *Statistics between the computed and experimental binding enthalpies for aromatic molecules*

CgenFF	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	-0.57	0.31	0.31	0.41	0.31	-0.35	-0.85	0.43
Intercept	-18.56	-11.07	-13.60	-7.86	-13.89	-18.84	37.01	-6.04
R2	0.11	0.08	0.07	0.14	0.04	0.03	0.14	0.16
Kendall's τ	-0.27	0.20	0.20	0.31	-0.02	-0.09	-0.20	0.31
RMSE	8.71	6.76	9.20	4.33	9.56	10.49	49.69	2.85
MSE	75.93	45.69	84.55	18.71	91.38	110.14	2468.70	8.12
MAE	7.97	6.45	8.96	3.97	9.23	9.77	49.49	2.60
GAFFv1	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	0.64	1.05	0.80	0.94	0.79	0.76	0.93	0.73
Intercept	-2.38	0.15	-2.36	1.27	-4.84	-5.68	-8.63	1.59
R2	0.40	0.54	0.25	0.48	0.53	0.39	0.32	0.37
Kendall's τ	0.49	0.67	0.38	0.60	0.56	0.49	0.42	0.38
RMSE	1.37	1.54	2.45	2.26	3.61	4.37	8.40	3.73
MSE	1.87	2.37	6.00	5.13	13.05	19.13	70.48	13.89
MAE	1.18	1.21	2.11	1.98	3.40	4.09	8.12	3.38
GAFFv2	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	-1.79	-0.60	-0.56	-0.50	-1.43	-1.34	-1.01	-0.50
Intercept	-16.19	-8.92	-10.10	-7.39	-17.81	-17.16	-21.23	-4.23
R2	0.36	0.35	0.23	0.27	0.33	0.28	0.26	0.35
Kendall's τ	-0.45	-0.53	-0.35	-0.38	-0.53	-0.45	-0.35	-0.56
RMSE	6.30	3.38	2.97	3.81	5.21	5.21	8.74	6.41
MSE	39.72	11.43	8.81	14.54	27.19	27.14	76.45	41.11
MAE	4.16	2.60	2.36	2.95	4.55	4.51	7.68	5.86
Parsley	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	1.58	2.26	2.21	2.11	2.11	1.93	1.70	1.93
Intercept	-2.56	2.79	-0.10	1.88	-2.00	-3.22	-12.59	4.03
R2	0.63	0.68	0.72	0.70	0.69	0.68	0.54	0.61
Kendall's τ	0.56	0.64	0.64	0.56	0.64	0.64	0.45	0.53
RMSE	6.82	6.48	8.76	6.29	9.91	9.86	17.49	3.63
MSE	46.58	41.97	76.73	39.52	98.21	97.12	305.95	13.19
MAE	6.48	5.68	8.27	5.62	9.50	9.52	17.28	2.96
SwissParam	**opc**	**spc**	**spce**	**tip3p**	**tip4p**	**tip4pew**	**tip5p**	**bind3p**
Slope	-0.41	0.06	-0.04	-0.01	-0.10	-0.20	-0.43	-0.12
Intercept	-17.33	-13.05	-15.70	-8.64	-17.29	-18.89	-20.14	-6.72
R2	0.04	0.00	0.00	0.00	0.00	0.01	0.02	0.01
Kendall's τ	-0.20	-0.09	-0.16	-0.09	-0.05	-0.13	-0.24	-0.16
RMSE	8.78	7.28	9.21	3.20	10.34	11.45	11.84	2.91
MSE	77.06	52.98	84.90	10.27	106.89	131.21	140.25	8.46
MAE	7.79	6.72	8.66	2.40	9.86	10.81	10.50	2.31

Supplementary material for Chapter 4

Table S4.1: *Initial BRD4-1 set having PDB structures with ITC binding data.*

PDB ID	Cluster Size	Cluster ID	K_d (M)	ΔG (kcal/mol)	ITC Condition	Ref.
3MXF	1	2	4.9E-08	-9.64	50 mM HEPES, 150 mM NaCl, and pH 7.4	[412]
3P5O	3	3	9.5E-08	-9.59	20 mM HEPES, 100 mM NaCl, and pH 7.5	[413]
3U5J	3	3	0.00000246	-7.40	50 mM HEPES, 150 mM NaCl, and pH 7.4	[414]
3U5L	3	3	0.00000064	-8.16	50 mM HEPES, 150 mM NaCl, and pH 7.4	[414]
4E96	6	6	4.74E-08	-9.68	50 mM HEPES, 100 mM NaCl, 0.5 mM TCEP and pH 7.5	[415]
4J3I	2	7	1.142E-06	-7.84	50 mM HEPES, 150 mM NaCl, 0.05% DMSO and pH 7.5	[416]
4MR4	2	7	0.00000893	-6.87	50 mM HEPES, 150 mM NaCl, and pH 7.5	[417]
4LYS	2	8	0.000046	-5.90	Not available	[108]
4LYW	11	9	2.37E-07	-9.00	Not available	[108]
4LZR	2	8	0.00002	-6.40	Not available	[108]
4LZS	11	9	0.000016	-6.60	Not available	[108]
4OGI	1	12	3.7E-08	-9.79	50 mM HEPES, 150 mM NaCl, and pH 7.4	[418]
4OGJ	1	13	1.64E-07	-8.94	50 mM HEPES, 150 mM NaCl, and pH 7.4	[418]
4QB3	1	14	0.0000033	-7.55	50 mM Na_3PO_4, 150 mM NaCl, and pH 7.4	[419]
4XY9	2	15	4.651E-06	-7.03	50 mM HEPES, 150 mM NaCl, and pH 7.5	[420]
4XYA	2	15	0.00000137	-7.73	50 mM HEPES, 150 mM NaCl, and pH 7.5	[420]
5BT4	1	17	8.85E-07	-8.14	50 mM HEPES, 150 mM NaCl, and pH 7.5	[421]
5CP5	6	6	1.234E-06	-8.06	50 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[110]
5CQT	6	6	1.24E-07	-9.42	50 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[110]
5CY9	6	6	2.18E-07	-9.09	50 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[110]
5D0C	6	6	0.00000234	-7.68	50 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[110]
5D24	11	9	5.841E-06	-7.14	Not available	[422]
5D25	11	9	5.075E-07	-8.59	Not available	[422]
5D26	11	9	8.096E-07	-8.31	Not available	[422]
5D3H	11	9	0.00000793	-6.96	Not available	[422]
5D3J	11	9	5.733E-06	-7.15	Not available	[422]
5D3L	11	9	8.844E-07	-8.26	Not available	[422]
5D3N	11	9	9.202E-06	-6.87	Not available	[422]
5D3P	11	9	7.391E-06	-7.00	Not available	[422]
5D3S	11	9	1.272E-06	-8.04	Not available	[422]
5D3T	11	9	0.00001206	-6.71	Not available	[422]
5DW2	2	7	6.95E-07	-8.00	50 mM HEPES, 150 mM NaCl, and pH 7.5	[423]
5DX4	6	6	1.37E-07	-9.36	50 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[110]
5EGU	1	1	0.0000014	-8.20	10 mM HEPES, 150 mM NaCl, and pH 7.5	[424]
5FBX	1	33	5.43E-09	-10.90	20 mM HEPES, 150 mM NaCl, 0.5 mM TCEP and pH 7.5	[425]
5IGK	1	34	4.18E-08	-9.73	50 mM HEPES, 150 mM NaCl, and pH 7.5	[426]

Table S4.2: *Initial binding enthalpy calculations*

PDB	$E_{Complex}$	$E_{Receptor}$	E_{Ligand}	ΔH_{MD}	ΔH_{Exp}
3MXF	-194542.80 ± 0.47	-194552.58 ± 0.24	-19104.10 ± 0.09	-9.15 ± 0.54	-8.42
3U5L	-194517.73 ± 0.40	-194552.58 ± 0.24	-19077.60 ± 0.07	-10.57 ± 0.47	-6.16
4LZR	-194515.99 ± 0.64	-194552.58 ± 0.24	-19076.84 ± 0.07	-9.60 ± 0.69	-9.00
4QB3	-194625.18 ± 0.71	-194552.58 ± 0.24	-19185.19 ± 0.07	-10.43 ± 0.75	-6.62
4XY9	-194805.96 ± 1.07	-194552.58 ± 0.24	-19369.69 ± 0.07	-6.71 ± 1.10	-6.09
5D0C	-194662.30 ± 0.64	-194552.58 ± 0.24	-19220.96 ± 0.07	-11.78 ± 0.69	-10.20
5D3S	-194622.99 ± 0.88	-194552.58 ± 0.24	-19184.16 ± 0.07	-9.28 ± 0.92	-9.77
5DW2	-194670.10 ± 0.58	-194552.58 ± 0.24	-19231.58 ± 0.07	-8.96 ± 0.63	-10.10
5FBX	-194559.37 ± 1.09	-194552.58 ± 0.24	-19113.12 ± 0.09	-16.69 ± 1.13	-15.57
5IGK	-194782.42 ± 0.58	-194552.58 ± 0.24	-19337.15 ± 0.07	-15.71 ± 0.64	-11.09

All values are in kcal/mol. Outliers are highlighted in bold. The error in ΔH_{MD} is taken as the maximum value of the error from the blocking analysis (Figure S2).

$\mathbf{E_{Water}}$ = -19123.03 ±0.07

Table S4.3: *Binding enthalpy calculations for 3UL5, 4BQ3 and 5IGK with optimized charges, differing buffer conditions and ZA-loop conformations.*

PDB	$E_{Complex}$	$E_{Receptor}$	E_{Ligand}	ΔH_{MD}	$\Delta H_{Exp.}$
3U5L[a]	-194482.19 ± 0.41	-194552.58 ± 0.24	-19038.78 ± 0.07	-13.85 ± 0.48	-6.16
4QB3[a]	-194507.71 ± 0.68	-194552.58 ± 0.24	-19073.03 ± 0.07	-5.13 ± 0.72	-6.62
5IGK[a]	-194492.98 ± 0.53	-194552.58 ± 0.24	-19049.07 ± 0.07	-14.35 ± 0.59	-11.09
3U5L[b]	-202904.22 ± 1.82	-202979.60 ± 1.85	-19038.78 ± 0.07	-8.86 ± 2.59	-6.16
5IGK[b]	-202920.04 ± 2.97	-202979.60 ± 1.85	-19049.07 ± 0.07	-14.39 ± 3.50	-11.09
3U5L[c]	-194474.86 ± 0.58	-194552.58 ± 0.24	-19038.78 ± 0.07	-6.52 ± 0.64	-6.16
3U5L[d]	-194479.27 ± 3.21	-194552.58 ± 0.24	-19038.78 ± 0.07	-10.93 ± 3.22	-6.16
5IGK[c]	-194485.25 ± 0.60	-194552.58 ± 0.24	-19049.07 ± 0.07	-6.62 ± 0.66	-11.09
5IGK[d]	-194489.92 ± 3.06	-194552.58 ± 0.24	-19049.07 ± 0.07	-11.29 ± 3.07	-11.09

[a] Calculated by using simulations with optimized ligand parameters of 3UL5, 4BQ3, 5IGK
[b] Calculated by simulations with three ionization states of HEPES and NaCl
[c] Simulations were started using receptor with ZA2 conformation for complex simulations
[d] Calculated by using combined simulations data for both ZA1 and ZA2 from 40 simulations
All values are in kcal/mol. $\mathbf{E_{Water}}$ = -19123.03 ±0.07

Table S4.4: *Binding enthalpy calculations from specific ZA loop conformations*

PDB	$E_{Complex}$	ΔH_{MD}	$\Delta H_{Exp.}$
4LZR_ZA1	-194517.48 ± 0.22	-10.96 ± 0.25	-9.00
4QB3_ZA1	-194507.94 ± 0.20	-5.23 ± 0.23	-6.62
4XY9_ZA1	-194806.62 ± 0.23	-7.25 ± 0.25	-6.09
5DW2_ZA1	-194668.91 ± 0.23	-7.66 ± 0.25	-10.10
4LZR_ZA2	-194513.13 ± 0.32	-6.61 ± 0.34	-9.00
4QB3_ZA2	-194505.80 ± 0.45	-3.09 ± 0.46	-6.62
4XY9_ZA2	-194803.31 ± 0.31	-3.93 ± 0.33	-6.09
5DW2_ZA2	-194671.73 ± 0.31	-10.48 ± 0.33	-10.10

All values are in kcal/mol.

Table S4.5: *Percentage occupation of ZA loop states*

System	ZA1 (%)	ZA2 (%)	Intermediate (%)
3MXF	92.70	3.02	4.28
3U5L	99.69	0.00	0.31
3U5L_ion	91.70	6.20	2.10
3U5L_qm	99.33	0.00	0.67
3U5L_ZA	0.00	93.22	6.78
4LZR	64.75	26.94	8.32
4QB3	66.79	28.70	4.51
4QB3_qm	80.89	15.95	3.16
4XY9	65.03	31.49	3.48
5D0C	99.79	0.00	0.21
5D3S	99.97	0.00	0.03
5DW2	58.97	31.98	9.06
5FBX	99.88	0.00	0.12
5IGK	99.97	0.00	0.03
5IGK_ion	99.97	0.00	0.03
5IGK_qm	99.96	0.00	0.04
5IGK_ZA	0.66	88.27	11.06
receptor(20)[a]	19.74	74.53	5.73
receptor(100)[b]	20.98	73.27	5.76
receptor_ion	49.38	42.21	8.41

_ion; contains 50 mM HEPES and 150 mM NaCl, and also optimized ligand parameters
_qm; performed using optimized ligand parameters
_ZA: ZA2 was used as a starting structure
[a] Percentages for the first 20 repeats
[b] Percentages for all 100 repeats

ZA1 state is defined as: If the Val90:H-Val87:O and Lys91:H-Val87:O hydrogen bond distances are more than 4 Å and 5 Å respectively. ZA2 state is defined as: if the Val90:H-Val87:O and Lys91:H-Val87:O hydrogen bond distances are both less than 3 Å. Everything else is defined as intermediate.

Table S4.6: *Physical determinants of Binding Enthalpies*

PDB	Val	Coul	LJ
3MXF	0.38	-5.09	-4.35
3U5L_ZA1	1.63	-12.10	-3.21
3U5L_ZA2	-1.10	0.64	-5.90
4LZR	-1.04	-7.25	-1.31
4QB3	4.95	-6.70	-3.04
4XY9	-0.94	-2.13	-3.04
5D0C	1.68	-9.24	-3.78
5D3S	1.77	-5.81	-4.99
5DW2	0.90	-6.10	-3.28
5FBX	-0.94	-11.85	-3.90
5IGK_ZA1	3.53	-16.00	-1.75
5IGK_ZA2	1.06	-2.20	-5.37

Coul*: Coulombic electrostatic contribution.* ***Val****: contribution from changes in bond-stretch, angle-bend, and dihedral terms.* ***LJ****: Lennard-Jones contribution.*

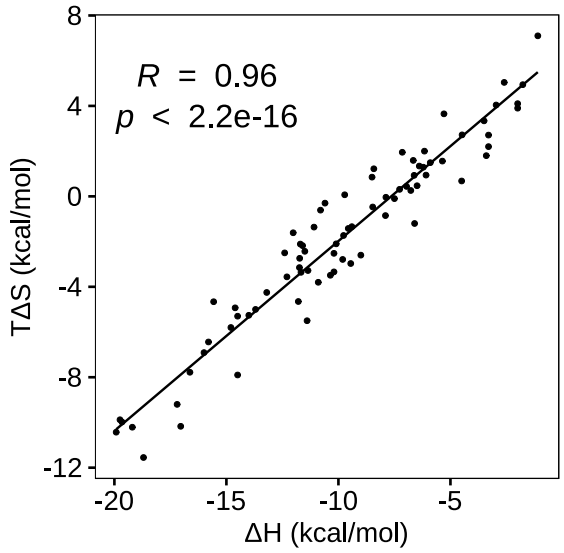

Figure S4.1: *Enthalpy-Entropy Compensation in BRD4-1 calorimetry data.*
The plot indicates a clear correlation between ΔH and TΔS components of the binding free energy.

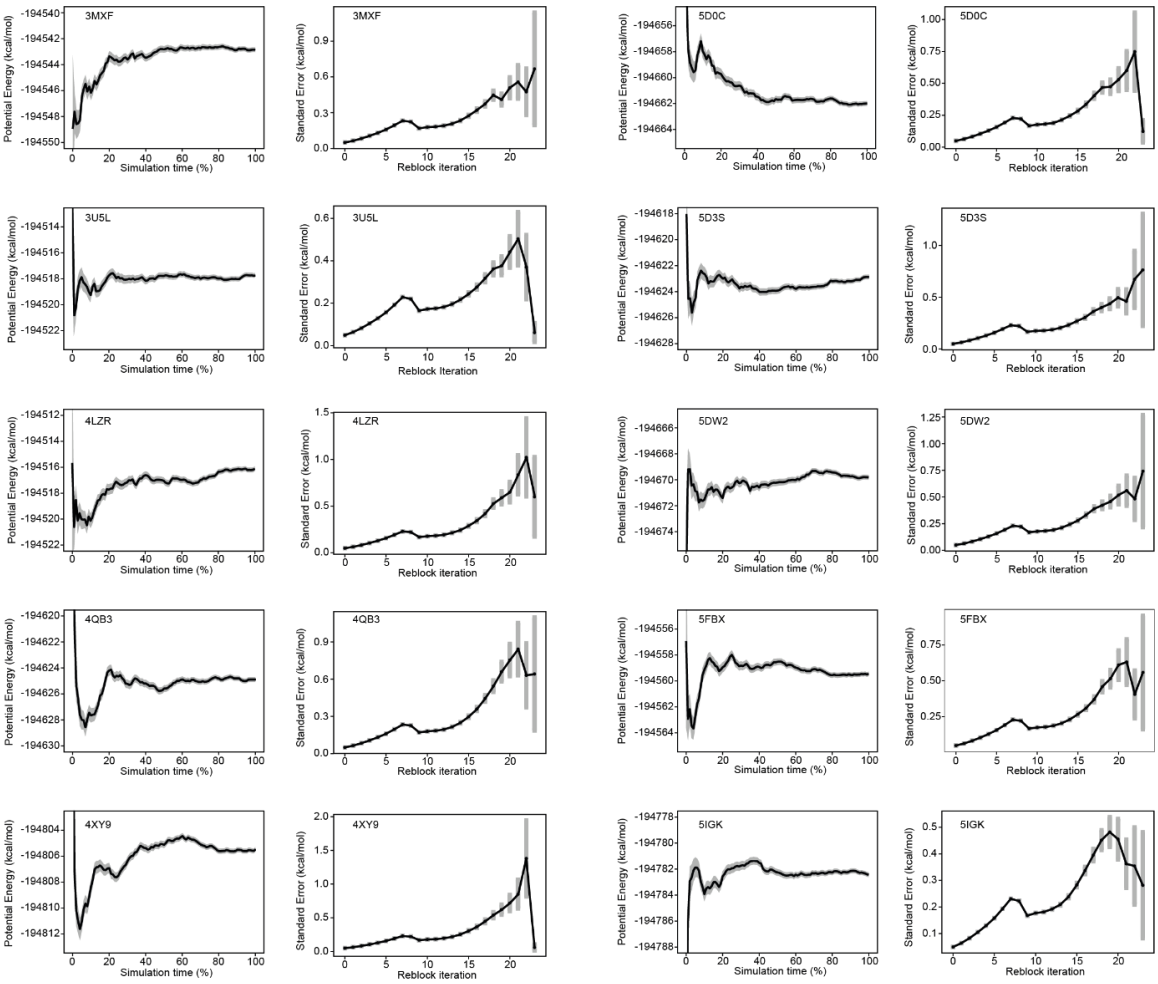

Figure S4.2: *The change in the average potential energy of cumulative simulation data from 20 replicates for all complexes along with the respective blocking analysis curves.*
Grey shade represents the standard error of the mean.

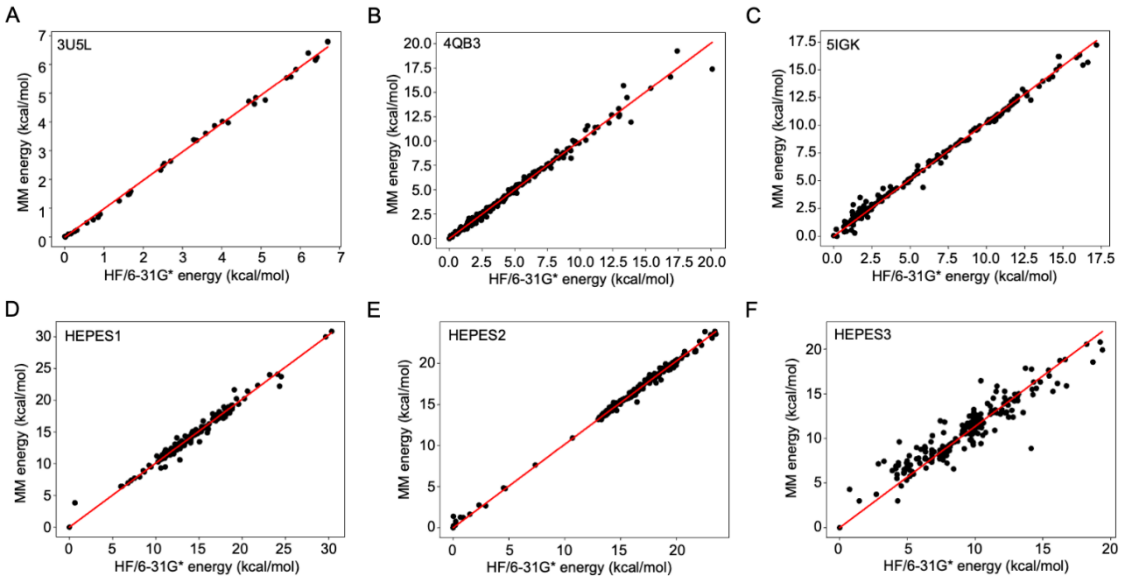

Figure S4.3: *Effects of parameters and buffer.*
(A-C) Comparison of HF/6-31G and MM energies for 3UL5, 4BQ3, 5IGK derived from the parameterize software tool (https://software.acellera.com/docs/latest/parameterize[290] and also (D-F) three ionization states of HEPES (see Figure S4). Points represent different conformations and the energy as computed by the two methods.*

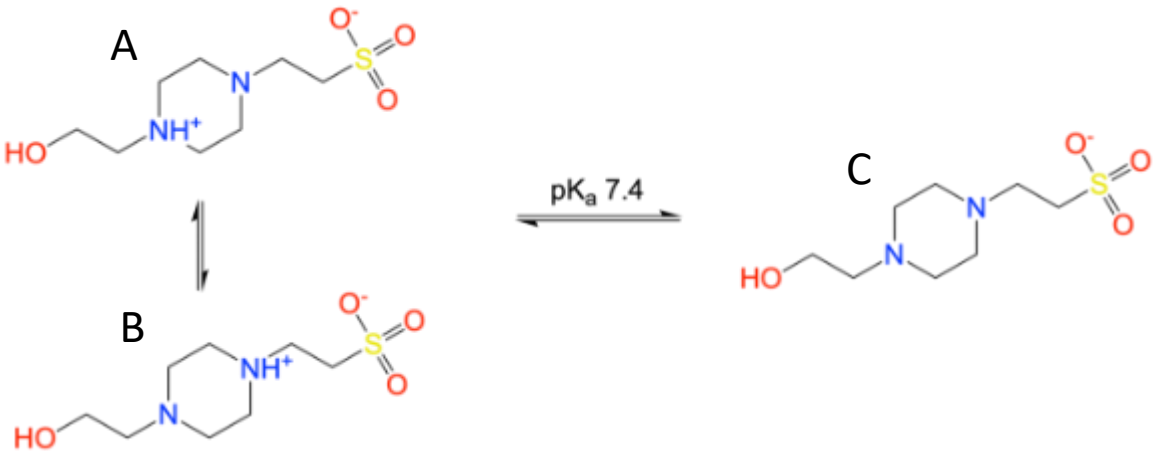

Figure S4.4: *Three different ionization states of HEPES buffer.*

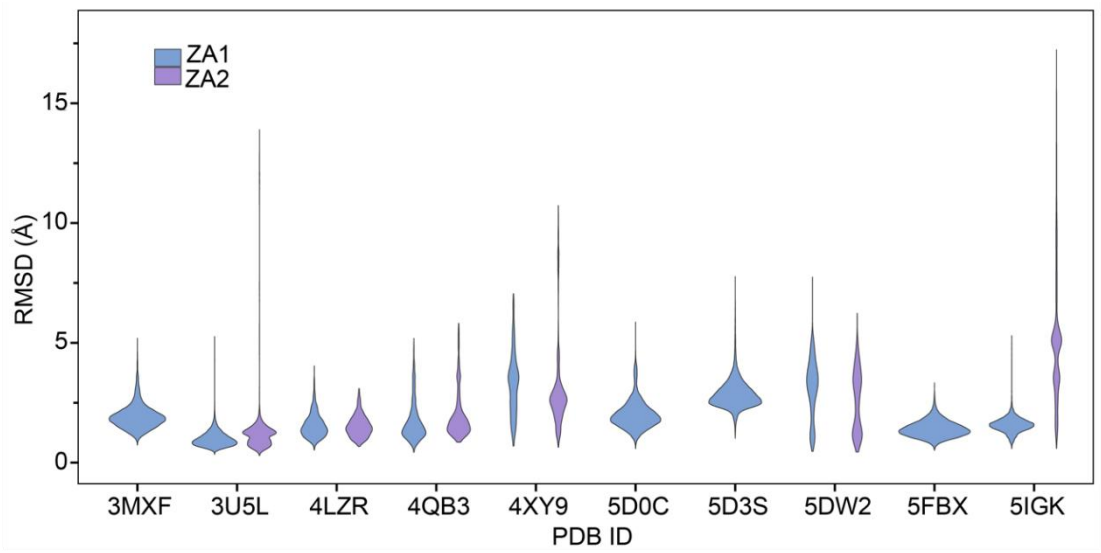

Figure S4.5: *Violin plots for all-atom RMSD of ligands with respect to their starting conformations.*

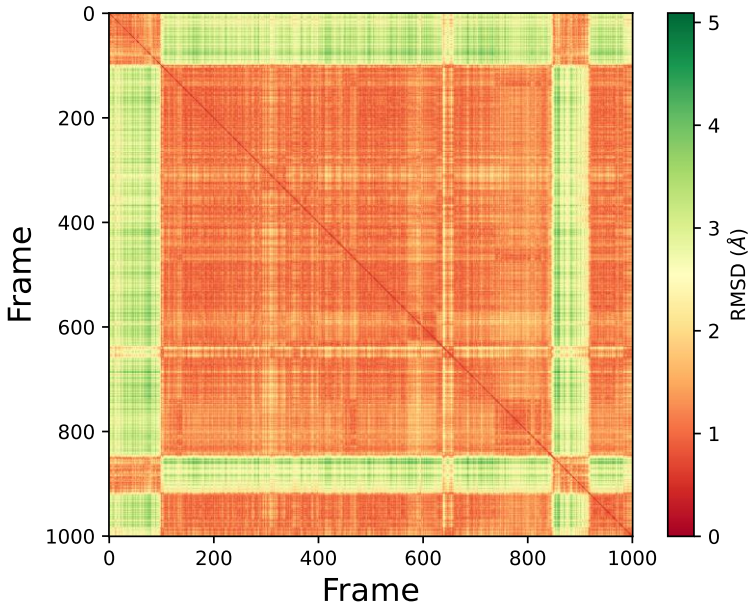

Figure S4.6: *Pairwise RMSD for ZA-loop (76-106 residues) backbone atoms of the simulation in Figure 4.6A of the main manuscript.*

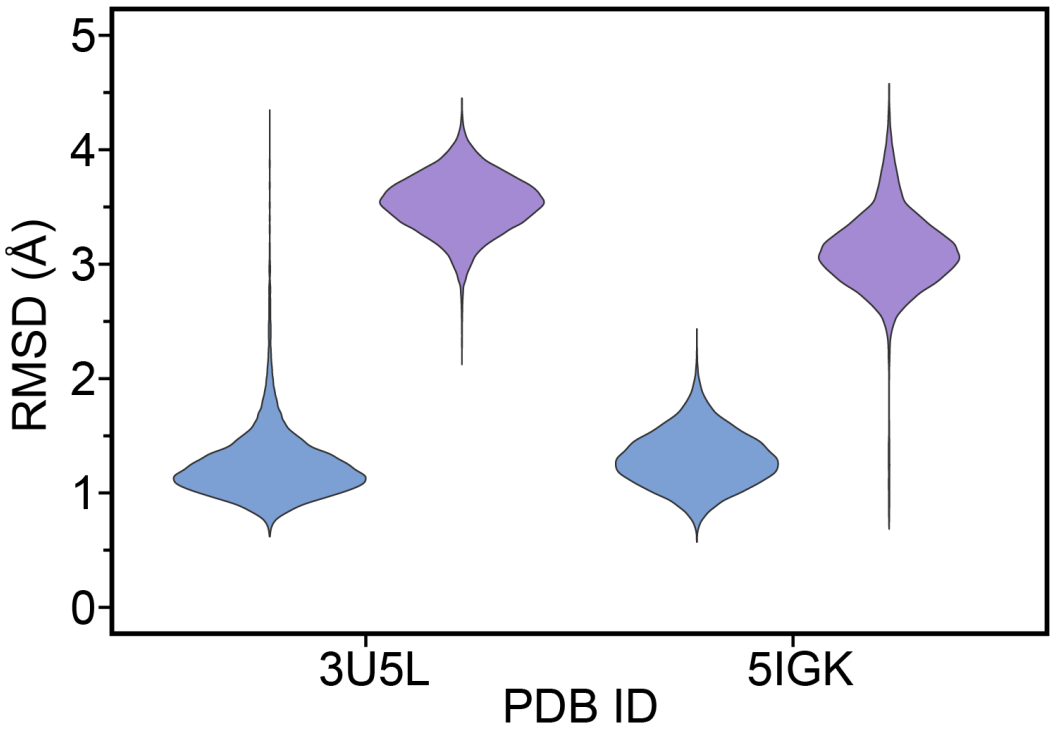

Figure S4.7: *RMSD violin plots for ZA-loop (76-106 residues) backbone atoms of 3U5L and 5IGK for simulations having ZA1 (blue) and ZA2 (purple) in the starting structure*.

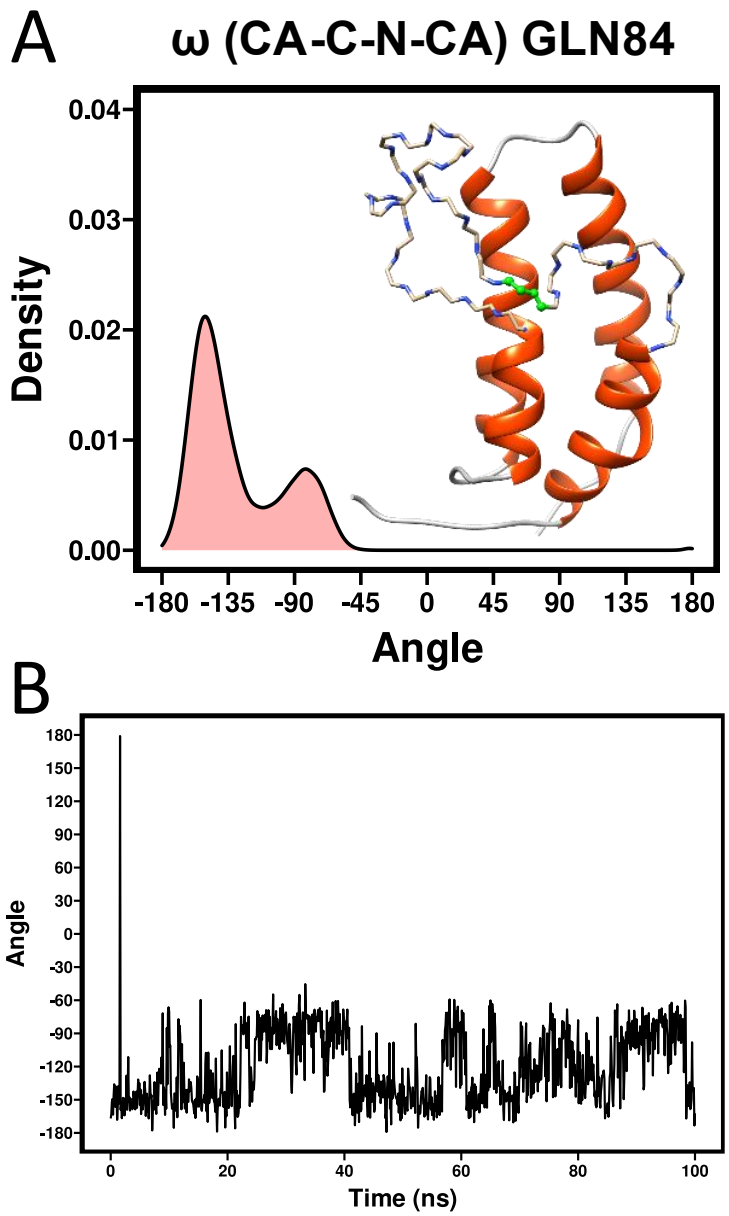

Figure S4.8: *A) The ω (CA-C-N-CA) angle distribution of Gln84 (in the ZA loop) across 100 apo-receptor simulations.*
Atoms are highlighted in green ball & stick in the inset cartoon. B) The angle change of the ω (CA-C-N-CA) Gln84 from the simulation in Figure 6A.

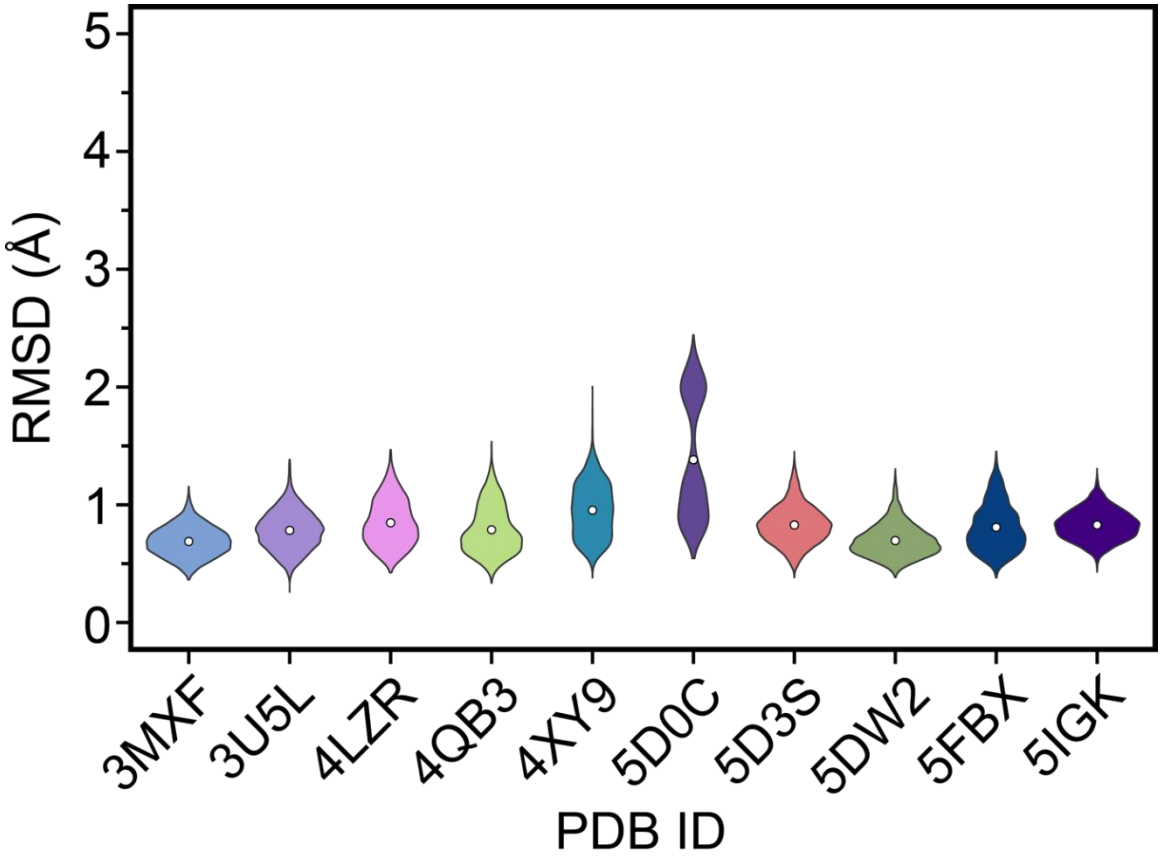

Figure S4.9: *RMSD violin plots for the ZA-loop (76-106 residues) backbone atoms in crystal lattice simulations.*

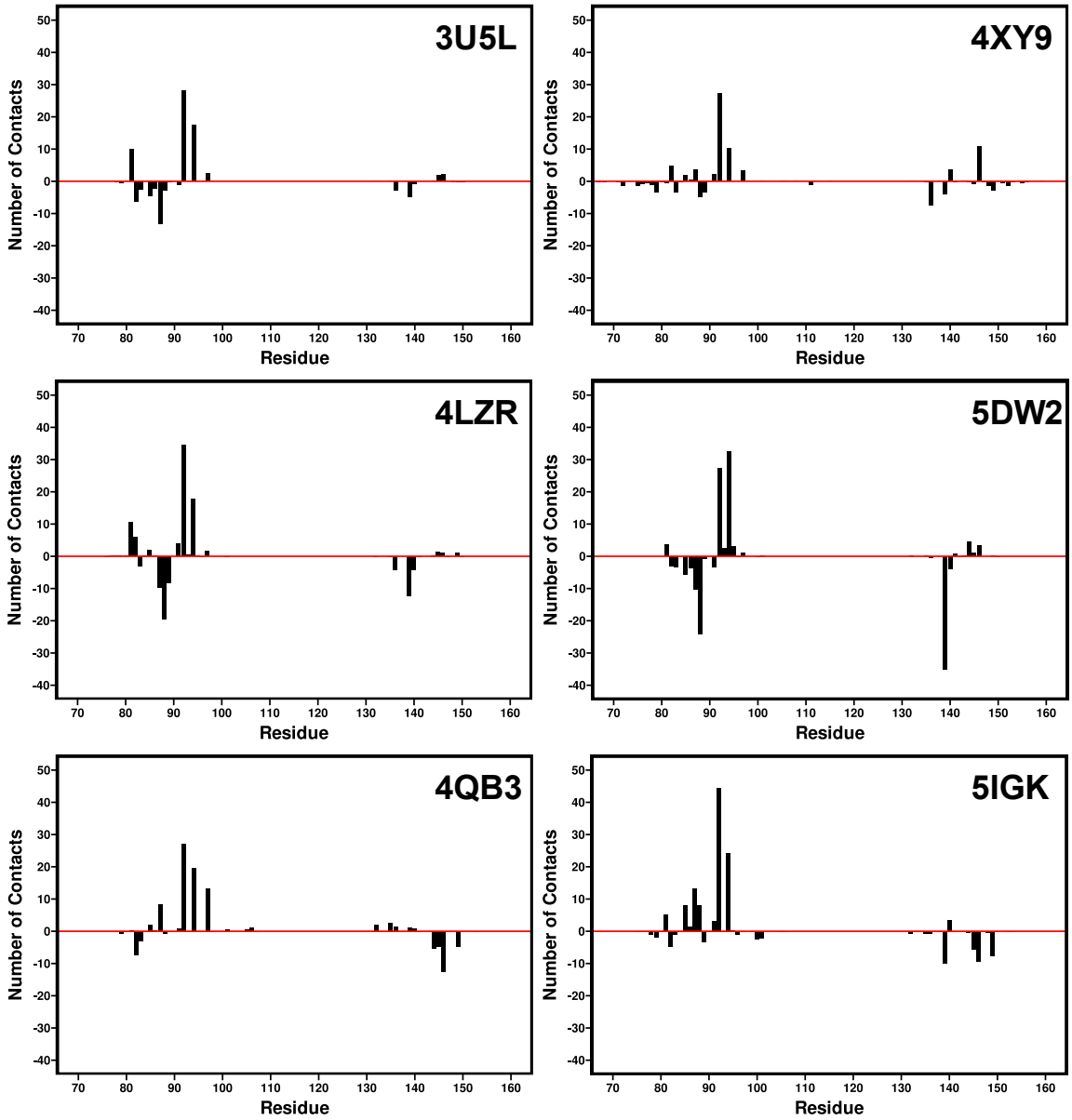

Figure S4.10: *Difference (ZA1 – ZA2) of the average number of contacts between conformations.*
Note that 5DW2 for which the ZA2 loop conformation gives the best agreement with experiment exhibits substantial contact preferences for Asp88 and Tyr139.

Supplementary material for Chapter 5

Table S5.1: *Initial binding enthalpy values for each PDB, and their potential energies for all simulation setups.*

PDB	COMPLEX	RECEPTOR	PEPTIDE	WATER
1DPU	-109091.26 ± 0.67	-107733.26 ± 0.42	-97202.80 ± 0.51	-95854.55 ± 0.22
1RST	-155956.57 ± 0.63	-155575.17 ± 0.43	-48298.92 ± 0.27	-47927.58 ± 0.21
2LQC	-122705.54 ± 1.17	-122100.59 ± 1.04	-96451.86 ± 0.43	-95854.55 ± 0.22
2MNU	-194074.04 ± 0.80	-193324.03 ± 0.73	-96600.76 ± 0.66	-95854.55 ± 0.22
2MWY	-107800.93 ± 0.92	-107200.34 ± 0.33	-48520.13 ± 0.55	-47927.58 ± 0.21
4F14	-100078.30 ± 0.33	-100023.01 ± 0.56	-95901.86 ± 0.22	-95854.55 ± 0.22
4Q6F	-123636.04 ± 0.90	-123001.41 ± 0.99	-48559.70 ± 0.22	-47927.58 ± 0.21
5E0M	-102729.31 ± 0.57	-102119.96 ± 0.50	-96457.28 ± 0.55	-95854.55 ± 0.22
5OVC	-128689.74 ± 0.68	-128196.94 ± 0.34	-48411.65 ± 0.21	-47927.58 ± 0.21
6EVO	-110970.19 ± 0.53	-110750.40 ± 0.58	-48134.57 ± 0.18	-47927.58 ± 0.21
6H8C	-117771.75 ± 1.77	-117017.16 ± 1.44	-96614.80 ± 0.75	-95854.55 ± 0.22

All values are in kcal/mol. The uncertainties were calculated using blocking analysis while potential energy values using ensemble averaging.

Table S5.1: *Initial binding enthalpy values for each PDB, and their potential energies for all simulation setups.*

PDB	ΔH_{EXP}	ΔH_{CAL}
1DPU	−16.80 ± 0.28	-9.75 ± 0.97
1RST	−12.56 ± 0.09	-10.06 ± 0.83
2LQC	−6.91 ± 0.07	-7.63 ± 1.64
2MNU	−4.60 ± 0.10	-3.80 ± 1.29
2MWY	−17.50 ± 0.30	-8.04 ± 1.14
4F14	–8.46 ± 0.48	-7.97 ± 0.72
4Q6F	−9.81 ± 0.04	-2.51 ± 1.37
5E0M	–8.20 ± 0.10	-6.62 ± 0.97
5OVC	−6.80 ± 0.04	-8.73 ± 0.82
6EVO	−8.70 ± 0.80	-12.80 ± 0.83
6H8C	−5.91± 0.09	5.66 ± 2.41

Table S5.2: *Experimental buffer conditions of ITC experiment for each PDB entries*

PDB	ITC Condition	Ref.
1DPU		[427]
1RST	50 mM Potassium Phosphate, pH 7.6	[352]
2LQC	20 mM HEPES, 100 mM KCl, and 5 mM CaCl2, pH 7.0	[353]
2MNU	20 mM Sodium Phosphate, pH 6.0	[354]
2MWY	10 mM Sodium Phosphate, 200 mM NaCl and 1 mM TCEP, pH 6.5	[355]
4F14	Tris 12.5 mM, NaCl 150 mM, pH 8.0	[356]
4Q6F	20 mM HEPES, 150 mM NaCl, 0.5 mM TCEP, pH 8.0	[357]
5E0M	50 mM Sodium Phosphate, 50 mM NaCl, 0.5 mM NaN_3, pH 7.5	[358]
5OVC	50 mM HEPES, 150 mM NaCl, pH 7.0	[359]
6EVO	20 mM TRIS, 50 mM NaCl, and 50 mM Glycine, pH 8.0	[360]
6H8C	50 mM TRIS, 100 mM NaCl, pH 7.5	[361]

Table S5.3: *Binding enthalpy values for additional calculations, and their potential energies for all simulation setups.*

PDB	COMPLEX	RECEPTOR	PEPTIDE	WATER
1DPU_tail	-109096.76 ± 0.44	-107733.26 ± 0.42	-97202.80 ± 0.51	-95854.55 ± 0.22
1DPU_AAA	-108821.20 ± 0.48	-107733.26 ± 0.42	-96931.90 ± 0.70	-95854.55 ± 0.22
2MWY_helix	-107808.23 ± 0.51	-107200.34 ± 0.33	-48520.13 ± 0.55	-47927.58 ± 0.21
4Q6F_zaff	-122372.79 ± 0.30	-121733.00 ± 0.43	-48559.70 ± 0.22	-47927.58 ± 0.21
6EVO_tail	-110970.19 ± 0.53	-110755.24 ± 0.42	-48134.57 ± 0.18	-47927.58 ± 0.21
6H8C_helix	-117784.64 ± 0.72	-117017.16 ± 1.44	-96614.80 ± 0.75	-95854.55 ± 0.22
6EVN	-110791.47 ± 0.46	-110750.40 ± 0.58	-47957.55 ± 0.21	-47927.58 ± 0.21
6EVN_tail	-110791.47 ± 0.46	-110755.24 ± 0.42	-47957.55 ± 0.21	-47927.58 ± 0.21
6EVO_ion	-112488.62 ± 1.09	-112274.59 ± 0.57	-48134.57 ± 0.18	-47927.58 ± 0.21
6EVN_ion	-112311.51 ± 0.67	-112274.59 ± 0.57	-47957.55 ± 0.21	-47927.58 ± 0.21

All values are in kcal/mol.
The uncertainties were calculated using blocking analysis while potential energy values using ensemble averaging.
_ion: 20 mM TRIS (tris(hydroxymethyl)aminomethane), 50 mM NaCl, and 50 mM glycine in the solution
_zaff: Zinc AMBER Force Field (ZAFF) for Zn^{2+}
_tail: considering alternative tail conformations
_helix: considering helix formation in the peptide
_AAA: the peptide has RNK/AAA mutation.

Table S5.3: *Binding enthalpy values for additional calculations, and their potential energies for all simulation setups.*

PDB	ΔH_{EXP}	ΔH_{CAL}
1DPU_tail	−16.80 ± 0.28	-15.25 ± 0.83
1DPU_AAA	-9.84 ± 0.28	-10.58 ± 0.97
2MWY_helix	−17.50 ± 0.30	-15.33 ± 0.84
4Q6F_zaff	−9.81 ± 0.04	-7.67 ± 0.60
6EVO_tail	−8.70 ± 0.80	-7.97 ± 0.73
6H8C_helix	−5.91± 0.09	-7.23 ± 1.78
6EVN	−6.60 ± 0.80	-11.10 ± 0.80
6EVN_tail	−6.60 ± 0.80	-6.27 ± 0.69
6EVO_ion	−8.70 ± 0.80	-7.04 ± 1.26
6EVN_ion	−6.60 ± 0.80	-6.95 ± 0.93

Table S5.4: *The solvent-accessible surface area (SASA) for bound and unbound states. Hydrophobic portions of SASA for bound (HB) and unbound (HU) states. Average number of hydrogen bonds between receptor and peptide (HBonds). HBA: average number of hydrogen bonds per SASA change between bound (HB) and unbound (HU) states. ΔHBond; absolute hydrogen bond difference between bound and unbound states.*

PDB ID	Bound (nm^2)	Unbound (nm^2)	HB (%)	HU (%)	HBonds	HBA	ΔHBond
1DPU	61.70	73.83	31.32	37.54	6.87	0.57	-0.65
1RST	81.35	90.75	44.47	49.58	5.24	0.56	0.40
2LQC	64.27	77.44	32.92	41.42	4.83	0.37	1.22
2MNU	69.79	85.39	36.94	46.70	8.04	0.52	1.51
2MWY	64.91	80.04	32.52	41.41	4.14	0.27	2.82
4F14	52.00	62.81	24.93	31.66	4.67	0.43	0.66
4Q6F	52.52	60.26	24.25	31.31	8.12	1.05	-2.57
5E0M	63.35	73.51	28.90	33.92	14.60	1.44	1.23
5OVC	61.11	71.63	32.09	37.71	13.37	1.27	0.66
6EVO	62.68	73.19	30.27	35.80	8.50	0.81	0.70
6H8C	80.41	99.25	38.67	50.34	11.89	0.63	1.59

All calculated by using simulation sets giving good agreement with experimental for the binding enthalpies.

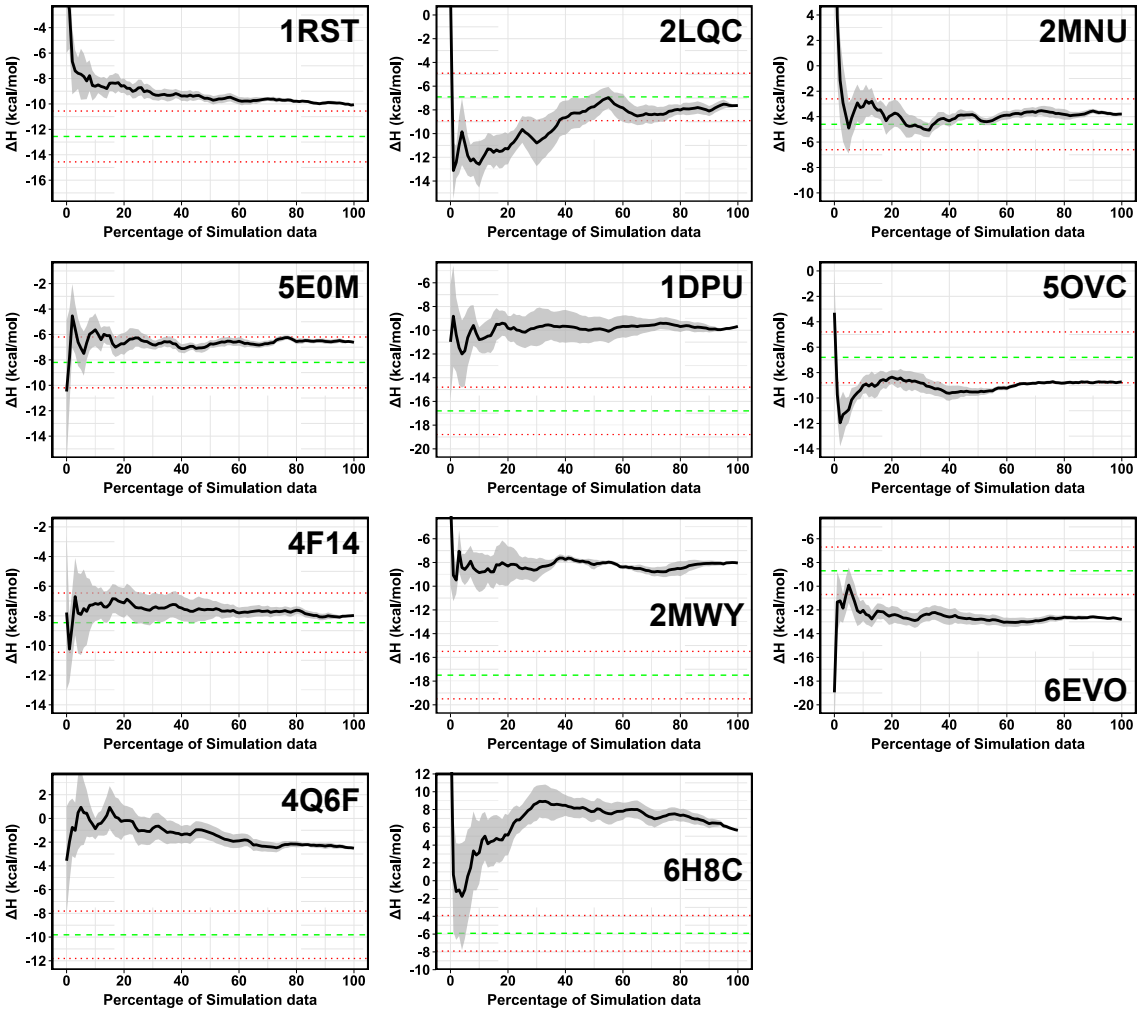

Figure S5.1: *Convergence pattern of the calculated ΔH by using 60 trajectories from the initial simulations of each PDB. Green dashed-line is the experimental ΔH while red dotted-lines indicate the 2 kcal/mol error limit.*

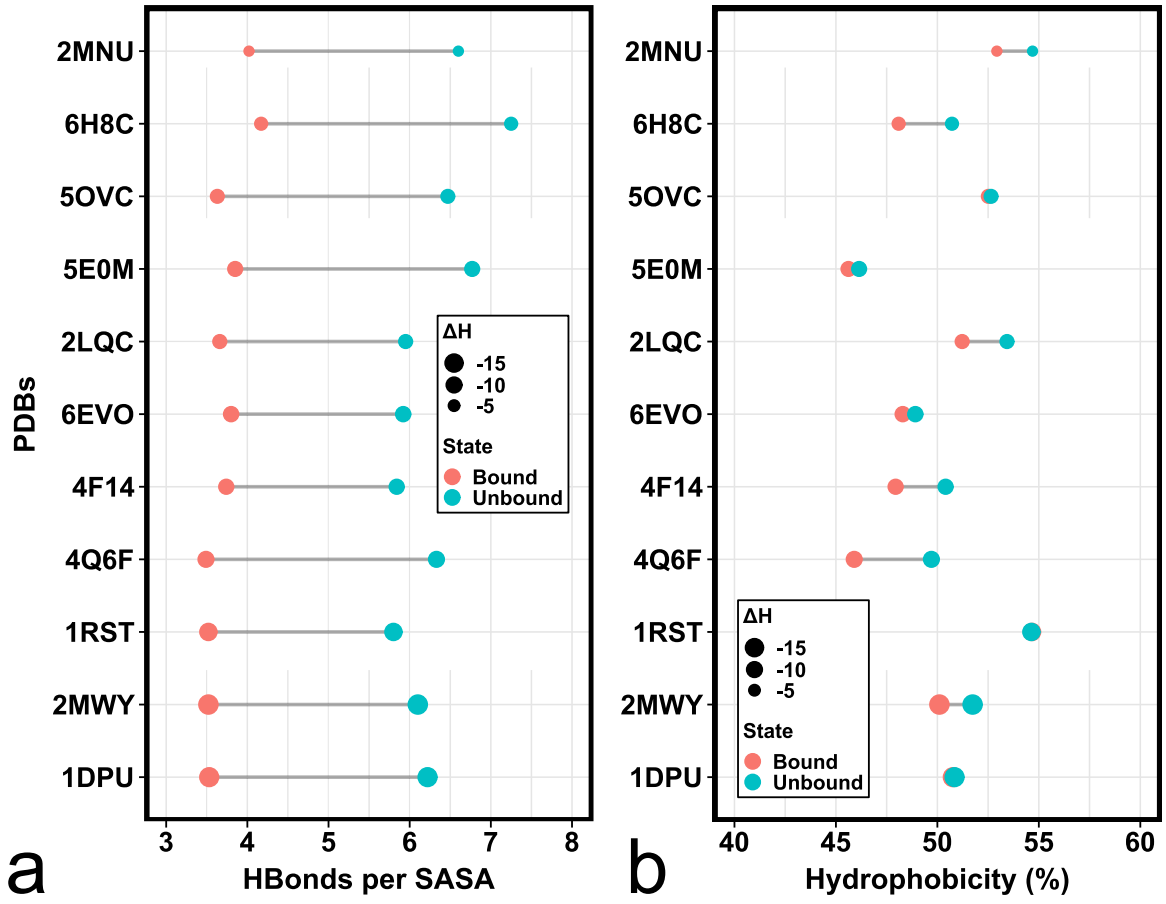

Figure S4: a) *The number hydrogen bonds per SASA between solute and solvent.* ***b)*** *Overall change in hydrophobic area between unbound and bound states. PDB IDs were sorted based on experimental ΔH values.*

References

[1] V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis, and E. Kazlauskas, "Binding affinity in drug design: experimental and computational techniques," *Expert Opinion on Drug Discovery*, vol. 14, no. 8, pp. 755-768, 2019/08/03 2019, doi: 10.1080/17460441.2019.1623202.

[2] N. M. Henriksen, A. T. Fenley, and M. K. Gilson, "Computational Calorimetry: High-Precision Calculation of Host-Guest Binding Thermodynamics," *Journal of Chemical Theory and Computation*, Article vol. 11, no. 9, pp. 4377-4394, Sep 2015, doi: 10.1021/acs.jctc.5b00405.

[3] E. Persch, O. Dumele, and F. Diederich, "Molecular Recognition in Chemical and Biological Systems," *Angewandte Chemie International Edition*, vol. 54, no. 11, pp. 3290-3327, 2015/03/09 2015, doi: 10.1002/anie.201408487.

[4] S.-F. Zhou and W.-Z. Zhong, "Drug Design and Discovery: Principles and Applications," *Molecules*, vol. 22, no. 2, p. 279, 2017.

[5] E. H. Houssein, M. E. Hosney, D. Oliva, N. Ortega-Sánchez, W. M. Mohamed, and M. Hassaballah, "Drug Design and Discovery: Theory, Applications, Open Issues and Challenges," in *Metaheuristics in Machine Learning: Theory and Applications*, D. Oliva, E. H. Houssein, and S. Hinojosa Eds. Cham: Springer International Publishing, 2021, pp. 337-358.

[6] A. Artese *et al.*, "Molecular interaction fields in drug discovery: recent advances and future perspectives," *WIREs Computational Molecular Science*, vol. 3, no. 6, pp. 594-613, 2013/11/01 2013, doi: 10.1002/wcms.1150.

[7] R. Sawada, H. Iwata, S. Mizutani, and Y. Yamanishi, "Target-Based Drug Repositioning Using Large-Scale Chemical–Protein Interactome Data," *Journal of Chemical Information and Modeling*, vol. 55, no. 12, pp. 2717-2730, 2015/12/28 2015, doi: 10.1021/acs.jcim.5b00330.

[8] M. H. S. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, vol. 555, no. 7698, pp. 604-610, 2018/03/01 2018, doi: 10.1038/nature25978.

[9] S. Szymkuć *et al.*, "Computer-Assisted Synthetic Planning: The End of the Beginning," *Angewandte Chemie International Edition*, vol. 55, no. 20, pp. 5904-5937, 2016/05/10 2016, doi: 10.1002/anie.201506101.

[10] N. Berdigaliyev and M. Aljofan, "An overview of drug discovery and development," *Future Medicinal Chemistry*, vol. 12, no. 10, pp. 939-947, 2020/05/01 2020, doi: 10.4155/fmc-2019-0307.

[11] Z. Zhang *et al.*, "Overcoming cancer therapeutic bottleneck by drug repurposing," *Signal Transduction and Targeted Therapy*, vol. 5, no. 1, p. 113, 2020/07/02 2020, doi: 10.1038/s41392-020-00213-8.

[12] X. Lin, X. Li, and X. Lin, "A Review on Applications of Computational Methods in Drug Screening and Design," *Molecules*, vol. 25, no. 6, doi: 10.3390/molecules25061375.

[13] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," *Molecular Diversity*, vol. 25, no. 3, pp. 1315-1360, 2021/08/01 2021, doi: 10.1007/s11030-021-10217-3.

[14] F. D. Prieto-Martínez, E. López-López, K. Eurídice Juárez-Mercado, and J. L. Medina-Franco, "Chapter 2 - Computational Drug Design Methods—Current and Future Perspectives," in *In Silico Drug Design*, K. Roy Ed.: Academic Press, 2019, pp. 19-44.

[15] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten, "Accelerating molecular modeling applications with graphics processors," *Journal of Computational Chemistry,* vol. 28, no. 16, pp. 2618-2640, 2007, doi: 10.1002/jcc.20829.

[16] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, and K. Schulten, "GPU-accelerated molecular modeling coming of age," *Journal of Molecular Graphics and Modelling,* vol. 29, no. 2, pp. 116-125, 2010/09/01/ 2010, doi: 10.1016/j.jmgm.2010.06.010.

[17] E. N. Muratov *et al.,* "A critical overview of computational approaches employed for COVID-19 drug discovery," *Chemical Society Reviews,* 10.1039/D0CS01065K vol. 50, no. 16, pp. 9121-9151, 2021, doi: 10.1039/D0CS01065K.

[18] B. Shaker, S. Ahmad, J. Lee, C. Jung, and D. Na, "In silico methods and tools for drug discovery," *Computers in Biology and Medicine,* vol. 137, p. 104851, 2021/10/01/ 2021, doi: 10.1016/j.compbiomed.2021.104851.

[19] V. T. Sabe *et al.,* "Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review," *European Journal of Medicinal Chemistry,* vol. 224, p. 113705, 2021/11/15/ 2021, doi: 10.1016/j.ejmech.2021.113705.

[20] R. Shukla and T. Tripathi, "Molecular Dynamics Simulation in Drug Discovery: Opportunities and Challenges," in *Innovations and Implementations of Computer Aided Drug Discovery Strategies in Rational Drug Design*, S. K. Singh Ed. Singapore: Springer Singapore, 2021, pp. 295-316.

[21] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, "Molecular Docking and Structure-Based Drug Design Strategies," *Molecules,* vol. 20, no. 7, pp. 13384-13421doi: 10.3390/molecules200713384.

[22] J. Vamathevan *et al.,* "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery,* vol. 18, no. 6, pp. 463-477, 2019/06/01 2019, doi: 10.1038/s41573-019-0024-5.

[23] F. Fogolari, A. Corazza, and G. Esposito, "Free Energy, Enthalpy and Entropy from Implicit Solvent End-Point Simulations," *Front Mol Biosci,* vol. 5, pp. 11-11, 2018 2018, doi: 10.3389/fmolb.2018.00011.

[24] Z. Sun, Y. N. Yan, M. Yang, and J. Z. H. Zhang, "Interaction entropy for protein-protein binding," *J. Chem. Phys.,* Article vol. 146, no. 12, Mar 28 2017, Art no. 124124, doi: 10.1063/1.4978893.

[25] A. Pan, T. Kar, A. K. Rakshit, and S. P. Moulik, "Enthalpy-Entropy Compensation (EEC) Effect: Decisive Role of Free Energy," *J. Phys. Chem. B,* Article vol. 120, no. 40, pp. 10531-10539, Oct 13 2016, doi: 10.1021/acs.jpcb.6b05890.

[26] J. D. Chodera and D. L. Mobley, "Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design," in *Annual Review of Biophysics, Vol 42*, vol. 42, K. A. Dill Ed., (Annual Review of Biophysics, 2013, pp. 121-142.

[27] C. Schonbeck and R. Holm, "Exploring the Origins of Enthalpy-Entropy Compensation by Calorimetric Studies of Cyclodextrin Complexes," *J. Phys. Chem. B,* Article vol. 123, no. 31, pp. 6686-6693, Aug 8 2019, doi: 10.1021/acs.jpcb.9b03393.

[28] S. Geschwindner, J. Ulander, and P. Johansson, "Ligand Binding Thermodynamics in Drug Discovery: Still a Hot Tip?," *Journal of Medicinal Chemistry,* Article vol. 58, no. 16, pp. 6321-6335, Aug 27 2015, doi: 10.1021/jm501511f.

[29] M. S. Friedrichs *et al.*, "Accelerating molecular dynamic simulation on graphics processing units," *Journal of Computational Chemistry,* vol. 30, no. 6, pp. 864-872, 2009, doi: 10.1002/jcc.21209.
[30] M. J. Abraham *et al.*, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX,* vol. 1-2, pp. 19-25, 2015/09/01/ 2015, doi: 10.1016/j.softx.2015.06.001.
[31] A. T. Fenley, N. M. Henriksen, H. S. Muddana, and M. K. Gilson, "Bridging Calorimetry and Simulation through Precise Calculations of Cucurbituril-Guest Binding Enthalpies," *Journal of Chemical Theory and Computation,* Article vol. 10, no. 9, pp. 4069-4078, Sep 2014, doi: 10.1021/ct5004109.
[32] C. H. Reynolds and M. K. Holloway, "Thermodynamics of Ligand Binding and Efficiency," *ACS Medicinal Chemistry Letters,* vol. 2, no. 6, pp. 433-437, 2011/06/09 2011, doi: 10.1021/ml200010k.
[33] J. E. DeLorbe *et al.*, "Thermodynamic and Structural Effects of Conformational Constraints in Protein-Ligand Interactions. Entropic Paradoxy Associated with Ligand Preorganization," *Journal of the American Chemical Society,* Review vol. 131, no. 46, pp. 16758-16770, Nov 25 2009, doi: 10.1021/ja904698q.
[34] A. Roy, D. P. Hua, J. M. Ward, and C. B. Post, "Relative binding enthalpies from molecular dynamics simulations using a direct method," *Journal of chemical theory and computation,* vol. 10, no. 7, pp. 2759-2768, 2014.
[35] A. Li and M. K. Gilson, "Protein-ligand binding enthalpies from near-millisecond simulations: Analysis of a preorganization paradox," *J. Chem. Phys.,* Article vol. 149, no. 7, Aug 21 2018, Art no. 072311, doi: 10.1063/1.5027439.
[36] J. M. Fox, M. Zhao, M. J. Fink, K. Kang, and G. M. Whitesides, "The Molecular Origin of Enthalpy/Entropy Compensation in Biomolecular Recognition," in *Annual Review of Biophysics, Vol 47*, vol. 47, K. A. Dill Ed., (Annual Review of Biophysics, 2018, pp. 223-250.
[37] D. A. Winkler, "Ligand Entropy Is Hard but Should Not Be Ignored," *Journal of Chemical Information and Modeling,* vol. 60, no. 10, pp. 4421-4423, 2020/10/26 2020, doi: 10.1021/acs.jcim.0c01146.
[38] G. G. Ferenczy and G. M. Keserű, "Thermodynamics guided lead discovery and optimization," *Drug Discovery Today,* vol. 15, no. 21, pp. 919-932, 2010/11/01/ 2010, doi: 10.1016/j.drudis.2010.08.013.
[39] A. Umuhire Juru, N. N. Patwardhan, and A. E. Hargrove, "Understanding the Contributions of Conformational Changes, Thermodynamics, and Kinetics of RNA–Small Molecule Interactions," *ACS Chemical Biology,* vol. 14, no. 5, pp. 824-838, 2019/05/17 2019, doi: 10.1021/acschembio.8b00945.
[40] J.-P. Renaud *et al.*, "Biophysics in drug discovery: impact, challenges and opportunities," *Nature Reviews Drug Discovery,* vol. 15, no. 10, pp. 679-698, 2016/10/01 2016, doi: 10.1038/nrd.2016.123.
[41] M. K. Gilson and H.-X. Zhou, "Calculation of Protein-Ligand Binding Affinities," *Annual Review of Biophysics and Biomolecular Structure,* vol. 36, no. 1, pp. 21-42, 2007/06/01 2007, doi: 10.1146/annurev.biophys.36.040306.132550.
[42] J. B. Chaires, "Possible origin of differences between van't Hoff and calorimetric enthalpy estimates," *Biophysical Chemistry,* vol. 64, no. 1, pp. 15-23, 1997/02/28/ 1997, doi: 10.1016/S0301-4622(96)02205-3.
[43] S. A. Kantonen, N. M. Henriksen, and M. K. Gilson, "Accounting for apparent deviations between calorimetric and van't Hoff enthalpies," *Biochimica et Biophysica Acta (BBA) - General Subjects,* vol. 1862, no. 3, pp. 692-704, 2018/03/01/ 2018, doi: 10.1016/j.bbagen.2017.11.020.

[44] C. D. Wick, J. I. Siepmann, and M. R. Schure, "Temperature Dependence of Transfer Properties: Importance of Heat Capacity Effects," *The Journal of Physical Chemistry B,* vol. 107, no. 38, pp. 10623-10627, 2003/09/01 2003, doi: 10.1021/jp0223556.

[45] E. Wang *et al.*, "End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design," *Chem. Rev.,* vol. 119, no. 16, pp. 9478-9508, 2019/08/28 2019, doi: 10.1021/acs.chemrev.9b00055.

[46] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities," *Expert Opinion on Drug Discovery,* vol. 10, no. 5, pp. 449-461, 2015/05/04 2015, doi: 10.1517/17460441.2015.1032936.

[47] H. Liu and T. Hou, "CaFE: a tool for binding affinity prediction using end-point free energy methods," *Bioinformatics,* vol. 32, no. 14, pp. 2216-2218, 2016, doi: 10.1093/bioinformatics/btw215.

[48] H. Sun *et al.*, "Assessing the performance of MM/PBSA and MM/GBSA methods. 7. Entropy effects on the performance of end-point binding free energy calculation approaches," *Physical Chemistry Chemical Physics,* 10.1039/C7CP07623A vol. 20, no. 21, pp. 14450-14460, 2018, doi: 10.1039/C7CP07623A.

[49] W. M. Menzer, C. Li, W. Sun, B. Xie, and D. D. L. Minh, "Simple Entropy Terms for End-Point Binding Free Energy Calculations," *Journal of Chemical Theory and Computation,* vol. 14, no. 11, pp. 6035-6049, 2018/11/13 2018, doi: 10.1021/acs.jctc.8b00418.

[50] M. Almlöf, J. Carlsson, and J. Åqvist, "Improving the Accuracy of the Linear Interaction Energy Method for Solvation Free Energies," *Journal of Chemical Theory and Computation,* vol. 3, no. 6, pp. 2162-2175, 2007/11/01 2007, doi: 10.1021/ct700106b.

[51] T. Hansson, J. Marelius, and J. Åqvist, "Ligand binding affinity prediction by linear interaction energy methods," *J. Comput.-Aided Mol. Des.,* vol. 12, no. 1, pp. 27-35, 1998/01/01 1998, doi: 10.1023/A:1007930623000.

[52] H. Gutiérrez-de-Terán and J. Åqvist, "Linear Interaction Energy: Method and Applications in Drug Design," in *Computational Drug Discovery and Design*, R. Baron Ed. New York, NY: Springer New York, 2012, pp. 305-323.

[53] K. Gao, J. Yin, N. M. Henriksen, A. T. Fenley, and M. K. Gilson, "Binding Enthalpy Calculations for a Neutral Host-Guest Pair Yield Widely Divergent Salt Effects across Water Models," *Journal of Chemical Theory and Computation,* Article vol. 11, no. 10, pp. 4555-4564, Oct 2015, doi: 10.1021/acs.jctc.5b00676.

[54] N. M. Henriksen and M. K. Gilson, "Evaluating Force Field Performance in Thermodynamic Calculations of Cyclodextrin Host-Guest Binding: Water Models, Partial Charges, and Host Force Field Parameters," *Journal of Chemical Theory and Computation,* Article vol. 13, no. 9, pp. 4253-4269, Sep 2017, doi: 10.1021/acs.jctc.7b00359.

[55] J. Yin *et al.*, "Overview of the SAMPL5 host-guest challenge: Are we doing better?," *J. Comput.-Aided Mol. Des.,* Article vol. 31, no. 1, pp. 1-19, Jan 2017, doi: 10.1007/s10822-016-9974-4.

[56] J. Yin, N. M. Henriksen, D. R. Slochower, and M. K. Gilson, "The SAMPL5 host–guest challenge: computing binding free energies and enthalpies from explicit solvent simulations by the attach-pull-release (APR) method," *J. Comput.-Aided Mol. Des.,* vol. 31, no. 1, pp. 133-145, 2017/01/01 2017, doi: 10.1007/s10822-016-9970-8.

[57] J. Yin, N. M. Henriksen, H. S. Muddana, and M. K. Gilson, "Bind3P: Optimization of a Water Model Based on Host-Guest Binding Data," *Journal of Chemical*

Theory and Computation, Article vol. 14, no. 7, pp. 3621-3632, Jul 2018, doi: 10.1021/acs.jctc.8b00318.

[58] S. S. Cinaroglu and P. C. Biggin, "Evaluating the Performance of Water Models with Host-Guest Force Fields in Binding Enthalpy Calculations for Cucurbit 7 uril-Guest Systems," *J. Phys. Chem. B,* Article vol. 125, no. 6, pp. 1558-1567, Feb 18 2021, doi: 10.1021/acs.jpcb.0c11383.

[59] J.-M. Lehn, "Supramolecular Chemistry—Scope and Perspectives Molecules, Supermolecules, and Molecular Devices (Nobel Lecture)," *Angewandte Chemie International Edition in English,* vol. 27, no. 1, pp. 89-112, 1988/01/01 1988, doi: 10.1002/anie.198800891.

[60] X. Ma and Y. Zhao, "Biomedical Applications of Supramolecular Systems Based on Host–Guest Interactions," *Chem. Rev.,* vol. 115, no. 15, pp. 7794-7839, 2015/08/12 2015, doi: 10.1021/cr500392w.

[61] L. You, D. Zha, and E. V. Anslyn, "Recent Advances in Supramolecular Analytical Chemistry Using Optical Sensing," *Chem. Rev.,* vol. 115, no. 15, pp. 7840-7892, 2015/08/12 2015, doi: 10.1021/cr5005524.

[62] G. Yu, K. Jie, and F. Huang, "Supramolecular Amphiphiles Based on Host–Guest Molecular Recognition Motifs," *Chem. Rev.,* vol. 115, no. 15, pp. 7240-7303, 2015/08/12 2015, doi: 10.1021/cr5005315.

[63] D.-H. Qu, Q.-C. Wang, Q.-W. Zhang, X. Ma, and H. Tian, "Photoresponsive Host–Guest Functional Systems," *Chem. Rev.,* vol. 115, no. 15, pp. 7543-7588, 2015/08/12 2015, doi: 10.1021/cr5006342.

[64] J. W. Steed, "First- and second-sphere coordination chemistry of alkali metal crown ether complexes," *Coordination Chemistry Reviews,* vol. 215, no. 1, pp. 171-221, 2001/05/01/ 2001, doi: 10.1016/S0010-8545(01)00317-4.

[65] M. Erdos, R. Hartkamp, T. J. H. Vlugt, and O. A. Moultos, "Inclusion Complexation of Organic Micropollutants with beta-Cyclodextrin," *J. Phys. Chem. B,* Article vol. 124, no. 7, pp. 1218-1228, Feb 20 2020, doi: 10.1021/acs.jpcb.9b10122.

[66] S. K. Albert, H. V. P. Thelu, M. Golla, N. Krishnan, and R. Varghese, "Modular synthesis of supramolecular DNA amphiphiles through host–guest interactions and their self-assembly into DNA-decorated nanovesicles," *Nanoscale,* 10.1039/C6NR08370F vol. 9, no. 17, pp. 5425-5432, 2017, doi: 10.1039/C6NR08370F.

[67] A. S. Braegelman and M. J. Webber, "Integrating Stimuli-Responsive Properties in Host-Guest Supramolecular Drug Delivery Systems," (in eng), *Theranostics,* vol. 9, no. 11, pp. 3017-3040, 2019, doi: 10.7150/thno.31913.

[68] M. P. van der Helm, G. Li, M. Hartono, and R. Eelkema, "Transient Host–Guest Complexation To Control Catalytic Activity," *Journal of the American Chemical Society,* vol. 144, no. 21, pp. 9465-9471, 2022/06/01 2022, doi: 10.1021/jacs.2c02695.

[69] R. Kuhn, F. Stoecklin, and F. Erni, "Chiral separations by host-guest complexation with cyclodextrin and crown ether in capillary zone electrophoresis," *Chromatographia,* vol. 33, no. 1, pp. 32-36, 1992/01/01 1992, doi: 10.1007/BF02276847.

[70] J. An *et al.*, "Purification of protein therapeutics via high-affinity supramolecular host–guest interactions," *Nature Biomedical Engineering,* vol. 4, no. 11, pp. 1044-1052, 2020/11/01 2020, doi: 10.1038/s41551-020-0589-7.

[71] H. J. Kim, M. H. Lee, L. Mutihac, J. Vicens, and J. S. Kim, "Host–guest sensing by calixarenes on the surfaces," *Chemical Society Reviews*, 10.1039/C1CS15169J vol. 41, no. 3, pp. 1173-1190, 2012, doi: 10.1039/C1CS15169J.
[72] R. Gupta and B. Rai, "Molecular dynamics simulation study of translocation of fullerene C60 through skin bilayer: effect of concentration on barrier properties," *Nanoscale*, 10.1039/C6NR09186E vol. 9, no. 12, pp. 4114-4127, 2017, doi: 10.1039/C6NR09186E.
[73] J. J. Karnes and I. Benjamin, "Structure and Dynamics of Host/Guest Complexation at the Liquid/Liquid Interface: Implications for Inverse Phase Transfer Catalysis," *The Journal of Physical Chemistry C*, vol. 121, no. 9, pp. 4999-5011, 2017/03/09 2017, doi: 10.1021/acs.jpcc.6b11715.
[74] S. B. Khan and S.-L. Lee, "Supramolecular Chemistry: Host–Guest Molecular Complexes," *Molecules*, vol. 26, no. 13, doi: 10.3390/molecules26133995.
[75] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," (in eng), *Proc Natl Acad Sci U S A*, vol. 115, no. 21, pp. E4758-e4766, May 22 2018, doi: 10.1073/pnas.1800690115.
[76] D. L. Mobley and M. K. Gilson, "Predicting Binding Free Energies: Frontiers and Benchmarks," in *Annual Review of Biophysics, Vol 46*, vol. 46, K. A. Dill Ed., (Annual Review of Biophysics, 2017, pp. 531-558.
[77] A. Nicholls *et al.*, "Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry," *Journal of Medicinal Chemistry*, vol. 51, no. 4, pp. 769-779, 2008/02/01 2008, doi: 10.1021/jm070549+.
[78] D. L. Mobley *et al.*, "Blind prediction of HIV integrase binding from the SAMPL4 challenge," *J. Comput.-Aided Mol. Des.*, vol. 28, no. 4, pp. 327-345, 2014/04/01 2014, doi: 10.1007/s10822-014-9723-5.
[79] H. Grosjean *et al.*, "SAMPL7 protein-ligand challenge: A community-wide evaluation of computational methods against fragment screening and pose-prediction," *J. Comput.-Aided Mol. Des.*, vol. 36, no. 4, pp. 291-311, 2022/04/01 2022, doi: 10.1007/s10822-022-00452-7.
[80] H. S. Muddana *et al.*, "Blind prediction of host–guest binding affinities: a new SAMPL3 challenge," *J. Comput.-Aided Mol. Des.*, vol. 26, no. 5, pp. 475-487, 2012/05/01 2012, doi: 10.1007/s10822-012-9554-1.
[81] A. Rizzi *et al.*, "Overview of the SAMPL6 host–guest binding affinity prediction challenge," *J. Comput.-Aided Mol. Des.*, vol. 32, no. 10, pp. 937-963, 2018/10/01 2018, doi: 10.1007/s10822-018-0170-6.
[82] M. Amezcua, J. Setiadi, Y. Ge, and D. L. Mobley, "An overview of the SAMPL8 host–guest binding challenge," *J. Comput.-Aided Mol. Des.*, vol. 36, no. 10, pp. 707-734, 2022/10/01 2022, doi: 10.1007/s10822-022-00462-5.
[83] H. S. Muddana, A. T. Fenley, D. L. Mobley, and M. K. Gilson, "The SAMPL4 host–guest blind prediction challenge: an overview," *J. Comput.-Aided Mol. Des.*, vol. 28, no. 4, pp. 305-317, 2014/04/01 2014, doi: 10.1007/s10822-014-9735-1.
[84] M. Amezcua, L. El Khoury, and D. L. Mobley, "SAMPL7 Host–Guest Challenge Overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations," *J. Comput.-Aided Mol. Des.*, vol. 35, no. 1, pp. 1-35, 2021/01/01 2021, doi: 10.1007/s10822-020-00363-5.
[85] N. Zaware and M.-M. Zhou, "Bromodomain biology and drug discovery," *Nature Structural & Molecular Biology*, vol. 26, no. 10, pp. 870-879, 2019/10/01 2019, doi: 10.1038/s41594-019-0309-8.

[86] M. Pervaiz, P. Mishra, and S. Günther, "Bromodomain Drug Discovery – the Past, the Present, and the Future," *The Chemical Record*, vol. 18, no. 12, pp. 1808-1817, 2018, doi: 10.1002/tcr.201800074.
[87] J. Shi and Christopher R. Vakoc, "The Mechanisms behind the Therapeutic Activity of BET Bromodomain Inhibition," *Molecular Cell*, vol. 54, no. 5, pp. 728-736, 2014, doi: 10.1016/j.molcel.2014.05.016.
[88] N.-H. Guo, J.-F. Zheng, F.-M. Zi, and J. Cheng, "I-BET151 suppresses osteoclast formation and inflammatory cytokines secretion by targetting BRD4 in multiple myeloma," *Bioscience Reports*, vol. 39, no. 5, 2019, doi: 10.1042/bsr20181245.
[89] J.-F. Zheng, N.-H. Guo, F.-M. Zi, and J. Cheng, "Long Noncoding RNA H19 Promotes Tumorigenesis of Multiple Myeloma by Activating BRD4 Signaling by Targeting MicroRNA 152-3p," *Molecular and Cellular Biology*, vol. 40, no. 3, pp. e00382-19, 2020/01/16 2020, doi: 10.1128/MCB.00382-19.
[90] J. Zuber *et al.*, "RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia," *Nature*, vol. 478, no. 7370, pp. 524-528, 2011/10/01 2011, doi: 10.1038/nature10334.
[91] K. P. Eagen and C. A. French, "Supercharging BRD4 with NUT in carcinoma," *Oncogene*, vol. 40, no. 8, pp. 1396-1408, 2021/02/01 2021, doi: 10.1038/s41388-020-01625-0.
[92] V. V. Singh and S. Alauddin, "Review on: BRD4 inhibitors for anticancer research," *Human Gene*, vol. 37, p. 201196, 2023/09/01/ 2023, doi: 10.1016/j.humgen.2023.201196.
[93] A. Dey *et al.*, "BRD4 directs hematopoietic stem cell development and modulates macrophage inflammatory responses," *The EMBO Journal*, vol. 38, no. 7, p. e100293, 2019, doi: 10.15252/embj.2018100293.
[94] A. Hajmirza, A. Emadali, A. Gauthier, O. Casasnovas, R. Gressin, and M. B. Callanan, "BET Family Protein BRD4: An Emerging Actor in NFκB Signaling in Inflammation and Cancer," *Biomedicines*, vol. 6, no. 1, p. 16, 2018.
[95] N. P. S. Crawford *et al.*, "Bromodomain 4 activation predicts breast cancer survival," *Proceedings of the National Academy of Sciences*, vol. 105, no. 17, pp. 6380-6385, 2008, doi: doi:10.1073/pnas.0710331105.
[96] D. A. Bisgrove, T. Mahmoudi, P. Henklein, and E. Verdin, "Conserved P-TEFb-interacting domain of BRD4 inhibits HIV transcription," *Proceedings of the National Academy of Sciences*, vol. 104, no. 34, pp. 13690-13695, 2007, doi: doi:10.1073/pnas.0705053104.
[97] R. Kohnken *et al.*, "Diminished microRNA-29b level is associated with BRD4-mediated activation of oncogenes in cutaneous T-cell lymphoma," *Blood*, vol. 131, no. 7, pp. 771-781, 2018, doi: 10.1182/blood-2017-09-805663.
[98] T. Iftner, J. Haedicke-Jarboui, S.-Y. Wu, and C.-M. Chiang, "Involvement of Brd4 in different steps of the papillomavirus life cycle," *Virus Research*, vol. 231, pp. 76-82, 2017/03/02/ 2017, doi: 10.1016/j.virusres.2016.12.006.
[99] M. E. White, J. M. Fenger, and W. E. Carson, "Emerging roles of and therapeutic strategies targeting BRD4 in cancer," *Cellular Immunology*, vol. 337, pp. 48-53, 2019/03/01/ 2019, doi: 10.1016/j.cellimm.2019.02.001.
[100] Y. Duan, Y. Guan, W. Qin, X. Zhai, B. Yu, and H. Liu, "Targeting Brd4 for cancer therapy: inhibitors and degraders," *MedChemComm*, 10.1039/C8MD00198G vol. 9, no. 11, pp. 1779-1802, 2018, doi: 10.1039/C8MD00198G.
[101] Z. Liu *et al.*, "Discovery, X-ray Crystallography, and Anti-inflammatory Activity of Bromodomain-containing Protein 4 (BRD4) BD1 Inhibitors Targeting a Distinct

New Binding Site," *Journal of Medicinal Chemistry*, vol. 65, no. 3, pp. 2388-2408, 2022/02/10 2022, doi: 10.1021/acs.jmedchem.1c01851.
[102] J.-l. Yu *et al.*, "NMR-based platform for fragment-based lead discovery used in screening BRD4-targeted compounds," *Acta Pharmacologica Sinica*, Article vol. 37, no. 7, pp. 984-993, Jul 2016, doi: 10.1038/aps.2016.19.
[103] T. Yokoyama, K. Matsumoto, A. Ostermann, T. E. Schrader, Y. Nabeshima, and M. Mizuguchi, "Structural and thermodynamic characterization of the binding of isoliquiritigenin to the first bromodomain of BRD4," *The FEBS Journal*, vol. 286, no. 9, pp. 1656-1667, 2019/05/01 2019, doi: 10.1111/febs.14736.
[104] C. Muvva, E. R. A. Singam, S. S. Raman, and V. Subramanian, "Structure-based virtual screening of novel, high-affinity BRD4 inhibitors," *Molecular BioSystems*, 10.1039/C4MB00243A vol. 10, no. 9, pp. 2384-2397, 2014, doi: 10.1039/C4MB00243A.
[105] V. K. Bhardwaj, P. Das, and R. Purohit, "Integrating microsecond timescale classical and biased molecular dynamics simulations to screen potential molecules for BRD4-BD1," *Chaos, Solitons & Fractals*, vol. 167, p. 113061, 2023/02/01/ 2023, doi: 10.1016/j.chaos.2022.113061.
[106] K. Wu *et al.*, "Virtual Screening of Antitumor Inhibitors Targeting BRD4 Based on Machine Learning Methods," *ChemistrySelect*, vol. 7, no. 5, p. e202104054, 2022, doi: 10.1002/slct.202104054.
[107] W. Lu, R. Zhang, H. Jiang, H. Zhang, and C. Luo, "Computer-Aided Drug Design in Epigenetics," *Frontiers in Chemistry*, Review vol. 6, 2018, doi: 10.3389/fchem.2018.00057.
[108] X. Lucas *et al.*, "4-Acyl Pyrroles: Mimicking Acetylated Lysines in Histone Code Reading," *Angewandte Chemie International Edition*, vol. 52, no. 52, pp. 14055-14059, 2013/12/23 2013, doi: 10.1002/anie.201307652.
[109] B. K. Allen, S. Mehta, S. W. Ember, E. Schonbrunn, N. Ayad, and S. C. Schürer, "Large-Scale Computational Screening Identifies First in Class Multitarget Inhibitor of EGFR Kinase and BRD4," (in eng), *Sci Rep*, vol. 5, p. 16924, Nov 24 2015, doi: 10.1038/srep16924.
[110] X. Xue *et al.*, "Discovery of Benzo[cd]indol-2(1H)-ones as Potent and Specific BET Bromodomain Inhibitors: Structure-Based Virtual Screening, Optimization, and Biological Evaluation," *Journal of Medicinal Chemistry*, vol. 59, no. 4, pp. 1565-1579, 2016/02/25 2016, doi: 10.1021/acs.jmedchem.5b01511.
[111] I. Ali, J. Lee, A. Go, G. Choi, and K. Lee, "Discovery of novel [1,2,4]triazolo[4,3-a]quinoxaline aminophenyl derivatives as BET inhibitors for cancer treatment," *Bioorganic & Medicinal Chemistry Letters*, vol. 27, no. 20, pp. 4606-4613, 2017/10/15/ 2017, doi: 10.1016/j.bmcl.2017.09.025.
[112] M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knappcd, and P. C. Biggin, "Accurate calculation of the absolute free energy of binding for drug molecules," *Chemical Science*, Article vol. 7, no. 1, pp. 207-218, 2016 2016, doi: 10.1039/c5sc02678d.
[113] M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp, and P. C. Biggin, "Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations," *Journal of the American Chemical Society*, Article vol. 139, no. 2, pp. 946-957, Jan 18 2017, doi: 10.1021/jacs.6b11467.
[114] G. Heinzelmann, N. M. Henriksen, and M. K. Gilson, "Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain," *Journal of Chemical Theory and Computation*, Article vol. 13, no. 7, pp. 3260-3275, Jul 2017, doi: 10.1021/acs.jctc.7b00275.

[115] D. J. Huggins, "Comparing the Performance of Different AMBER Protein Forcefields, Partial Charge Assignments, and Water Models for Absolute Binding Free Energy Calculations," *Journal of Chemical Theory and Computation*, vol. 18, no. 4, pp. 2616-2630, 2022/04/12 2022, doi: 10.1021/acs.jctc.1c01208.
[116] B. M. Dickson, P. W. de Waal, Z. H. Ramjan, H. E. Xu, and S. B. Rothbart, "A fast, open source implementation of adaptive biasing potentials uncovers a ligand design strategy for the chromatin regulator BRD4," *The Journal of Chemical Physics*, vol. 145, no. 15, p. 154113, 2016/10/21 2016, doi: 10.1063/1.4964776.
[117] E. Van Quickelberghe, D. De Sutter, G. van Loo, S. Eyckerman, and K. Gevaert, "A protein-protein interaction map of the TNF-induced NF-κB signal transduction pathway," *Scientific Data*, vol. 5, no. 1, p. 180289, 2018/12/18 2018, doi: 10.1038/sdata.2018.289.
[118] D. V. Arkhipov, S. N. Lomin, Y. A. Myakushina, E. M. Savelieva, D. I. Osolodkin, and G. A. Romanov, "Modeling of Protein–Protein Interactions in Cytokinin Signal Transduction," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2096, 2019.
[119] V. Perovic, N. Sumonja, B. Gemovic, E. Toska, S. G. Roberts, and N. Veljkovic, "TRI_tool: a web-tool for prediction of protein–protein interactions in human transcriptional regulation," *Bioinformatics*, vol. 33, no. 2, pp. 289-291, 2016, doi: 10.1093/bioinformatics/btw590.
[120] A. Yasgar, A. Jadhav, A. Simeonov, and N. P. Coussens, "AlphaScreen-Based Assays: Ultra-High-Throughput Screening for Small-Molecule Inhibitors of Challenging Enzymes and Protein-Protein Interactions," in *High Throughput Screening: Methods and Protocols*, W. P. Janzen Ed. New York, NY: Springer New York, 2016, pp. 77-98.
[121] Y. Qiu, X. Li, X. He, J. Pu, J. Zhang, and S. Lu, "Computational methods-guided design of modulators targeting protein-protein interactions (PPIs)," *European Journal of Medicinal Chemistry*, vol. 207, p. 112764, 2020/12/01/ 2020, doi: 10.1016/j.ejmech.2020.112764.
[122] A. E. Modell, S. L. Blosser, and P. S. Arora, "Systematic Targeting of Protein–Protein Interactions," *Trends in Pharmacological Sciences*, vol. 37, no. 8, pp. 702-713, 2016/08/01/ 2016, doi: 10.1016/j.tips.2016.05.008.
[123] L. Mabonga and A. P. Kappo, "Protein-protein interaction modulators: advances, successes and remaining challenges," *Biophysical Reviews*, vol. 11, no. 4, pp. 559-581, 2019/08/01 2019, doi: 10.1007/s12551-019-00570-x.
[124] T. Siebenmorgen and M. Zacharias, "Computational prediction of protein–protein binding affinities," *WIREs Computational Molecular Science*, vol. 10, no. 3, p. e1448, 2020/05/01 2020, doi: 10.1002/wcms.1448.
[125] M. M. Gromiha, K. Yugandhar, and S. Jemimah, "Protein–protein interactions: scoring schemes and binding affinity," *Current Opinion in Structural Biology*, vol. 44, pp. 31-38, 2017/06/01/ 2017, doi: 10.1016/j.sbi.2016.10.016.
[126] J. Oláh, T. Szénási, A. Lehotzky, V. Norris, and J. Ovádi, "Challenges in Discovering Drugs That Target the Protein–Protein Interactions of Disordered Proteins," *International Journal of Molecular Sciences*, vol. 23, no. 3, p. 1550, 2022.
[127] M. D. S. Kumar and M. M. Gromiha, "PINT: Protein–protein Interactions Thermodynamic Database," *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D195-D198, 2006, doi: 10.1093/nar/gkj017.

[128] P. L. Kastritis and A. M. Bonvin, "On the binding affinity of macromolecular interactions: daring to ask why proteins interact," (in eng), *J R Soc Interface,* vol. 10, no. 79, p. 20120835, Feb 2013, doi: 10.1098/rsif.2012.0835.
[129] S. Y. Noskov and C. Lim, "Free Energy Decomposition of Protein-Protein Interactions," *Biophysical Journal,* vol. 81, no. 2, pp. 737-750, 2001, doi: 10.1016/S0006-3495(01)75738-4.
[130] M. Bastos *et al.*, "Isothermal titration calorimetry," *Nature Reviews Methods Primers,* vol. 3, no. 1, p. 17, 2023/03/09 2023, doi: 10.1038/s43586-023-00199-x.
[131] R. J. Falconer, B. Schuur, and A. K. Mittermaier, "Applications of isothermal titration calorimetry in pure and applied research from 2016 to 2020," *J. Mol. Recognit.,* vol. 34, no. 10, p. e2901, 2021, doi: 10.1002/jmr.2901.
[132] G. Klebe, "Applying thermodynamic profiling in lead finding and optimization," *Nature Reviews Drug Discovery,* vol. 14, no. 2, pp. 95-110, 2015/02/01 2015, doi: 10.1038/nrd4486.
[133] D. G. Myszka *et al.*, "The ABRF-MIRG'02 study: assembly state, thermodynamic, and kinetic analysis of an enzyme/inhibitor interaction," (in eng), *J Biomol Tech,* vol. 14, no. 4, pp. 247-69, Dec 2003.
[134] J. Tellinghuisen and J. D. Chodera, "Systematic errors in isothermal titration calorimetry: Concentrations and baselines," *Analytical Biochemistry,* vol. 414, no. 2, pp. 297-299, 2011/07/15/ 2011, doi: 10.1016/j.ab.2011.03.024.
[135] D. Hahn *et al.*, "Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]," *Living Journal of Computational Molecular Science,* vol. 4, no. 1, p. 1497, 08/30 2022, doi: 10.33011/livecoms.4.1.1497.
[136] A. R. Leach, *Molecular Modelling: Principles and Applications*. Longman, 1996.
[137] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*. OUP Oxford, 2010.
[138] K. Zhou and B. Liu, "Chapter 1 - Fundamentals of classical molecular dynamics simulation," in *Molecular Dynamics Simulation,* K. Zhou and B. Liu Eds.: Elsevier, 2022, pp. 1-40.
[139] K. Zhou and B. Liu, "Chapter 2 - Potential energy functions," in *Molecular Dynamics Simulation,* K. Zhou and B. Liu Eds.: Elsevier, 2022, pp. 41-65.
[140] K. Zhou and B. Liu, "Chapter 3 - Control techniques of molecular dynamics simulation," in *Molecular Dynamics Simulation,* K. Zhou and B. Liu Eds.: Elsevier, 2022, pp. 67-96.
[141] A. R. Leach and A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.
[142] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," *Journal of Computational Chemistry,* Review vol. 26, no. 16, pp. 1701-1718, Dec 2005, doi: 10.1002/jcc.20291.
[143] A.-P. Hynninen and M. F. Crowley, "New faster CHARMM molecular dynamics engine," *Journal of Computational Chemistry,* vol. 35, no. 5, pp. 406-413, 2014/02/15 2014, doi: 10.1002/jcc.23501.
[144] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the Amber biomolecular simulation package," *WIREs Computational Molecular Science,* vol. 3, no. 2, pp. 198-210, 2013/03/01 2013, doi: 10.1002/wcms.1121.
[145] J. A. Rackers *et al.*, "Tinker 8: Software Tools for Molecular Design," *Journal of Chemical Theory and Computation,* vol. 14, no. 10, pp. 5273-5289, 2018/10/09 2018, doi: 10.1021/acs.jctc.8b00529.

[146] H. Chen *et al.*, "Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD," *Journal of Chemical Information and Modeling,* vol. 60, no. 11, pp. 5301-5307, 2020/11/23 2020, doi: 10.1021/acs.jcim.0c00745.
[147] M. J. Abraham and J. E. Gready, "Optimization of parameters for molecular dynamics simulation using smooth particle-mesh Ewald in GROMACS 4.5," *Journal of Computational Chemistry,* vol. 32, no. 9, pp. 2031-2040, 2011, doi: 10.1002/jcc.21773.
[148] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A Smooth Particle Mesh Ewald Method," *J. Chem. Phys.,* Article vol. 103, no. 19, pp. 8577-8593, Nov 15 1995, doi: 10.1063/1.470117.
[149] L. Verlet, "Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules," (in English), *Physical Review,* Article vol. 159, no. 1, pp. 98-103, 1967, doi: 10.1103/PhysRev.159.98.
[150] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics,* Article vol. 76, no. 1, pp. 637-649, 1982, doi: 10.1063/1.442716.
[151] W. F. Van Gunsteren and H. J. C. Berendsen, "A Leap-frog Algorithm for Stochastic Dynamics," *Molecular Simulation,* vol. 1, no. 3, pp. 173-185, 1988/03/01 1988, doi: 10.1080/08927028808080941.
[152] P. H. Hünenberger, "Thermostat Algorithms for Molecular Dynamics Simulations," in *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*, C. Dr. Holm and K. Prof. Dr. Kremer Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 105-149.
[153] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics,* vol. 23, no. 3, pp. 327-341, 1977/03/01/ 1977, doi: 10.1016/0021-9991(77)90098-5.
[154] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, "LINCS: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry,* Article vol. 18, no. 12, pp. 1463-1472, Sep 1997, doi: 10.1002/(sici)1096-987x(199709)18:12<1463::Aid-jcc4>3.3.Co;2-l.
[155] P. Yi and G. C. Rutledge, "Molecular simulation of crystal nucleation in n-octane melts," *The Journal of Chemical Physics,* vol. 131, no. 13, p. 134902, 2009/10/07 2009, doi: 10.1063/1.3240202.
[156] D. Zahn and S. Leoni, "Nucleation and Growth in Pressure-Induced Phase Transitions from Molecular Dynamics Simulations: Mechanism of the Reconstructive Transformation of NaCl to the CsCl-Type Structure," *Physical Review Letters,* vol. 92, no. 25, p. 250201, 06/24/ 2004, doi: 10.1103/PhysRevLett.92.250201.
[157] J. Zhao *et al.*, "A structure model for phase separated fluoroaluminosilicate glass system by molecular dynamic simulations," *Journal of the European Ceramic Society,* vol. 39, no. 15, pp. 5018-5029, 2019/12/01/ 2019, doi: 10.1016/j.jeurceramsoc.2019.06.042.
[158] Q. Shi and G. A. Voth, "Multi-Scale Modeling of Phase Separation in Mixed Lipid Bilayers," *Biophysical Journal,* vol. 89, no. 4, pp. 2385-2394, 2005/10/01/ 2005, doi: 10.1529/biophysj.105.063784.
[159] P. Lafourcade, C. Denoual, and J.-B. Maillet, "Dislocation Core Structure at Finite Temperature Inferred by Molecular Dynamics Simulations for 1,3,5-Triamino-

2,4,6-trinitrobenzene Single Crystal," *The Journal of Physical Chemistry C,* vol. 121, no. 13, pp. 7442-7449, 2017/04/06 2017, doi: 10.1021/acs.jpcc.6b11576.
[160] L. Deng and J. Du, "Effects of system size and cooling rate on the structure and properties of sodium borosilicate glasses from molecular dynamics simulations," *The Journal of Chemical Physics,* vol. 148, no. 2, p. 024504, 2018/01/14 2018, doi: 10.1063/1.5007083.
[161] E. Paci and M. Marchi, "Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure," *Proceedings of the National Academy of Sciences,* vol. 93, no. 21, pp. 11609-11614, 1996/10/15 1996, doi: 10.1073/pnas.93.21.11609.
[162] F. Lv, P. Liu, H. Qi, J. Liu, R. Sun, and W. Wang, "The early stage of the thermal pulse explosions of aluminum nanowires under different energy deposition levels," *Computational Materials Science,* vol. 170, p. 109142, 2019/12/01/ 2019, doi: 10.1016/j.commatsci.2019.109142.
[163] H. Morita, K. Tanaka, T. Kajiyama, T. Nishi, and M. Doi, "Study of the Glass Transition Temperature of Polymer Surface by Coarse-Grained Molecular Dynamics Simulation," *Macromolecules,* vol. 39, no. 18, pp. 6233-6237, 2006/09/01 2006, doi: 10.1021/ma052632h.
[164] S. Hu, W. Setyawan, V. V. Joshi, and C. A. Lavender, "Atomistic simulations of thermodynamic properties of Xe gas bubbles in U10Mo fuels," *Journal of Nuclear Materials,* vol. 490, pp. 49-58, 2017/07/01/ 2017, doi: 10.1016/j.jnucmat.2017.04.016.
[165] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics,* vol. 72, no. 4, pp. 2384-2393, 1980/02/15 1980, doi: 10.1063/1.439486.
[166] Q. An, W.-G. Liu, W. A. Goddard, III, T. Cheng, S. V. Zybin, and H. Xiao, "Initial Steps of Thermal Decomposition of Dihydroxylammonium 5,5′-bistetrazole-1,1′-diolate Crystals from Quantum Mechanics," *The Journal of Physical Chemistry C,* vol. 118, no. 46, pp. 27175-27181, 2014/11/20 2014, doi: 10.1021/jp509582x.
[167] X.-z. Deng, L. Lang, Y.-f. Mo, K.-j. Dong, Z.-a. Tian, and W.-y. Hu, "Solid−solid phase transition of tungsten induced by high pressure: A molecular dynamics simulation," *Transactions of Nonferrous Metals Society of China,* vol. 30, no. 11, pp. 2980-2993, 2020/11/01/ 2020, doi: 10.1016/S1003-6326(20)65436-X.
[168] M. A. González, "Force fields and molecular dynamics simulations," *JDN,* vol. 12, pp. 169-200, 2011, doi: 10.1051/sfn/201112009.
[169] H. C. Urey and C. A. Bradley, "The Vibrations of Pentatonic Tetrahedral Molecules," *Physical Review,* vol. 38, no. 11, pp. 1969-1978, 12/01/ 1931, doi: 10.1103/PhysRev.38.1969.
[170] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids," *Journal of the American Chemical Society,* vol. 118, no. 45, pp. 11225-11236, 1996/11/13 1996, doi: 10.1021/ja9621760.
[171] E. Sigfridsson and U. Ryde, "Comparison of methods for deriving atomic charges from the electrostatic potential and moments," *Journal of Computational Chemistry,* vol. 19, no. 4, pp. 377-395, 1998/03/01 1998.
[172] D. F. Green and B. Tidor, "Evaluation of ab Initio Charge Determination Methods for Use in Continuum Solvation Calculations," *The Journal of Physical Chemistry B,* vol. 107, no. 37, pp. 10261-10273, 2003/09/01 2003, doi: 10.1021/jp0350971.
[173] B. Guillot, "A reappraisal of what we have learnt during three decades of computer simulations on water," *J. Mol. Liq.,* Article; Proceedings Paper vol. 101, no. 1-3,

pp. 219-260, Nov 2002, Art no. Pii s0167-7322(02) 00094-6, doi: 10.1016/s0167-7322(02)00094-6.
[174] S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building Water Models: A Different Approach," *J. Phys. Chem. Lett.*, Article vol. 5, no. 21, pp. 3863-3871, Nov 6 2014, doi: 10.1021/jz501780a.
[175] A. V. Onufriev and S. Izadi, "Water models for biomolecular simulations," *WIREs Computational Molecular Science*, vol. 8, no. 2, p. e1347, 2018/03/01 2018, doi: 10.1002/wcms.1347.
[176] J. F. Ouyang and R. P. A. Bettens, "Modelling Water: A Lifetime Enigma," *Chimia*, Article vol. 69, no. 3, pp. 104-111, 2015 2015, doi: 10.2533/chimia.2015.104.
[177] M. Chaplin. "Water Models." https://water.lsbu.ac.uk/water/water_models.html (accessed 20 December, 2022).
[178] J. D. Bernal and R. H. Fowler, "A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions," (in English), *J. Chem. Phys.*, Article vol. 1, no. 8, pp. 515-548, Aug 1933, doi: 10.1063/1.1749327.
[179] G. W. Robinson, *Water in biology, chemistry, and physics: experimental overviews and computational methodologies*. World Scientific, 1996.
[180] P. G. Kusalik and I. M. Svishchev, "The Spatial Structure in Liquid Water," (in English), *Science*, Article vol. 265, no. 5176, pp. 1219-1221, Aug 1994, doi: 10.1126/science.265.5176.1219.
[181] L. A. Baez and P. Clancy, "Existence of a Density Maximum in Extended Simple Point-Charge Water," (in English), *J. Chem. Phys.*, Article vol. 101, no. 11, pp. 9837-9840, Dec 1994, doi: 10.1063/1.467949.
[182] H. W. Horn *et al.*, "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew," (in English), *J. Chem. Phys.*, Article vol. 120, no. 20, pp. 9665-9678, May 2004, doi: 10.1063/1.1683075.
[183] N. W. Silver *et al.*, "Efficient Computation of Small-Molecule Configurational Binding Entropy and Free Energy Changes by Ensemble Enumeration," *Journal of Chemical Theory and Computation*, vol. 9, no. 11, pp. 5098-5115, 2013/11/12 2013, doi: 10.1021/ct400383v.
[184] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, "The statistical-thermodynamic basis for computation of binding affinities: a critical review," (in eng), *Biophys J*, vol. 72, no. 3, pp. 1047-69, Mar 1997, doi: 10.1016/s0006-3495(97)78756-3.
[185] H. S. Hansen and P. H. Hünenberger, "Using the local elevation method to construct optimized umbrella sampling potentials: Calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water," *Journal of Computational Chemistry*, vol. 31, no. 1, pp. 1-23, 2010/01/15 2010, doi: 10.1002/jcc.21253.
[186] M. K. Gilson and K. K. Irikura, "Symmetry Numbers for Rigid, Flexible, and Fluxional Molecules: Theory and Applications," (in English), *J. Phys. Chem. B*, Article vol. 114, no. 49, pp. 16304-16317, Dec 2010, doi: 10.1021/jp110434s.
[187] M. T. Record, "Cooperativity Theory in Biochemistry - Steady-State and Equilibrium Systems," (in English), *Science*, Book Review vol. 229, no. 4718, pp. 1080-1081, 1985, doi: 10.1126/science.229.4718.1080.
[188] M. Aldeghi, J. P. Bluck, and P. C. Biggin, "Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide," in *Computational Drug Discovery and Design*, M. Gore and U. B. Jagtap Eds. New York, NY: Springer New York, 2018, pp. 199-232.

[189] I. Alibay, A. Magarkar, D. Seeliger, and P. C. Biggin, "Evaluating the use of absolute binding free energy in the fragment optimisation process," *Communications Chemistry*, vol. 5, no. 1, p. 105, 2022/09/05 2022, doi: 10.1038/s42004-022-00721-4.
[190] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, "Absolute Binding Free Energies: A Quantitative Approach for Their Calculation," *The Journal of Physical Chemistry B*, vol. 107, no. 35, pp. 9535-9551, 2003/09/01 2003, doi: 10.1021/jp0217839.
[191] P. Hall, "Chapter 39 Methodology and theory for the bootstrap," in *Handbook of Econometrics*, vol. 4: Elsevier, 1994, pp. 2341-2381.
[192] H. Flyvbjerg, "Error estimates on averages of correlated data," in *Eotvos Summer School on Advances in Computer Simulation*, Budapest, Hungary, 1996, vol. 501, in Lecture Notes in Physics, 1998, pp. 88-103.
[193] H. Flyvbjerg and H. G. Petersen, "Error estimates on averages of correlated data," *The Journal of Chemical Physics*, vol. 91, no. 1, pp. 461-466, 1989/07/01 1989, doi: 10.1063/1.457480.
[194] A. Grossfield and D. M. Zuckerman, "Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations," in *Annual Reports in Computational Chemistry*, vol. 5, R. A. Wheeler Ed.: Elsevier, 2009, pp. 23-48.
[195] R. M. Lee, G. J. Conduit, N. Nemec, P. L. Rios, and N. D. Drummond, "Strategies for improving the efficiency of quantum Monte Carlo calculations," *Physical Review E*, Article vol. 83, no. 6, Jun 27 2011, Art no. 066706, doi: 10.1103/PhysRevE.83.066706.
[196] U. Wolff and A. Collaboration, "Monte Carlo errors with less errors," *Computer Physics Communications*, Article vol. 156, no. 2, pp. 143-153, Jan 1 2004, doi: 10.1016/s0010-4655(03)00467-3.
[197] J. Spencer. "pyblock." https://pyblock.readthedocs.io/en/latest/ (accessed 27 December, 2022).
[198] K. Han, P. S. Hudson, M. R. Jones, N. Nishikawa, F. Tofoleanu, and B. R. Brooks, "Prediction of CB 8 host-guest binding free energies in SAMPL6 using the double-decoupling method," *J. Comput.-Aided Mol. Des.*, Article vol. 32, no. 10, pp. 1059-1073, Oct 2018, doi: 10.1007/s10822-018-0144-8.
[199] D. R. Bell *et al.*, "Calculating binding free energies of host-guest systems using the AMOEBA polarizable force field," *Physical Chemistry Chemical Physics*, Article vol. 18, no. 44, pp. 30261-30269, 2016 2016, doi: 10.1039/c6cp02509a.
[200] L. Wickstrom *et al.*, "Parameterization of an effective potential for protein-ligand binding from host-guest affinity data," *J. Mol. Recognit.*, Article vol. 29, no. 1, pp. 10-21, Jan 2016, doi: 10.1002/jmr.2489.
[201] S. Moghaddam, Y. Inoue, and M. K. Gilson, "Host-Guest Complexes with Protein-Ligand-like Affinities: Computational Analysis and Design," *Journal of the American Chemical Society*, Article vol. 131, no. 11, pp. 4012-4021, Mar 25 2009, doi: 10.1021/ja808175m.
[202] K. N. Houk, A. G. Leach, S. P. Kim, and X. Y. Zhang, "Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes," *Angew. Chem.-Int. Edit.*, Review vol. 42, no. 40, pp. 4872-4897, 2003 2003, doi: 10.1002/anie.200200565.
[203] S. Fan, B. I. Iorga, and O. Beckstein, "Prediction of octanol-water partition coefcients for the SAMPL6-log P molecules using molecular dynamics simulations with OPLS-AA, AMBER and CHARMM force felds," (in English), *J. Comput.-*

Aided Mol. Des., Article vol. 34, no. 5, pp. 543-560, May 2020, doi: 10.1007/s10822-019-00267-z.

[204] R. B. Best, "Atomistic Force Fields for Proteins," in *Biomolecular Simulations: Methods and Protocols*, vol. 2022, M. Bonomi and C. Camilloni Eds., (Methods in Molecular Biology, 2019, pp. 3-19.

[205] B. A. Bauer and S. Patel, "Properties of water along the liquid-vapor coexistence curve via molecular dynamics simulations using the polarizable TIP4P-QDP-LJ water model," *J. Chem. Phys.*, Article vol. 131, no. 8, Aug 28 2009, Art no. 084709, doi: 10.1063/1.3200869.

[206] R. Sakamaki, A. K. Sum, T. Narumi, and K. Yasuoka, "Molecular dynamics simulations of vapor/liquid coexistence using the nonpolarizable water models," *J. Chem. Phys.*, Article vol. 134, no. 12, Mar 28 2011, Art no. 124708, doi: 10.1063/1.3574038.

[207] C. Vega and E. de Miguel, "Surface tension of the most popular models of water by using the test-area simulation method," *J. Chem. Phys.*, Article vol. 126, no. 15, Apr 21 2007, Art no. 154707, doi: 10.1063/1.2715577.

[208] P. M. de Hijes, E. Sanz, L. Joly, C. Valeriani, and F. Caupin, "Viscosity and self-diffusion of supercooled and stretched water from molecular dynamics simulations," *J. Chem. Phys.*, Article vol. 149, no. 9, Sep 7 2018, Art no. 094503, doi: 10.1063/1.5042209.

[209] Y. Xiong, P. S. Shabane, and A. V. Onufriev, "Melting Points of OPC and OPC3 Water Models," *ACS Omega*, vol. 5, no. 39, pp. 25087-25094, 2020/10/06 2020, doi: 10.1021/acsomega.0c02638.

[210] C. I. Lynch, S. Rao, and M. S. P. Sansom, "Water in Nanopores and Biological Channels: A Molecular Simulation Perspective," *Chem. Rev.*, Review vol. 120, no. 18, pp. 10298-10335, Sep 23 2020, doi: 10.1021/acs.chemrev.9b00830.

[211] D. R. Nutt and J. C. Smith, "Molecular dynamics simulations of proteins: Can the explicit water model be varied?," *Journal of Chemical Theory and Computation*, Article vol. 3, no. 4, pp. 1550-1560, Jul-Aug 2007, doi: 10.1021/ct700053u.

[212] Z. Tang and C.-e. A. Chang, "Binding Thermodynamics and Kinetics Calculations Using Chemical Host and Guest: A Comprehensive Picture of Molecular Recognition," *Journal of Chemical Theory and Computation*, Article vol. 14, no. 1, pp. 303-318, Jan 2018, doi: 10.1021/acs.jctc.7b00899.

[213] M. Erdos, M. Frangou, T. J. H. Vlugt, and O. A. Moultos, "Diffusivity of alpha-, beta-, gamma-cyclodextrin and the inclusion complex of beta-cyclodextrin: Ibuprofen in aqueous solutions; A molecular dynamics simulation study," *Fluid Phase Equilib.*, Article vol. 528, Jan 15 2021, Art no. 112842, doi: 10.1016/j.fluid.2020.112842.

[214] S. Moghaddam *et al.*, "New Ultrahigh Affinity Host-Guest Complexes of Cucurbit 7 uril with Bicyclo 2.2.2 octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations," *Journal of the American Chemical Society*, Article vol. 133, no. 10, pp. 3570-3581, Mar 16 2011, doi: 10.1021/ja109904u.

[215] H. Yin, R. Wang, J. Wan, Y. Zheng, D. Ouyang, and R. Wang, "Molecular Encapsulation of Histamine H-2-Receptor Antagonists by Cucurbit 7 Uril: An Experimental and Computational Study," *Molecules*, Article vol. 21, no. 9, Sep 2016, Art no. 1178, doi: 10.3390/molecules21091178.

[216] N. A. Tcyrulnikov, R. Varadharajan, A. A. Tikhomirova, M. Pattabiraman, V. Ramamurthy, and R. M. Wilson, "Modulation of Reduction Potentials of Bis(pyridinium)alkane Dications through Encapsulation within Cucurbit 7 uril," *J.*

Org. Chem., Article vol. 84, no. 13, pp. 8759-8765, Jul 5 2019, doi: 10.1021/acs.joc.9b01049.

[217] J.-S. Yu, F.-G. Wu, L.-F. Tao, J.-J. Luo, and Z.-W. Yu, "Mechanism of the fast exchange between bound and free guests in cucurbit 7 uril-guest systems," *Physical Chemistry Chemical Physics,* Article vol. 13, no. 9, pp. 3638-3641, 2011 2011, doi: 10.1039/c0cp02349c.

[218] Z. Miskolczy and L. Biczok, "Kinetics and Thermodynamics of Berberine Inclusion in Cucurbit 7 uril," *J. Phys. Chem. B,* Article vol. 118, no. 9, pp. 2499-2505, Mar 6 2014, doi: 10.1021/jp500603g.

[219] Z. Miskolczy, M. Megyesi, O. Toke, and L. Biczok, "Change of the kinetics of inclusion in cucurbit 7 uril upon hydrogenation and methylation of palmatine," *Physical Chemistry Chemical Physics,* Article vol. 21, no. 9, pp. 4912-4919, Mar 7 2019, doi: 10.1039/c8cp07231k.

[220] S. Z. Ndendjio, W. Liu, N. Yvanez, Z. Meng, P. Y. Zavalij, and L. Isaacs, "Triptycene walled glycoluril trimer: synthesis and recognition properties," *New J. Chem.,* Article vol. 44, no. 2, pp. 338-345, Jan 14 2020, doi: 10.1039/c9nj05336k.

[221] K. Vanommeslaeghe and A. D. MacKerell, Jr., "Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing," *Journal of Chemical Information and Modeling,* Article vol. 52, no. 12, pp. 3144-3154, Dec 2012, doi: 10.1021/ci300363c.

[222] K. Vanommeslaeghe, E. P. Raman, and A. D. MacKerell, Jr., "Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges," *Journal of Chemical Information and Modeling,* Article vol. 52, no. 12, pp. 3155-3168, Dec 2012, doi: 10.1021/ci3003649.

[223] J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of Computational Chemistry,* Article vol. 25, no. 9, pp. 1157-1174, Jul 15 2004, doi: 10.1002/jcc.20035.

[224] Y. D. Qiu *et al.,* "Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field," (in English), *Journal of Chemical Theory and Computation,* Article vol. 17, no. 10, pp. 6262-6280, Oct 2021, doi: 10.1021/acs.jctc.1c00571.

[225] V. Zoete, M. A. Cuendet, A. Grosdidier, and O. Michielin, "SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules," *Journal of Computational Chemistry,* Article vol. 32, no. 11, pp. 2359-2368, Aug 2011, doi: 10.1002/jcc.21816.

[226] P. Mark and L. Nilsson, "Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K," *J. Phys. Chem. A,* Article vol. 105, no. 43, pp. 9954-9960, Nov 1 2001, doi: 10.1021/jp003020w.

[227] 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.,* Article vol. 79, no. 2, pp. 926-935, 1983 1983, doi: 10.1063/1.445869.

[228] M. W. Mahoney and W. L. Jorgensen, "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions," (in English), *J. Chem. Phys.,* Article vol. 112, no. 20, pp. 8910-8922, May 2000, Art no. Pii [s0021-9606(00)50820-4], doi: 10.1063/1.481505.

[229] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method," *Journal of Computational Chemistry,* Article vol. 21, no. 2, pp. 132-146, Jan 30 2000.

[230] A. Jakalian, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation," *Journal of Computational Chemistry,* Article vol. 23, no. 16, pp. 1623-1641, Dec 2002, doi: 10.1002/jcc.10128.

[231] J. M. Wang, W. Wang, and P. A. Kollman, "Antechamber: An accessory software package for molecular mechanical calculations," *Abstr. Pap. Am. Chem. Soc.,* Meeting Abstract vol. 222, pp. U403-U403, Aug 2001..

[232] K. Vanommeslaeghe *et al.*, "CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields," *Journal of Computational Chemistry,* Article vol. 31, no. 4, pp. 671-690, Mar 2010, doi: 10.1002/jcc.21367.

[233] H. J. C. Berendsen, D. Vanderspoel, and R. Vandrunen, "GROMACS - A Message-Passing Parallel Molecular-Dynamics Implementation," *Computer Physics Communications,* Article vol. 91, no. 1-3, pp. 43-56, Sep 1995, doi: 10.1016/0010-4655(95)00042-e.

[234] J. W. Ponder and F. M. Richards, "An Efficient Newton-Like Method for Molecular Mechanics Energy Minimization of Large Molecules," *Journal of Computational Chemistry,* Article vol. 8, no. 7, pp. 1016-1024, Oct-Nov 1987, doi: 10.1002/jcc.540080710.

[235] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.,* Article vol. 126, no. 1, Jan 7 2007, Art no. 014101, doi: 10.1063/1.2408420.

[236] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied physics,* vol. 52, no. 12, pp. 7182-7190, 1981.

[237] B. Hess, "P-LINCS: A parallel linear constraint solver for molecular simulation," *Journal of Chemical Theory and Computation,* Article vol. 4, no. 1, pp. 116-122, Jan 2008, doi: 10.1021/ct700200b.

[238] C. Cezard, X. Trivelli, F. Aubry, F. Djedaini-Pilard, and F.-Y. Dupradeau, "Molecular dynamics studies of native and substituted cyclodextrins in different media: 1. Charge derivation and force field performances," *Physical Chemistry Chemical Physics,* Article vol. 13, no. 33, pp. 15103-15121, 2011 2011, doi: 10.1039/c1cp20854c.

[239] V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly, and D. L. Mobley, "Benchmark assessment of molecular geometries and energies from small molecule force fields," (in eng), *F1000Res,* vol. 9, 2020, doi: 10.12688/f1000research.27141.1.

[240] J. Traeg and D. Zahn, "Improved GAFF2 parameters for fluorinated alkanes and mixed hydro- and fluorocarbons," *J. Mol. Model.,* Article vol. 25, no. 2, Feb 2019, Art no. 39, doi: 10.1007/s00894-018-3911-5.

[241] Y. Khalak, G. Tresadern, B. L. de Groot, and V. Gapsys, "Non-equilibrium approach for binding free energies in cyclodextrins in SAMPL7: force fields and software," *J. Comput.-Aided Mol. Des.,* Article vol. 35, no. 1, pp. 49-61, Jan 2021, doi: 10.1007/s10822-020-00359-1.

[242] S. Zhu, "Validation of the Generalized Force Fields GAFF, CGenFF, OPLS-AA, and PRODRGFF by Testing Against Experimental Osmotic Coefficient Data for Small Drug-Like Molecules," *Journal of Chemical Information and Modeling,* Article vol. 59, no. 10, pp. 4239-4247, Oct 2019, doi: 10.1021/acs.jcim.9b00552.

[243] V. Kumar, K. S. Rane, S. Wierzchowski, M. Shaik, and J. R. Errington, "Evaluation of the Performance of GAFF and CGenFF in the Prediction of Liquid-

Vapor Saturation Properties of Naphthalene Derivatives," *Ind. Eng. Chem. Res.*, Article vol. 53, no. 41, pp. 16072-16081, Oct 15 2014, doi: 10.1021/ie503346m.
[244] P. Dauber-Osguthorpe and A. T. Hagler, "Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there?," *J. Comput.-Aided Mol. Des.*, Article vol. 33, no. 2, pp. 133-203, Feb 2019, doi: 10.1007/s10822-018-0111-4.
[245] P. S. Nerenberg and T. Head-Gordon, "New developments in force fields for biomolecular simulations," *Current Opinion in Structural Biology*, Article vol. 49, pp. 129-138, Apr 2018, doi: 10.1016/j.sbi.2018.02.002.
[246] L. Mei *et al.*, "Supramolecular Host-Guest Inclusion for Distinguishing Cucurbit 7 uril-Based Pseudorotaxanes from Small-Molecule Ligands in Coordination Assembly with a Uranyl Center," *Chem.-Eur. J.*, Article vol. 23, no. 56, pp. 13995-14003, Oct 9 2017, doi: 10.1002/chem.201702752.
[247] I. M. Kenney, O. Beckstein, and B. I. Iorga, "Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field," *J. Comput.-Aided Mol. Des.*, Article vol. 30, no. 11, pp. 1045-1058, Nov 2016, doi: 10.1007/s10822-016-9949-5.
[248] P. Rao *et al.*, "Reckoning a fungal metabolite, Pyranonigrin A as a potential Main protease (M-pro) inhibitor of novel SARS-CoV-2 virus identified using docking and molecular dynamics simulation," *Biophysical Chemistry*, Article vol. 264, Sep 2020, Art no. 106425, doi: 10.1016/j.bpc.2020.106425.
[249] X. Jiang *et al.*, "Simulations of octapeptin-outer membrane interactions reveal conformational flexibility is linked to antimicrobial potency," *Journal of Biological Chemistry*, Article vol. 295, no. 47, pp. 15902-15912, Nov 20 2020, doi: 10.1074/jbc.RA120.014856.
[250] L. S. Dodda, I. C. de Vaca, J. Tirado-Rives, and W. L. Jorgensen, "LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands," *Nucleic Acids Research*, Article vol. 45, no. W1, pp. W331-W336, Jul 3 2017, doi: 10.1093/nar/gkx312.
[251] A. K. Malde *et al.*, "An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0," *Journal of Chemical Theory and Computation*, Article vol. 7, no. 12, pp. 4026-4037, Dec 2011, doi: 10.1021/ct200196m.
[252] B. Hess, D. van der Spoel, M. J. Abraham, and E. Lindahl, "On the importance of accurate algorithms for reliable molecular dynamics simulations," 2019.
[253] S. Kale and J. Herzfeld, "Natural polarizability and flexibility via explicit valency: The case of water," *J. Chem. Phys.*, Article vol. 136, no. 8, Feb 28 2012, Art no. 084109, doi: 10.1063/1.3688228.
[254] J. L. Finney, "The water molecule and its interactions: the interaction between theory, modelling, and experiment," *J. Mol. Liq.*, Article; Proceedings Paper vol. 90, no. 1-3, pp. 303-312, Feb 2001, doi: 10.1016/s0167-7322(01)00134-9.
[255] J. L. Finney, "Water? What's so special about it?," *Philos. Trans. R. Soc. B-Biol. Sci.*, Article; Proceedings Paper vol. 359, no. 1448, pp. 1145-1163, Aug 29 2004, doi: 10.1098/rstb.2004.1495.
[256] S. P. Kadaoluwa Pathirannahalage *et al.*, "Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations," *Journal of Chemical Information and Modeling*, vol. 61, no. 9, pp. 4521-4536, 2021/09/27 2021, doi: 10.1021/acs.jcim.1c00794.
[257] V. S. Neverov and A. V. Komolkin, "A study of the structural and thermodynamic properties of water by the molecular dynamics method," *Russian Journal of*

Physical Chemistry B, vol. 4, no. 2, pp. 217-226, 2010/04/01 2010, doi: 10.1134/S1990793110020065.

[258] C. Vega and J. L. F. Abascal, "Simulating water with rigid non-polarizable models: a general perspective," *Physical Chemistry Chemical Physics,* 10.1039/C1CP22168J vol. 13, no. 44, pp. 19663-19688, 2011, doi: 10.1039/C1CP22168J.

[259] S. S. Çınaroğlu and P. C. Biggin, "The role of loop dynamics in the prediction of ligand–protein binding enthalpy," *Chemical Science,* 10.1039/D2SC06471E vol. 14, no. 24, pp. 6792-6805, 2023, doi: 10.1039/D2SC06471E.

[260] A. S. J. S. Mey *et al.*, "Best practices for alchemical free energy calculations.," *LiveCoMS,* vol. 1, 2021.

[261] V. Gapsys, D. F. Hahn, G. Tresadern, D. L. Mobley, M. Rampp, and B. L. de Groot, "Pre-Exascale Computing of Protein–Ligand Binding Free Energies with Open Source Software for Drug Design," *Journal of Chemical Information and Modeling,* vol. 62, no. 5, pp. 1172-1177, 2022/03/14 2022, doi: 10.1021/acs.jcim.1c01445.

[262] H. Fu *et al.*, "Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations," *Nature Protocols,* vol. 17, no. 4, pp. 1114-1141, 2022/04/01 2022, doi: 10.1038/s41596-021-00676-1.

[263] J. Wang, A. Ishchenko, W. Zhang, A. Razavi, and D. Langley, "A highly accurate metadynamics-based Dissociation Free Energy method to calculate protein–protein and protein–ligand binding potencies," *Scientific Reports,* vol. 12, no. 1, p. 2024, 2022/02/07 2022, doi: 10.1038/s41598-022-05875-8.

[264] G. Heinzelmann and M. K. Gilson, "Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation," *Scientific Reports,* vol. 11, no. 1, p. 1116, 2021/01/13 2021, doi: 10.1038/s41598-020-80769-1.

[265] R. M. Levy and E. Gallicchio, "Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects," *Ann. Rev. Phys. Chem.,* vol. 49, no. 1, pp. 531-567, 1998/10/01 1998, doi: 10.1146/annurev.physchem.49.1.531.

[266] N. Lu, D. A. Kofke, and T. B. Woolf, "Staging is more important than perturbation method for computation of enthalpy and entropy changes in complex systems," *J. Phys. Chem. B,* vol. 107, no. 23, pp. 5598-5611, 2003/06/01 2003, doi: 10.1021/jp027627j.

[267] M. A. Wyczalkowski, A. Vitalis, and R. V. Pappu, "New estimators for calculating solvation entropy and enthalpy and comparative assessments of their accuracy and precision," *J. Phys. Chem. B,* vol. 114, no. 24, pp. 8166-8180, 2010/06/24 2010, doi: 10.1021/jp103050u.

[268] E. Freire, "Do enthalpy and entropy distinguish first in class from best in class?," *Drug Discovery Today,* Review vol. 13, no. 19-20, pp. 869-874, Oct 2008, doi: 10.1016/j.drudis.2008.07.005.

[269] P. Filippakopoulos and S. Knapp, "The bromodomain interaction module," *FEBS Letts.,* vol. 586, no. 17, pp. 2692-2704, 8/14/ 2012, doi: 10.1016/j.febslet.2012.04.045.

[270] S. Shu and K. Polyak, "BET bromodomain proteins as cancer therapeutic targets," *Cold Spr. Harb. Symp. Quant. Biol.,* vol. 81, pp. 123-129, 2016.

[271] C.-Y. Wang and P. Filippakopoulos, "Beating the odds: BETs in disease," *Trends in Biochemical Sciences,* Review vol. 40, no. 8, pp. 468-479, Aug 2015, doi: 10.1016/j.tibs.2015.06.002.

[272] M. Aldeghi, M. J. Bodkin, S. Knapp, and P. C. Biggin, "A statistical analysis on the performance of MMPBSA versus absolute binding free energy calculations: bromodomains as a case study," *J. Chem. Inf. Model.*, vol. 57, pp. 2203-2221, 2017/08/08 2017, doi: 10.1021/acs.jcim.7b00347.
[273] M. Aldeghi, G. A. Ross, M. J. Bodkin, J. W. Essex, S. Knapp, and P. C. Biggin, "Large-scale analysis of water stability in bromodomain binding pockets with grand canonical Monte Carlo," *Commun. Chem.*, vol. 1, no. 1, p. 19, 2018/04/05 2018, doi: 10.1038/s42004-018-0019-x.
[274] M. Brand *et al.*, "Controlling intramolecular interactions in the design of selective, high-affinity ligands for the CREBBP bromodomain," *J. Med. Chem.*, vol. 64, no. 14, pp. 10102-10123, 2021/07/22 2021, doi: 10.1021/acs.jmedchem.1c00348.
[275] A. Dickson, "Mapping the Ligand Binding Landscape," *Biophysical Journal*, vol. 115, no. 9, pp. 1707-1719, 2018/11/06/ 2018, doi: 10.1016/j.bpj.2018.09.021.
[276] E. E. Guest, L. F. Cervantes, S. D. Pickett, C. L. Brooks, and J. D. Hirst, "Alchemical free energy methods applied to complexes of the first bromodomain of BRD4," *J. Chem. Inform. Model.*, vol. 62, no. 6, pp. 1458-1470, 2022/03/28 2022, doi: 10.1021/acs.jcim.1c01229.
[277] L. E. Jennings *et al.*, "BET bromodomain ligands: Probing the WPF shelf to improve BRD4 bromodomain affinity and metabolic stability," *Bioorg. Med. Chem.*, vol. 26, pp. 2937-2957, 2018, doi: 10.1016/j.bmc.2018.05.003.
[278] C. M. C. Laurin *et al.*, "Fragment-based identification of ligands for bromodomain-containing factor 3 of Trypanosoma cruzi," *ACS Infect. Dis.*, 2020/11/17 2020, doi: 10.1021/acsinfecdis.0c00618.
[279] L. Raich, K. Meier, J. Günther, C. D. Christ, F. Noe, and S. Olsson, "Discovery of a hidden transient state in all bromodomain families," *bioRxiv*, p. 2020.04.01.019547, 2020, doi: 10.1101/2020.04.01.019547. bioRxiv.
[280] S. Wan *et al.*, "Rapid and reliable binding affinity prediction of bromodomain inhibitors: A computational study," *J. Chem. Theory Comput.*, vol. 13, no. 2, pp. 784-795, 2017/02/14 2017, doi: 10.1021/acs.jctc.6b00794.
[281] Z. Zhang, F. Liu, and J. Chen, "Conformational Changes of CFTR upon Phosphorylation and ATP Binding," *Cell*, Article vol. 170, no. 3, pp. 483-+, Jul 27 2017, doi: 10.1016/j.cell.2017.06.041.
[282] T. W. H. Backman, Y. Cao, and T. Girke, "ChemMine tools: an online service for analyzing and clustering small molecules," *Nucleic Acids Research*, Article vol. 39, pp. W486-W491, Jul 2011, doi: 10.1093/nar/gkr320.
[283] Y. Cao, A. Charisi, L.-C. Cheng, T. Jiang, and T. Girke, "ChemmineR: a compound mining framework for R," *Bioinformatics*, Article vol. 24, no. 15, pp. 1733-1734, Aug 1 2008, doi: 10.1093/bioinformatics/btn307.
[284] E. F. Pettersen *et al.*, "UCSF chimera - A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, Article vol. 25, no. 13, pp. 1605-1612, Oct 2004, doi: 10.1002/jcc.20084.
[285] S. Jo *et al.*, "CHARMM-GUI 10 years for biomolecular modeling and simulation," *J. Comp. Chem.*, vol. 38, no. 15, pp. 1114-1124, 2017/06/05 2017, doi: 10.1002/jcc.24660.
[286] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859-1865, 2008/08/01 2008, doi: 10.1002/jcc.20945.
[287] X. He, V. H. Man, W. Yang, T.-S. Lee, and J. Wang, "A fast and high-quality charge model for the next generation general AMBER force field," *J. Chem. Phys.*, vol. 153, no. 11, p. 114502, 2020/09/21 2020, doi: 10.1063/5.0019056.

[288] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *Journal of Chemical Theory and Computation,* Article vol. 11, no. 8, pp. 3696-3713, Aug 2015, doi: 10.1021/acs.jctc.5b00255.

[289] R. M. Parrish *et al.*, "Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability," *Journal of Chemical Theory and Computation,* vol. 13, no. 7, pp. 3185-3197, 2017/07/11 2017, doi: 10.1021/acs.jctc.7b00174.

[290] R. Galvelis, S. Doerr, J. M. Damas, M. J. Harvey, and G. De Fabritiis, "A Scalable Molecular Force Field Parameterization Method Based on Density Functional Theory and Quantum-Level Machine Learning," *Journal of Chemical Information and Modeling,* Article vol. 59, no. 8, pp. 3485-3493, Aug 2019, doi: 10.1021/acs.jcim.9b00439.

[291] S. Pall *et al.*, "Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS," *J. Chem. Phys.,* Article vol. 153, no. 13, Oct 7 2020, Art no. 134110, doi: 10.1063/5.0018516.

[292] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation.," *Comp. Phys. Comm.,* vol. 95, pp. 43-56, 1995.

[293] J. Huang *et al.*, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nat. Methods.,* vol. 14, no. 1, pp. 71-73, 2017/01/01 2017, doi: 10.1038/nmeth.4067.

[294] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen, "Improved peptide and protein torsional energetics with the OPLS-AA force field," *J. Chem. Theor. Comput.,* vol. 11, no. 7, pp. 3499-3509, 2015/07/14 2015, doi: 10.1021/acs.jctc.5b00356.

[295] M. Hügle *et al.*, "4-Acyl pyrroles as dual BET-BRD7/9 bromodomain inhibitors address BETi insensitive human cancer cell lines," *J. Med. Chem.,* vol. 63, no. 24, pp. 15603-15620, 2020/12/24 2020, doi: 10.1021/acs.jmedchem.0c00478.

[296] M. Zhang *et al.*, "Structure-based discovery and optimization of benzo[d]isoxazole derivatives as potent and selective BET inhibitors for potential treatment of castration-resistant prostate cancer (CRPC)," *J. Med. Chem.,* vol. 61, no. 7, pp. 3037-3058, 2018/04/12 2018, doi: 10.1021/acs.jmedchem.8b00103.

[297] T. Fiala, K. Sleziakova, K. Marsalek, K. Salvadori, and V. Sindelar, "Thermodynamics of Halide Binding to a Neutral Bambusuril in Water and Organic Solvents," *J. Org. Chem.,* Article vol. 83, no. 4, pp. 1903-1912, Feb 16 2018, doi: 10.1021/acs.joc.7b02846.

[298] V. H. Le, M. Yanney, M. McGuire, A. Sygula, and E. A. Lewis, "Thermodynamics of host–guest interactions between fullerenes and a buckycatcher," *J. Phys. Chem. B,* vol. 118, no. 41, pp. 11956-11964, 2014/10/16 2014, doi: 10.1021/jp5087152.

[299] G. Bertin, "Calculation of Binding Free Energy via Alchemical Transformations," Padua, 2020. [Online]. Available: https://thesis.unipd.it/handle/20.500.12608/21280

[300] G. Poncet-Montange *et al.*, "Observed bromodomain flexibility reveals histone peptide- and small molecule ligand-compatible forms of ATAD2," *Biochem. J.,* vol. 466, no. 2, pp. 337-346, 2015, doi: 10.1042/BJ20140933.

[301] H. Huang *et al.*, "Solution structure of the second bromodomain of Brd2 and its specific interaction with acetylated histone tails," *BMC Struct. Biol.,* vol. 7, no. 1, p. 57, 2007/09/12 2007, doi: 10.1186/1472-6807-7-57.

[302] C. Cheng, H. Diao, F. Zhang, Y. Wang, K. Wang, and R. Wu, "Deciphering the mechanisms of selective inhibition for the tandem BD1/BD2 in the BET-bromodomain family," *Physical Chemistry Chemical Physics*, Article vol. 19, no. 35, pp. 23934-23941, Sep 21 2017, doi: 10.1039/c7cp04608a.
[303] J. Zhu, C. Zhou, and A. Caflisch, "Structure-based discovery of selective BRPF1 bromodomain inhibitors," *Eur. J. Med. Chem.*, vol. 155, pp. 337-352, 2018/07/15/ 2018, doi: 10.1016/j.ejmech.2018.05.037.
[304] S. Steiner, A. Magno, D. Huang, and A. Caflisch, "Does bromodomain flexibility influence histone recognition?," *Febs Letters*, Article vol. 587, no. 14, pp. 2158-2163, Jul 11 2013, doi: 10.1016/j.febslet.2013.05.032.
[305] F. M. Ferguson *et al.*, "Binding Hotspots of BAZ2B Bromodomain: Histone Interaction Revealed by Solution NMR Driven Docking," *Biochemistry*, Article vol. 53, no. 42, pp. 6706-6716, Oct 28 2014, doi: 10.1021/bi500909d.
[306] N. H. Theodoulou *et al.*, "Discovery of I-BRD9, a selective cell active chemical probe for bromodomain containing protein 9 inhibition," *J. Med. Chem.*, vol. 59, no. 4, pp. 1425-1439, 2016/02/25 2016, doi: 10.1021/acs.jmedchem.5b00256.
[307] W. Shen *et al.*, "Solution structure of human Brg1 bromodomain and its specific binding to acetylated histone tails," *Biochemistry*, Article vol. 46, no. 8, pp. 2100-2110, Feb 27 2007, doi: 10.1021/bi0611208.
[308] L. R. Vidler, N. Brown, S. Knapp, and S. Hoelder, "Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites," *J. Med. Chem.*, vol. 55, no. 17, pp. 7346-7359, 2012/09/13 2012, doi: 10.1021/jm300346w.
[309] E. E. Guest, S. D. Pickett, and J. D. Hirst, "Structural variation of protein-ligand complexes of the first bromodomain of BRD4," *Organic & Biomolecular Chemistry*, Article vol. 19, no. 25, pp. 5632-5641, Jul 7 2021, doi: 10.1039/d1ob00658d.
[310] S. J. Eron *et al.*, "Structural characterization of degrader-induced ternary complexes using hydrogen–deuterium exchange mass spectrometry and computational modeling: Implications for structure-based design," *ACS Chem. Biol.*, vol. 16, no. 11, pp. 2228-2243, 2021/11/19 2021, doi: 10.1021/acschembio.1c00376.
[311] C. Langini, A. Caflisch, and A. Vitalis, "The ATAD2 bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes," *J. Biol. Chem.*, vol. 292, no. 40, pp. 16734-16745, 2017/10/06/ 2017, doi: 10.1074/jbc.M117.786350.
[312] J. T. Lloyd *et al.*, "Structural insights into the recognition of mono- and diacetylated histones by the ATAD2B bromodomain," *J. Med. Chem.*, vol. 63, no. 21, pp. 12799-12813, 2020/11/12 2020, doi: 10.1021/acs.jmedchem.0c01178.
[313] Y. Zhou, M. Hussain, G. Kuang, J. Zhang, and Y. Tu, "Mechanistic insights into peptide and ligand binding of the ATAD2-bromodomain via atomistic simulations disclosing a role of induced fit and conformational selection," *Physical Chemistry Chemical Physics*, Article vol. 20, no. 36, pp. 23222-23232, Sep 28 2018, doi: 10.1039/c8cp03860k.
[314] H. Sun *et al.*, "Solution structure of BRD7 bromodomain and its interaction with acetylated peptides from histone H3 and H4," *Biochem. Biophys. Res. Comms.*, vol. 358, no. 2, pp. 435-441, 2007/06/29/ 2007, doi: 10.1016/j.bbrc.2007.04.139.
[315] M. Kuang, J. Zhou, L. Wang, Z. Liu, J. Guo, and R. Wu, "Binding kinetics versus affinities in BRD4 inhibition," *J. Chem. Inf. Model.*, vol. 55, no. 9, pp. 1926-1935, 2015/09/28 2015, doi: 10.1021/acs.jcim.5b00265.

[316] J. Su, X. Liu, S. Zhang, F. Yan, Q. Zhang, and J. Chen, "A computational insight into binding modes of inhibitors XD29, XD35, and XD28 to bromodomain-containing protein 4 based on molecular dynamics simulations," *J. Biomol. Struct. Dyn.*, vol. 36, no. 5, pp. 1212-1224, 2018/04/04 2018, doi: 10.1080/07391102.2017.1317666.

[317] R. Tumdam, A. Kumar, N. Subbarao, and B. S. Balaji, "In silico study directed towards identification of novel high-affinity inhibitors targeting an oncogenic protein: BRD4-BD1," *SAR QSAR Env. Res.*, vol. 29, no. 12, pp. 975-996, 2018/12/02 2018, doi: 10.1080/1062936X.2018.1537301.

[318] X. Zhang, K. Chen, Y.-D. Wu, and O. Wiest, "Protein dynamics and structural waters in bromodomains," *PLOS ONE*, vol. 12, no. 10, p. e0186570, 2017, doi: 10.1371/journal.pone.0186570.

[319] G. Klebe, "Broad-scale analysis of thermodynamic signatures in medicinal chemistry: are enthalpy-favored binders the better development option?," (in English), *Drug Discovery Today*, Review vol. 24, no. 4, pp. 943-948, Apr 2019, doi: 10.1016/j.drudis.2019.01.014.

[320] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986-998, 2011.

[321] Y. Fukao, "Protein–protein interactions in plants," *Plant and Cell Physiology*, vol. 53, no. 4, pp. 617-625, 2012.

[322] S. Hayes, B. Malacrida, M. Kiely, and P. A. Kiely, "Studying protein-protein interactions: progress, pitfalls and solutions," *Biochemical Society Transactions*, Review vol. 44, pp. 994-1004, Aug 15 2016, doi: 10.1042/bst20160092.

[323] U. Kuzmanov and A. Emili, "Protein-protein interaction networks: probing disease mechanisms using model systems," *Genome medicine*, vol. 5, no. 4, pp. 1-12, 2013.

[324] M. Rosell and J. Fernandez-Recio, "Hot-spot analysis for drug discovery targeting protein-protein interactions," *Expert Opinion on Drug Discovery*, Review vol. 13, no. 4, pp. 327-338, 2018 2018, doi: 10.1080/17460441.2018.1430763.

[325] P. Watkins, "Drug safety sciences and the bottleneck in drug development," *Clinical Pharmacology & Therapeutics*, vol. 89, no. 6, pp. 788-790, 2011.

[326] M. Gatti and F. De Ponti, "Drug repurposing in the COVID-19 era: Insights from case studies showing pharmaceutical peculiarities," *Pharmaceutics*, vol. 13, no. 3, p. 302, 2021.

[327] C.-k. Chang *et al.*, "Targeting protein-protein interaction interfaces in COVID-19 drug discovery," *Computational and Structural Biotechnology Journal*, Article vol. 19, pp. 2246-2255, 2021 2021, doi: 10.1016/j.csbj.2021.04.003.

[328] N. Tonali, S. Nencetti, E. Orlandini, and L. Ciccone, "Application of PROTAC strategy to TTR-Aβ protein-protein interaction for the development of Alzheimer's disease drugs," *Neural Regeneration Research*, vol. 16, no. 8, p. 1554, 2021.

[329] K. B. Karunakaran, N. Yanamala, G. Boyce, M. J. Becich, and M. K. Ganapathiraju, "Malignant pleural mesothelioma interactome with 364 novel protein-protein interactions," *Cancers*, vol. 13, no. 7, p. 1660, 2021.

[330] S. Patel *et al.*, "Pyruvate kinase M2 in chronic inflammations: a potpourri of crucial protein–protein interactions," *Cell Biology and Toxicology*, vol. 37, no. 5, pp. 653-678, 2021.

[331] Y. Gu and D. Zhu, "nNOS-mediated protein-protein interactions: promising targets for treating neurological and neuropsychiatric disorders," *Journal of Biomedical Research*, vol. 35, no. 1, p. 1, 2021.

[332] F. Lasala *et al.*, "Identification of potential inhibitors of protein-protein interaction useful to fight against Ebola and other highly pathogenic viruses," *Antiviral research,* vol. 186, p. 105011, 2021.
[333] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling," *Nature chemistry,* vol. 9, no. 10, pp. 1005-1011, 2017.
[334] C. Tse, L. Wickstrom, M. Kvaratskhelia, E. Gallicchio, R. Levy, and N. Deng, "Exploring the Free-Energy Landscape and Thermodynamics of Protein-Protein Association," *Biophysical Journal,* Article vol. 119, no. 6, pp. 1226-1238, Sep 15 2020, doi: 10.1016/j.bpj.2020.08.005.
[335] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov, and J. Åqvist, "Free energy calculations and ligand binding," *Advances in protein chemistry,* vol. 66, pp. 123-158, 2003.
[336] A. Pohorille, C. Jarzynski, and C. Chipot, "Good practices in free-energy calculations," *The Journal of Physical Chemistry B,* vol. 114, no. 32, pp. 10235-10253, 2010.
[337] M. Lapelosa, E. Gallicchio, and R. M. Levy, "Conformational transitions and convergence of absolute binding free energy calculations," *Journal of Chemical Theory and Computation,* vol. 8, no. 1, pp. 47-60, 2012.
[338] A. P. Bhati and P. V. Coveney, "Large Scale Study of Ligand–Protein Relative Binding Free Energy Calculations: Actionable Predictions from Statistically Robust Protocols," *Journal of Chemical Theory and Computation,* vol. 18, no. 4, pp. 2687-2702, 2022.
[339] R. B. Russell *et al.*, "A structural perspective on protein–protein interactions," *Current opinion in structural biology,* vol. 14, no. 3, pp. 313-324, 2004.
[340] M. M. Pierce, C. Raman, and B. T. Nall, "Isothermal titration calorimetry of protein–protein interactions," *Methods,* vol. 19, no. 2, pp. 213-221, 1999.
[341] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, "Conformational entropy in molecular recognition by proteins," *Nature,* vol. 448, no. 7151, pp. 325-329, 2007.
[342] Z. Liu *et al.*, "PDB-wide collection of binding data: current status of the PDBbind database," *Bioinformatics,* Article vol. 31, no. 3, pp. 405-412, Feb 1 2015, doi: 10.1093/bioinformatics/btu626.
[343] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, "The PDBbind database: methodologies and updates," *Journal of medicinal chemistry,* vol. 48, no. 12, pp. 4111-4119, 2005.
[344] H. M. Berman *et al.*, "The protein data bank," *Nucleic acids research,* vol. 28, no. 1, pp. 235-242, 2000.
[345] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular systems biology,* vol. 7, no. 1, p. 539, 2011.
[346] A. R. Ortiz, C. E. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison," *Protein Science,* vol. 11, no. 11, pp. 2606-2621, 2002.
[347] A. Herbert and M. Sternberg, "MaxCluster: a tool for protein structure comparison and clustering," ed, 2008.
[348] Z. Yu, P. Li, and K. M. Merz, Jr., "Extended Zinc AMBER Force Field (EZAFF)," *Journal of Chemical Theory and Computation,* Article vol. 14, no. 1, pp. 242-254, Jan 2018, doi: 10.1021/acs.jctc.7b00773.

[349] M. B. Peters, Y. Yang, B. Wang, L. Fuesti-Molnar, M. N. Weaver, and K. M. Merz, Jr., "Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF)," *Journal of Chemical Theory and Computation*, Article vol. 6, no. 9, pp. 2935-2947, Sep 2010, doi: 10.1021/ct1002626.
[350] J. W. Ponder and F. M. Richards, "An efficient newton-like method for molecular mechanics energy minimization of large molecules," *Journal of computational chemistry*, vol. 8, no. 7, pp. 1016-1024, 1987.
[351] S. Xie, Y. Lu, J. Jakoncic, H. Sun, J. Xia, and C. Qian, "Structure of RPA 32 bound to the N-terminus of SMARCAL 1 redefines the binding interface between RPA 32 and its interacting proteins," *The FEBS journal*, vol. 281, no. 15, pp. 3382-3396, 2014.
[352] T. G. M. Schmidt, J. Koepke, R. Frank, and A. Skerra, "Molecular interaction between the Strep-tag affinity peptide and its cognate target, streptavidin," *Journal of Molecular Biology*, Article vol. 255, no. 5, pp. 753-766, Feb 9 1996, doi: 10.1006/jmbi.1996.0061.
[353] Z. Liu and H. J. Vogel, "Structural basis for the regulation of L-type voltage-gated calcium channels: interactions between the N-terminal cytoplasmic domain and Ca2+-calmodulin," *Frontiers in Molecular Neuroscience*, Article vol. 5, 2012 2012, Art no. 38, doi: 10.3389/fnmol.2012.00038.
[354] T. K. Yu *et al.*, "An unusual protein–protein interaction through coupled unfolding and binding," *Angewandte Chemie*, vol. 126, no. 37, pp. 9942-9945, 2014.
[355] C. R. Grace *et al.*, "Monitoring ligand-induced protein ordering in drug discovery," *Journal of molecular biology*, vol. 428, no. 6, pp. 1290-1303, 2016.
[356] S. Eulitz *et al.*, "Identification of Xin-repeat proteins as novel ligands of the SH3 domains of nebulin and nebulette and analysis of their interaction during myofibril formation and remodeling," *Molecular Biology of the Cell*, Article vol. 24, no. 20, pp. 3215-3226, Oct 15 2013, doi: 10.1091/mbc.E13-04-0202.
[357] C. Tallant *et al.*, "Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger and Bromodomain of the Chromatin Remodeling Complex NoRC," *Structure*, Article vol. 23, no. 1, pp. 80-92, Jan 6 2015, doi: 10.1016/j.str.2014.10.017.
[358] S. Clark, A. Nyarko, F. Loehr, P. A. Karplus, and E. Barbar, "The Anchored Flexibility Model in LC8 Motif Recognition: Insights from the Chica Complex," *Biochemistry*, Article vol. 55, no. 1, pp. 199-209, Jan 12 2016, doi: 10.1021/acs.biochem.5b01099.
[359] S. K. Ponna, S. Ruskamo, M. Myllykoski, C. Keller, T. M. Boeckers, and P. Kursula, "Structural basis for PDZ domain interactions in the post-synaptic density scaffolding protein Shank3," *Journal of Neurochemistry*, Article vol. 145, no. 6, pp. 449-463, Jun 2018, doi: 10.1111/jnc.14322.
[360] A. V. Murthy *et al.*, "Structural enzymology binding studies of the peptide-substrate-binding domain of human collagen prolyl 4-hydroxylase (type-II): High affinity peptides have a PxGP sequence motif," *Protein Science*, Article vol. 27, no. 9, pp. 1692-1703, Sep 2018, doi: 10.1002/pro.3450.
[361] J. Huber *et al.*, "An atypical LIR motif within UBA5 (ubiquitin like modifier activating enzyme 5) interacts with GABARAP proteins and mediates membrane localization of UBA5," *Autophagy*, Article vol. 16, no. 2, pp. 256-270, 2020 2020, doi: 10.1080/15548627.2019.1606637.
[362] M. Wade, Y.-C. Li, and G. M. Wahl, "MDM2, MDMX and p53 in oncogenesis and cancer therapy," *Nature Reviews Cancer*, vol. 13, no. 2, pp. 83-96, 2013.

[363] K. Takemura, N. Matubayasi, and A. Kitao, "Binding free energy analysis of protein-protein docking model structures by evERdock," *The Journal of chemical physics,* vol. 148, no. 10, p. 105101, 2018.
[364] J. W. Perthold and C. Oostenbrink, "Simulation of reversible protein–protein binding and calculation of binding free energies using perturbed distance restraints," *Journal of chemical theory and computation,* vol. 13, no. 11, pp. 5697-5708, 2017.
[365] V. M. Krishnamurthy, B. R. Bohall, V. Semetey, and G. M. Whitesides, "The paradoxical thermodynamic basis for the interaction of ethylene glycol, glycine, and sarcosine chains with bovine carbonic anhydrase II: An unexpected manifestation of enthalpy/entropy compensation," *Journal of the American Chemical Society,* Article vol. 128, no. 17, pp. 5802-5812, May 3 2006, doi: 10.1021/ja060070r.
[366] M. D. Tyka *et al.,* "Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping," *Journal of Molecular Biology,* Article vol. 405, no. 2, pp. 607-618, Jan 14 2011, doi: 10.1016/j.jmb.2010.11.008.
[367] A. Gershenson, L. M. Gierasch, A. Pastore, and S. E. Radford, "Energy landscapes of functional proteins are inherently risky," *Nature Chemical Biology,* Article vol. 10, no. 11, pp. 884-891, Nov 2014, doi: 10.1038/nchembio.1670.
[368] H. Fu, Y. Zhou, X. Jing, X. Shao, and W. Cai, "Meta-Analysis Reveals That Absolute Binding Free-Energy Calculations Approach Chemical Accuracy," *Journal of Medicinal Chemistry,* Article; Early Access 2022, doi: 10.1021/acs.jmedchem.2c00796.
[369] D. Suárez and N. Díaz, "Conformational and entropy analyses of extended molecular dynamics simulations of α-, β-and γ-cyclodextrins and of the β-cyclodextrin/nabumetone complex," *Physical Chemistry Chemical Physics,* vol. 19, no. 2, pp. 1431-1440, 2017.
[370] B. K. Ho and K. A. Dill, "Folding very short peptides using molecular dynamics," *PLoS computational biology,* vol. 2, no. 4, p. e27, 2006.
[371] M. Sattler *et al.,* "Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis," *Science,* vol. 275, no. 5302, pp. 983-986, 1997.
[372] A. M. Petros *et al.,* "Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies," *Protein Science,* vol. 9, no. 12, pp. 2528-2534, 2000.
[373] M. V. E. Botuyan, J. Momand, and Y. Chen, "Solution conformation of an essential region of the p53 transactivation domain," *Folding and Design,* vol. 2, no. 6, pp. 331-342, 1997.
[374] H. Lee *et al.,* "Local structural elements in the mostly unstructured transcriptional activation domain of human p53," *Journal of Biological Chemistry,* vol. 275, no. 38, pp. 29426-29432, 2000.
[375] R. Dawson, L. Müller, A. Dehner, C. Klein, H. Kessler, and J. Buchner, "The N-terminal domain of p53 is natively unfolded," *Journal of molecular biology,* vol. 332, no. 5, pp. 1131-1141, 2003.
[376] P. H. Kussie *et al.,* "Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain," *Science,* vol. 274, no. 5289, pp. 948-953, 1996.
[377] V. BoÈttger *et al.,* "Comparative study of the p53-mdm2 and p53-MDMX interfaces," *Oncogene,* vol. 18, no. 1, pp. 189-199, 1999.
[378] C. W. Lee, M. A. Martinez-Yamout, H. J. Dyson, and P. E. Wright, "Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator

binding domain of CREB binding protein," *Biochemistry,* vol. 49, no. 46, pp. 9964-9971, 2010.
[379] N. London, D. Movshovitz-Attias, and O. Schueler-Furman, "The Structural Basis of Peptide-Protein Binding Strategies," *Structure,* Article vol. 18, no. 2, pp. 188-199, Feb 10 2010, doi: 10.1016/j.str.2009.11.012.
[380] Q. Zhang, S. X. Zeng, and H. Lu, "Targeting p53-MDM2-MDMX loop for cancer therapy," *Mutant p53 and MDM2 in Cancer,* pp. 281-319, 2014.
[381] D. Spiegelberg, A. C. Mortensen, S. Lundsten, C. J. Brown, D. P. Lane, and M. Nestor, "The MDM2/MDMX-p53 antagonist PM2 radiosensitizes wild-type p53 tumors," *Cancer research,* vol. 78, no. 17, pp. 5084-5093, 2018.
[382] G. M. Popowicz, A. Dömling, and T. A. Holak, "The structure-based design of MDM2/MDMX–p53 inhibitors gets serious," *Angewandte Chemie International Edition,* vol. 50, no. 12, pp. 2680-2688, 2011.
[383] N. Soudah *et al.,* "An N-terminal extension to UBA5 adenylation domain boosts UFM1 activation: Isoform-specific differences in ubiquitin-like protein activation," *Journal of molecular biology,* vol. 431, no. 3, pp. 463-478, 2019.
[384] W. Oweis *et al.,* "Trans-binding mechanism of ubiquitin-like protein activation revealed by a UBA5-UFM1 complex," *Cell reports,* vol. 16, no. 12, pp. 3113-3120, 2016.
[385] R. B. Best, N. V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?," (in eng), *Biophys J,* vol. 95, no. 1, pp. L07-9, Jul 2008, doi: 10.1529/biophysj.108.132696.
[386] M. D. Polêto and J. A. Lemkul, "Integration of Experimental Data and Use of Automated Fitting Methods in Developing Protein Force Fields," (in eng), *Commun Chem,* vol. 5, 2022, doi: 10.1038/s42004-022-00653-z.
[387] G. A. Holdgate, "Making cool drugs hot: isothermal titration calorimetry as a tool to study binding energetics," *Biotechniques,* vol. 31, no. 1, pp. 164-6, 168, 170 passim, 2001.
[388] E. Harper and J. Black, "Histamine H3-receptor agonists and imidazole-based H3-receptor antagonists can be thermodynamically discriminated," *British journal of pharmacology,* vol. 151, no. 4, pp. 504-517, 2007.
[389] X. Zheng, W. Cheng, C. Ji, J. Zhang, and M. Yin, "Detection of metal ions in biological systems: A review," *Reviews in Analytical Chemistry,* Review vol. 39, no. 1, pp. 231-246, Jan 2020, doi: 10.1515/revac-2020-0118.
[390] A. J. Thomson and H. B. Gray, "Bio-inorganic chemistry," *Current opinion in chemical biology,* vol. 2, no. 2, pp. 155-158, 1998.
[391] K. J. Waldron, J. C. Rutherford, D. Ford, and N. J. Robinson, "Metalloproteins and metal sensing," *Nature,* vol. 460, no. 7257, pp. 823-830, 2009.
[392] S. M. Yannone, S. Hartung, A. L. Menon, M. W. Adams, and J. A. Tainer, "Metals in biology: defining metalloproteomes," *Current opinion in biotechnology,* vol. 23, no. 1, pp. 89-95, 2012.
[393] R. H. Stote and M. Karplus, "Zinc binding in proteins and solution: a simple but accurate nonbonded representation," *Proteins: Structure, Function, and Bioinformatics,* vol. 23, no. 1, pp. 12-31, 1995.
[394] P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz Jr, "Rational design of particle mesh Ewald compatible Lennard-Jones parameters for+ 2 metal cations in explicit solvent," *Journal of chemical theory and computation,* vol. 9, no. 6, pp. 2733-2748, 2013.

[395] P. Li and K. M. Merz, Jr., "Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions," *Journal of Chemical Theory and Computation,* Article vol. 10, no. 1, pp. 289-297, Jan 2014, doi: 10.1021/ct400751u.
[396] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, "High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing," *Journal of Chemical Information and Modeling,* vol. 50, no. 3, pp. 397-403, 2010/03/22 2010, doi: 10.1021/ci900455r.
[397] M. R. Shirts, D. L. Mobley, and J. D. Chodera, "Alchemical Free Energy Calculations: Ready for Prime Time?," in *Annual Reports in Computational Chemistry, Vol 3*, vol. 3, D. C. Spellmeyer and R. A. Wheeler Eds., (Annual Reports in Computational Chemistry, 2007, pp. 41-59.
[398] J. Yin, A. T. Fenley, N. M. Henriksen, and M. K. Gilson, "Toward Improved Force-Field Accuracy through Sensitivity Analysis of Host-Guest Binding Thermodynamics," *The Journal of Physical Chemistry B,* vol. 119, no. 32, pp. 10145-10155, 2015/08/13 2015, doi: 10.1021/acs.jpcb.5b04262.
[399] Y. Khalak, B. Baumeier, and M. Karttunen, "Improved general-purpose five-point model for water: TIP5P/2018," *The Journal of Chemical Physics,* vol. 149, no. 22, p. 224507, 2018/12/14 2018, doi: 10.1063/1.5070137.
[400] C.-L. Zhao, D.-X. Zhao, C.-C. Bei, X.-N. Meng, S. Li, and Z.-Z. Yang, "Seven-Site Effective Pair Potential for Simulating Liquid Water," *The Journal of Physical Chemistry B,* vol. 123, no. 21, pp. 4594-4603, 2019/05/30 2019, doi: 10.1021/acs.jpcb.9b03149.
[401] C. Lu *et al.*, "OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space," *Journal of Chemical Theory and Computation,* vol. 17, no. 7, pp. 4291-4300, 2021/07/13 2021, doi: 10.1021/acs.jctc.1c00302.
[402] T. Siebenmorgen and M. Zacharias, "Evaluation of Predicted Protein–Protein Complexes by Binding Free Energy Simulations," *Journal of Chemical Theory and Computation,* vol. 15, no. 3, pp. 2071-2086, 2019/03/12 2019, doi: 10.1021/acs.jctc.8b01022.
[403] S. Izadi and A. V. Onufriev, "Accuracy limit of rigid 3-point water models," *The Journal of Chemical Physics,* vol. 145, no. 7, p. 074501, 2016/08/21 2016, doi: 10.1063/1.4960175.
[404] J. W. Ponder *et al.*, "Current Status of the AMOEBA Polarizable Force Field," *The Journal of Physical Chemistry B,* vol. 114, no. 8, pp. 2549-2564, 2010/03/04 2010, doi: 10.1021/jp910674d.
[405] S. Azimi, S. Khuttan, J. Z. Wu, R. K. Pal, and E. Gallicchio, "Relative Binding Free Energy Calculations for Ligands with Diverse Scaffolds with the Alchemical Transfer Method," *Journal of Chemical Information and Modeling,* vol. 62, no. 2, pp. 309-323, 2022/01/24 2022, doi: 10.1021/acs.jcim.1c01129.
[406] N. Singh and W. Li, "Absolute Binding Free Energy Calculations for Highly Flexible Protein MDM2 and Its Inhibitors," *International Journal of Molecular Sciences,* vol. 21, no. 13, p. 4765, 2020.
[407] M. Shiroishi *et al.*, "Structural Consequences of Mutations in Interfacial Tyr Residues of a Protein Antigen-Antibody Complex: THE CASE OF HyHEL-10-HEL*," *Journal of Biological Chemistry,* vol. 282, no. 9, pp. 6783-6791, 2007/03/02/ 2007, doi: 10.1074/jbc.M605197200.
[408] F. Cozzolino, I. Iacobucci, V. Monaco, and M. Monti, "Protein–DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era," *Journal of Proteome Research,* vol. 20, no. 6, pp. 3018-3030, 2021/06/04 2021, doi: 10.1021/acs.jproteome.1c00074.

[409] P. d. C. Monroig, L. Chen, S. Zhang, and G. A. Calin, "Small molecule compounds targeting miRNAs for cancer therapy," *Advanced Drug Delivery Reviews,* vol. 81, pp. 104-116, 2015/01/01/ 2015, doi: 10.1016/j.addr.2014.09.002.
[410] M. Wang, Y. Yu, C. Liang, A. Lu, and G. Zhang, "Recent Advances in Developing Small Molecules Targeting Nucleic Acid," *International Journal of Molecular Sciences,* vol. 17, no. 6, p. 779, 2016.
[411] L. R. Rutledge, H. F. Durst, and S. D. Wetmore, "Evidence for Stabilization of DNA/RNA−Protein Complexes Arising from Nucleobase−Amino Acid Stacking and T-Shaped Interactions," *Journal of Chemical Theory and Computation,* vol. 5, no. 5, pp. 1400-1410, 2009/05/12 2009, doi: 10.1021/ct800567q.
[412] P. Filippakopoulos *et al.*, "Selective inhibition of BET bromodomains," *Nature,* vol. 468, no. 7327, pp. 1067-1073, 2010, doi: 10.1038/nature09504.
[413] M. G. J. Baud *et al.*, "A bump-and-hole approach to engineer controlled selectivity of BET bromodomain chemical probes," *Science,* vol. 346, no. 6209, pp. 638-641, 2014, doi: 10.1126/science.1249830.
[414] P. Filippakopoulos *et al.*, "Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family," *Bioorganic & Medicinal Chemistry,* vol. 20, no. 6, pp. 1878-1886, 2012, doi: 10.1016/j.bmc.2011.10.080.
[415] S. Picaud *et al.*, "PFI-1, a Highly Selective Protein Interaction Inhibitor, Targeting BET Bromodomains," *Cancer Research,* vol. 73, no. 11, pp. 3336-3346, 2013, doi: 10.1158/0008-5472.can-12-3292.
[416] S. Picaud *et al.*, "RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain," *Proceedings of the National Academy of Sciences,* vol. 110, no. 49, pp. 19754-19759, 2013, doi: 10.1073/pnas.1310658110.
[417] K. G. McLure *et al.*, "RVX-208, an Inducer of ApoA-I in Humans, Is a BET Bromodomain Antagonist," *PLoS ONE,* vol. 8, no. 12, p. e83190, 2013, doi: 10.1371/journal.pone.0083190.
[418] P. Ciceri *et al.*, "Dual kinase-bromodomain inhibitors for rationally designed polypharmacology," *Nature Chemical Biology,* vol. 10, no. 4, pp. 305-312, 2014, doi: 10.1038/nchembio.1471.
[419] M. Gacias *et al.*, "Selective Chemical Modulation of Gene Transcription Favors Oligodendrocyte Lineage Progression," *Chemistry & Biology,* vol. 21, no. 7, pp. 841-854, 2014, doi: 10.1016/j.chembiol.2014.05.009.
[420] S. Picaud *et al.*, "9H-Purine Scaffold Reveals Induced-Fit Pocket Plasticity of the BRD9 Bromodomain," *Journal of Medicinal Chemistry,* vol. 58, no. 6, pp. 2718-2736, 2015/03/26 2015, doi: 10.1021/jm501893k.
[421] A. Hammitzsch *et al.*, "CBP30, a selective CBP/p300 bromodomain inhibitor, suppresses human Th17 responses," *Proceedings of the National Academy of Sciences,* vol. 112, no. 34, pp. 10768-10773, 2015, doi: 10.1073/pnas.1501956112.
[422] M. Hügle *et al.*, "4-Acyl pyrrole derivatives yield novel vectors for designing inhibitors of the acetyl-lysine recognition site of BRD4 (1)," *Journal of medicinal chemistry,* vol. 59, no. 4, pp. 1518-1530, 2016.
[423] O. A. Kharenko *et al.*, "RVX-297- a novel BD2 selective inhibitor of BET bromodomains," *Biochemical and Biophysical Research Communications,* vol. 477, no. 1, pp. 62-67, 2016/08/12/ 2016, doi: 10.1016/j.bbrc.2016.06.021.
[424] B. Raux *et al.*, "Exploring selective inhibition of the first bromodomain of the human bromodomain and extra-terminal domain (BET) proteins," *Journal of medicinal chemistry,* vol. 59, no. 4, pp. 1634-1641, 2016.

[425] R. C. Montenegro *et al.*, "BET inhibition as a new strategy for the treatment of gastric cancer," *Oncotarget*, vol. 7, no. 28, pp. 43997-44012, 2016, doi: 10.18632/oncotarget.9766.

[426] S. Picaud *et al.*, "Promiscuous targeting of bromodomains by bromosporine identifies BET proteins as master regulators of primary transcription response in leukemia," *Sci. Adv.*, vol. 2, no. 10, p. e1600760, doi: 10.1126/sciadv.1600760.

[427] S. Xie, Y. Lu, J. Jakoncic, H. Sun, J. Xia, and C. Qian, "Structure of RPA32 bound to the N-terminus of SMARCAL1 redefines the binding interface between RPA32 and its interacting proteins," *Febs Journal*, Article vol. 281, no. 15, pp. 3382-3396, Aug 2014, doi: 10.1111/febs.12867.