November 2023

ATOMISTIC SIMULATIONS OF INTRINSICALLY DISORDERED PROTEIN FOLDING AND DYNAMICS

Xiping Gong
University of Massachusetts Amherst

ATOMISTIC SIMULATIONS OF INTRINSICALLY DISORDERED PROTEIN FOLDING AND DYNAMICS

A Dissertation Presented

by

XIPING GONG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2023

Department of Chemistry

ATOMISTIC SIMULATIONS OF INTRINSICALLY DISORDERED PROTEIN FOLDING AND DYNAMICS

A Dissertation Presented

by

XIPING GONG

Approved as to style and content by:

_____
Jianhan Chen, Chair

_____
Scott Auerbach, Member

_____
Craig Martin, Member

_____
Matthew D. Moore, Member

_____
Ricardo Metz, Department Head
Department of Chemistry

DEDICATION

To my parents, wife, advisor and other people who helped me.

ACKNOWLEDGMENTS

Throughout my five-year Ph.D. career, I have received a great deal of support, assistance and guidance. First of all, I would like to give my special thanks to my advisor, Dr. Jianhan Chen. I remember that I had little knowledge of molecular dynamics five years ago, but it is he who first guided me to learn from the basis to be an expert, which is invaluable and undoubtedly enriches my knowledge in the field of computational chemistry. In addition, he is always requesting me to critically think and resolve the problems in the way of science, which enables me to build up sufficient skills and confidence in resolving my scientific challenges. Therefore, it is he who systematically trained me to be like a scientist, which will play a fundamental role in helping me to be a great professor in the future. I also really appreciate my dissertation committee members, Dr. Scott Auerbach, Dr. Craig Martin and Dr. Matthew D. Moore, for their academic guidance and insightful comments.

It is challenging for me to complete my first project. It was Kuo Hao Lee who gave me the scripts and showed me how to successfully run the GBMV2 implicit solvent simulations in CHARMM. It was Mara Chiricotto who taught me how to compile and implement the GPU-GBMV2/SA into the CHARMM/OpenMM program. It was Zhiguang Jia who helped me to set up my personal accounts and gave me countless technical assistance. It was Xiaorong Liu who allowed me to ask her many questions on implicit solvent models and replica exchange simulations. It was Mahdieh Yazdani who showed me her excellent BK project with a lot of patient, encouraged me to be more confident in many aspects, and helped me attend my first conference in the United States. I also gained a lot of help and knowledge by enjoying our journal club and monthly

research update from many people, including Shrishti Barethiya, Juni Campbell, Mara Chiricotto, David DoCoeur, Kairong Dong, Azar Farjamnia, Qianlan Jia, Zhiguang Jia, Aron Korsunsky, Kuo Hao Lee, Shanlong Li, Xiaorong Liu, Erik Nordquist, Samantha Schultz, Mahdieh Yazdani, etc. I also would like to give my additional thanks to the people who contributed or revised the articles I published, especially Yumeng Zhang and Jianhan Chen. In addition, I gained a lot of fun from my friends in Dr. Chen's lab and felt so comfortable during this five-year period. I really enjoyed our lab activities to reduce the stress, like hiking, biking, eating, gaming, celebrating, cleaning, decorating, etc. I really appreciate your joining, especially Erik Nordquist who guided us to propose the plans and have fun together. I am also really glad to create our fitness team, and together go swimming, running, weightlifting, etc.

Finally, I would like to give my deepest appreciation to my family. It is my wife, Hualu Zhou, who is not only a great mother, but also a happy wife who is always encouraging me to overcome many challenges and bringing me a lot of confidence and support. I feel so happy that we had such a memorable period in our college and graduate school, which will become an invaluable experience we will often recall once we get old. Additionally, I would like to give special thanks to my parents who provided their valuable 5-year time to support us. Their selfless dedication undoubtedly plays a crucial role in the success of my Ph.D. graduation. Your help increased my happiness and reduced a lot of stress from me. Of course, I also want to give my lovely thanks to my classmates, teachers, and the people I met at UMass who helped and supported me, and the chemistry department and graduate school which provided me with financial support and a great learning environment.

ABSTRACT

ATOMISTIC SIMULATIONS OF INTRINSICALLY DISORDERED PROTEIN FOLDING AND DYNAMICS

SEPTEMBER 2023

XIPING GONG, B.S., NANCHANG UNIVERSITY

M.S., XIAMEN UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jianhan Chen

Intrinsically disordered proteins (IDPs) are crucial in biology and human diseases, necessitating a comprehensive understanding of their structure, dynamics, and interactions. Atomistic simulations have emerged as a key tool for unraveling the molecular intricacies and establishing mechanistic insights into how these proteins facilitate diverse biological functions. However, achieving accurate simulations requires both an appropriate protein force field capable of describing the energy landscape of functionally relevant IDP conformations and sufficient conformational sampling to capture the free energy landscape of IDP dynamics. These factors are fundamental in comprehending potential IDP structures, dynamics, and interactions.

I first conducted explicit solvent simulations to assess the performance of two state-of-the-art protein force fields, namely CHARMM36m and a99SB-disp, in capturing the stability of small protein-protein interactions. To evaluate their accuracy, I selected a set of 46 amino acid backbone and side chain pairs with representative configurations and computed the free energy profiles of their interactions. The results demonstrated that CHARMM36m consistently predicted stronger protein-protein interactions compared to

a99SB-disp. Notably, the most significant overestimation in CHARMM36m occurred in charged pairs involving Arg and Glu side chains, with an overestimation of up to 2.9 kcal/mol. Through free energy decomposition analysis, I determined that these overestimations were primarily driven by protein-water electrostatic interactions rather than van der Waals (vdW) interactions. Consequently, these findings suggest that careful rebalancing of electrostatic interactions should be considered in the further optimization of protein force fields.

In order to enhance the conformational sampling of IDPs, I developed an integrated approach that combines an improved implicit solvent model called Generalized Born with molecular volume and solvent accessible surface area (GBMV2/SA) with a multiscale enhanced sampling (MSES) technique. To make this approach more efficient, I implemented it as a standalone OpenMM plugin on Graphics Processing Units (GPUs). The results demonstrated that the GPU-GBMV2/SA model achieved numerical equivalence to the original CPU-GBMV2/SA models, while providing a remarkable ~60x speedup on a single NVIDIA TITAN X (Pascal) graphics card for molecular dynamic simulations of both folded and unstructured proteins. This significant acceleration greatly facilitated the application of the approach in biomolecular simulations.

In addition, I conducted an evaluation of the reliability of GBMV2/SA models in simulating both folded and unfolded proteins. The results revealed that the GBMV2/SA model accurately describes small proteins, but its applicability is limited when it comes to larger proteins such as KID and p53-TAD proteins. This limitation can be attributed to the absence of long-range solute-solvent dispersion interactions in the model. To address this issue, I introduced a comprehensive treatment of nonpolar solvation free energy

called GBMV2/NP model. Unfortunately, the GBMV2/NP model exhibited a destabilizing effect on well-folded proteins, particularly larger ones, due to an inaccurate representation of the repulsive solvent accessible surface area (SASA) model caused by the utilization of unphysical van der Waals volume. This observation highlights the need for further improvements in accurately describing the nonpolar term in the model.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION [1]

1.1 Intrinsically disordered proteins: structures, folding and dynamics

Intrinsically disordered proteins (IDPs) or regions (IDRs), compared to well-structured proteins, do not have stable tertiary structures under physiological conditions. Nevertheless, IDPs or IDRs can be found in nearly a third of proteins encoded in the human proteome [1], and they play key roles in a variety of biological processes that underlie vital cellular functions ranging from signaling, regulation to transport [2, 3]. The inherent thermodynamic instability of an IDP's conformation allows it to respond sensitively to numerous stimuli, including binding, changes in cellular environments (*e.g.*, pH), and post-translational modifications [4-8]. Such conformational plasticity arguably enables IDPs to interact with multiple signaling pathways and serve as scaffolds to form multi-protein complexes [9]. Importantly, IDPs and IDRs house around 25% of disease-associated missense mutations [10]. They have been considered promising therapeutic targets for treating various diseases (such as chronic diseases) [11-13]. While many IDPs have been shown to undergo binding-induced folding transitions upon specific binding [3], many examples are also emerging to demonstrate that IDPs can remain unstructured even in specific complexes and functional assemblies [14-20]. Such a dynamic mode of specific protein interactions seems much more prevalent than previously thought [21-23].

[1] Gong, X., Y. Zhang, and J. Chen Advanced Sampling Methods for Multiscale Simulation of Disordered Proteins and Dynamic Interactions. Biomolecules, **2021**. 11

One of the key points is to provide a reliable description of the conformational ensembles of IDPs and IDRs. A disordered state does not lend itself to traditional structural determination methods that are geared toward describing a coherent set of similar structures. Biophysical techniques such as NMR, SAXS and FRET can provide complementary information on various local and long-range structural organizations [7]. However, these ensemble-averaged measurements alone are not sufficient to unambiguously define the heterogeneous ensemble, due to the severely underdetermined nature of the structure calculation problem [8, 24, 25]. As a result, studies of IDPs have relied heavily in the traditional structure-function paradigm, by solving the folded structure of the bound state, analyzing coupled binding and folding mechanisms, or identifying putative pre-existing functional structures in the unbound state [3]. However, the disordered ensemble itself is arguably the central conduit of cellular signaling. The functional mechanism of an IDP is encoded in how the disordered ensemble as a whole responds to various stimuli, may it be cooperative binding-induced folding or redistribution of conformational sub-states in dynamic interactions. Multiple cellular signals can be naturally integrated through cooperative responses of the whole dynamic ensemble [26-28]. Therefore, there is a critical need for reliable characterization of disordered protein conformation ensembles, in both bound and unbound states, in order to establish the molecular basis of IDPs and IDRs in various physiological and pathophysiological processes.

Given the fundamental challenges of characterizing disordered protein states based on ensemble-averaged measurements alone, molecular modeling and simulations have a crucial and unique role to play in mechanistic studies of IDPs and IDRs [29-33].

This is reflected in continuously increasing numbers of research articles that contain keywords “intrinsically disordered” and “molecular dynamics” published in the last 10 years (Figure 1.1). A particularly attractive approach is to first generate the disordered ensemble using transferable, physics-based force fields without any experimental restraints and then use the later for independent validation [7]. Such *de novo* simulations of disordered protein ensembles require both high force field accuracy and adequate sampling of relevant conformational space, pushing the limit of these two central ingredients of molecular dynamics (MD) and Monte Carlo (MC) simulations. The challenges of simulating disordered proteins have driven significant interest in developing better protein force fields and advanced sampling methods. In particular, important advances have been made in the state-of-the-art atomistic force fields for describing the conformational equilibria of ordered and disordered proteins [13]. Enhanced sampling techniques have played crucial roles in both the development and application of atomistic force fields, by allowing one to cross energy barriers faster and accelerate the conformational sampling of IDPs [34-41]. Nonetheless, atomistic simulations still have limited capability for describing large systems such as biological condensates [42]. For this, the multi-scale approaches are necessary to bridge the gaps in experimental and computational time- and length-scales, including implicit solvent models, which removes the solvent degrees of freedom [8], and various coarse-grained models, which significantly reduce both proteins and solvent degrees of freedom [43].

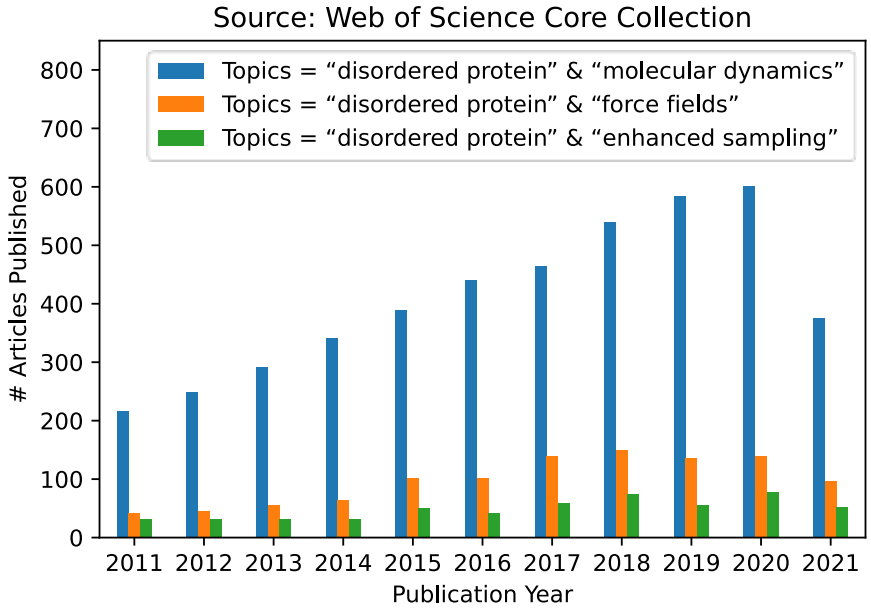

Figure 1.1 Number of articles identified with three different search keywords published from 2011 to 2021 based on a Web of Science core collection source (as of August 15, 2021).

1.2 Challenges of simulating IDP conformational equilibria

Compared to the globular proteins that have one or a few well-defined global energy minima, the energy landscape of an IDP is flatter and generally includes many local energy minima separated by modest energy barriers [44]. IDPs and IDRs typically have fewer hydrophobic residues, but a larger number of polar or charged as well as disorder-promoting residues (such as glycine and proline) [45]. These sequence features hamper the formation of hydrophobic cores that drive protein folding and thus prevent the formation of stable tertiary structures. Instead, IDPs and IDRs favor forming an ensemble of unfolded or partially folded states. This presents a major challenge for simulation and depends critically on the ability of the force fields to accurately describe the energetics of relevant conformational states, especially for capturing both folded and unfolded states of an IDP. For example, one recent study tested atomistic simulations of

IDPs for eight force fields and found marked differences in describing the conformational ensembles of IDPs, in particular the secondary structure content [46]. Similar observations have also been made in other benchmark studies, consistently showing that protein force fields previously optimized for folded proteins are not suitable for simulating disordered protein states, largely due to over-stabilization of protein-protein interactions [47]. These benchmark studies also suggested that the key towards better protein force field was to rebalance protein-protein, protein-water, and water-water interactions.

Besides accurate force fields, reliable simulation of IDPs hinges on sufficient sampling of many relevant conformation states within a reasonable simulation time. Standard MD simulations are generally insufficient to generate representative conformational ensembles, even using the most accurate protein force fields coupled with advance of GPU computing or specialized hardware such as ANTON supercomputer [48]. For example, a recent reanalysis of 30-μs ANTON trajectory of 40-residue Aβ40 peptide in explicit solvent revealed very limited convergence even at the secondary structure level [13]. This can be attributed to the diverse and large accessible conformational space of an IDP and the potentially high free energy barriers separating various sub-states that require exponentially longer time to cross. Note that typical simulation times on conventional hardware (such as GPUs) are at least one-order of magnitude shorter. There is thus great danger in relying on standard MD to calculate disordered protein conformational ensembles at the atomistic level. There is a critical need to develop and leverage so-called enhanced sampling techniques, which aim to

generate statistically meaningful conformational ensembles with dramatically less computation.

Computational studies of IDP interaction and assembly are even more demanding. The conformational equilibrium of an IDP can respond sensitively to specific and nonspecific binding, potentially shifting from a disordered to somewhat ordered state or fully folded state. In principle, simulations could provide the much-needed spatial and time resolutions to elucidate the kinetics and thermodynamics of coupled folding and binding processes and characterize the mechanistic features. However, the challenge is that this coupled process of folding and binding is a complex reaction involving the formation of many noncovalent interactions, which requires extremely long simulations generally beyond the current capabilities at the atomistic level. As such, coarse-grained models are generally required for computational studies of IDP interaction and assembly.

1.3 The state-of-the-art protein force fields for describing IDP conformations

Empirical protein force fields are potential energy functions that typically include physics-motivated bonded and non-bonded terms carefully parameterized based on a wide range of theoretical and experimental data [49]. These force fields can in principle be transferable between folded proteins and IDPs. To achieve this, it is also critical to develop suitable water models and better describe the water-protein interactions [50, 51]. Two recent review articles have already provided comprehensive descriptions on the latest development of better protein force fields [49, 52]. We therefore briefly summarize the status of the state-of-the-art nonpolarizable and polarizable force fields for IDP dynamics and interactions.

1.3.1 Nonpolarizable protein force fields

Many previous nonpolarizable force fields have significant shortcomings for describing unfolded or disordered proteins. For example, they typically provide a poor description of the secondary structure content for IDPs and have a preference to give too compact conformations with respect to the experimentally measured dimension of IDPs [46, 53]. These problems were likely attributed to the unbalanced parameterization of dihedral torsion space and description of protein-protein and protein-water interactions [54]. As a result, most of improved force fields managed to give more accurate secondary structure propensities by adjusting dihedral parameters or adding grid-based energy correction map (CMAP) parameters [52]. The over-compactness of disordered proteins can be alleviated by modifying protein-water van der Waals interactions or combining with refined water models [50]. Representative state-of-the-art force fields includes the latest CHARMM36m/TIP3P* [55], ff19SB/OPC [56]and a99SB-disp/TIP4P-D [48]. Many benchmark studies have consistently demonstrated that these refined force fields do provide significant improvements in describing not only single folded and disordered proteins, but also the multiprotein systems that are either soluble or aggregate in the solution [53, 57-60]. At the same time, these studies also identified significant remaining limitations in description of the noncovalent interactions in the multiprotein systems [58]. Recognizing limitations in the ability of a99SB-disp/TIP4P-D force field to accurately describe the protein-protein interactions, a new force field, DES-Amber, was recently developed to provide more accurate simulations of protein-protein complexes while maintaining reliable descriptions of both ordered and disordered single-chain proteins [59]. However, DES-Amber is still limited in reproducing the experimental protein-

protein association free energies of some protein complexes, in particular for the systems with highly polar interfaces [59]. In the latter case, it was found that the charged sidechains were buried at the protein-protein interface instead of being solvent-exposed. It was further suggested that nonpolarizable force fields were fundamentally limited in achieving a balanced description of charged groups that were solvent-exposed or buried at a protein-protein interface.

1.3.2 Polarizable protein force fields

Polarizable force fields explicitly consider electronic polarization using various empirical models to provide better description of charged and polar protein motifs in heterogeneous biomolecular environments [61]. Exciting progresses have been made in the last few years and several polarizable force fields are now available for stable simulation of proteins in both aqueous and membrane environments [62, 63]. Simulations using the latest polarizable force fields have also showed a high level of consistency with experimental observations, particularly the ion solvation and binding thermodynamics, permeation free energy of ions or small charged molecules into the cell membrane, and protein-ligand binding [61]. For example, the Drude-2013 polarizable force field, compared to CHARMM36 force field, is more accurate to describe folding cooperativity of $(AAQAA)_3$ peptide, which can be attributed to enhanced backbone dipole moments in the helix state [64]. Additional studies are still needed to show the necessity of considering polarizable force fields in IDP simulations, where the significantly higher computational cost adds to the challenge of generating converged ensembles [61]. Existing comparisons suggest that polarizable force fields, including AMOEBA and

Drude models, still frequently have problems in reproducing the nature structures and folding of proteins [65-67]. For example, stronger protein-water interactions in polarizable force fields can destabilize the native protein structure, opposite to the observations from nonpolarizable force fields where protein-water interactions have traditionally been underestimated [42]. Nonetheless, it can be anticipated that polarizable force fields will continue to improve and become increasingly important for simulating IDP structure and interactions.

1.4 Enhanced sampling methods for sampling IDP conformational ensembles

Enhanced sampling techniques generally accelerate the crossing of energy barriers to achieve better sampling efficiency, such as by introducing bias potentials, modifying the potential energy itself, and changing the effective temperature. These techniques have proven essential in atomistic simulations of IDPs [68, 69], yielding levels of convergence that could not be achieved even with drastically longer standard constant-temperature MD simulations [13]. The central idea of biased MD simulations is similar to importance sampling in MC simulations, where a biased potential is introduced to construct a flat free energy landscape along single or multiple collective variables of interest, such that many states can be readily sampled due to the removal of free energy barriers. The replica-exchange (REX) class of sampling methods, in particular, replica exchange molecular dynamics (REMD), has been one of the most popular methods for simulating protein conformations. Figure 1.2 shows the general scheme of REMD simulations, where the key point is to first set up multiple replicas with different unitless unbiased or biased potentials, given as the energy over $k_{\mathrm{B}}T$ (T is the temperature), and

then use the Metropolis rule to allow MC to exchange the replicas and maintain the detailed balance. A key advantage of using multiple replicas and maintaining detailed balance is avoiding the reweighting problem generally required for biased simulations. Note that virtually all biased sampling strategies can be readily incorporated within the REX framework to benefit from both classes of enhanced sampling, including metadynamics (MTD) [70, 71], accelerated MD (aMD) [72], umbrella sampling (US) [73, 74], integrated tempering sampling [75]. In practice, effective REMD protocols require proper choices of 1) the optimal number of replicas and proper distributions of conditions, to ensure a uniform exchange acceptance rate and efficient random walk in the condition space, and 2) the choice of those unitless (biased) potentials for effective conformational diffusion at each condition [76]. Here, we divide various enhanced sampling strategies into two general groups depending on the need for collective variables and discuss their recent applications to IDP conformational sampling. These methods are summarized Table 1.1.

Table 1.1 Summary of enhanced sampling methods for IDP simulations.

Types	Sampling Methods	Key Features	References
CV-based	WT-MTD	History-based adaptive bias potentials	[70, 71]
	Bias-exchange MTD	Multiple replicas with bias on different CVs	[77]
	Umbrella sampling	Pre-determined bias potentials	[78]
	Machine learning	On-the-fly discovery of optimal CVs	[79, 80]
Tempering-based	Simulated tempering	Random walk in the temperature space	[81]
	Parallel tempering	Multiple replicas to avoid the need for estimating the density of states	[36]
	Integrated tempering	Integral of Boltzmann distributions over a range of temperatures as the bias	[75]
	Solute tempering	Scaling the energies of only selected atoms or terms to achieve effective tempering	[37, 82]

Accelerated MD	GaMD	Boost potentials to accelerate barrier crossing	[83]
Combinations	MSES	Temperature/Hamiltonian replica exchange to couple CG and AT models accelerate sampling	[34]
	REUS/REST	Combined REUS and REST	[84]
	REUS/GaMD	Combined REUS and GaMD	[85]
	Integrated aMD	Integrated aMD and integrated tempering	[67, 86]
	PT-MTD	Combined the WT-MTD with PT	[77]

1.4.1 Collective variables-based sampling methods and optimization

MTD and its variants have been considered one of the most important collective variables (CV)-based sampling methods for protein simulations [87]. MTD uses a history-dependent bias potential, which is generally a sum of Gaussians, to eventually construct a flat free energy landscape along the predetermined CV(s). A well-tempered MTD (WT-MTD) was later developed to increase the convergence, by gradually reducing the size of Gaussians based on the total accumulated bias potential [70, 71]. Furthermore, the parallel tempering MTD (PT-MTD) and the combinations with other biased sampling methods have also been developed to increase the sampling efficiency and convergence of free energy calculations [88, 89]. Representative examples include the PT-MTD that combines WT-MTD with PT or bias-exchange MTD that uses a different CV in each replica, rather than exchanging the temperatures. For example, the PT-WTD and bias-exchange MTD have been employed to obtain the conformational ensembles and coupled binding and folding of disordered pKID and KID proteins, using the α-score of helical structures as CVs [77]. It has also shown that the REMD-based MTD, compared to conventional MTD or T-REMD, can enhance the conformational sampling of N-Glycans using dihedral angles as CVs to characterize the global motions [90]. The binding mechanism of two disordered peptides, NRF2 and PTMA, were simulated by the WT-MTD, and the results showed that the WT-MTD method could provide converged free energy profiles with 1.5 μs sampling time [91]. Together, these applications have shown that MTD-class sampling methods can be effectively applied to IDP simulations. Beside MTD, another important class of CV-based sampling strategy is the US method [74]. US is not strictly an enhanced sampling method like MTD. It

typically uses multiple harmonic potentials to focus sampling various states along the collective variables of interest. US is often combined with REMD in studies of IDPs, as illustrated in a recent 2D window-exchange US simulation of the coupled folding and binding mechanism of HdeA homodimer [78]. The simulation was able to capture rare unfolding transitions of the dimer at neutral pH and provided detailed description of the transition pathways.

A central limitation of CV-based sampling methods is that the efficiency strongly depends on the quality of selected CV(s). For diffusion processes such as protein conformational fluctuation, it is often not clear which CVs can best capture large-scale transitions or even if these transitions could be effectively described using one or a few CVs [92-94]. Another practical limitation is that the computational cost of MTD and US grows exponentially as a function of the number of CVs, generally limiting the maximum to 3. Parallel bias metadynamics (PBMetD) approaches have been proposed to overcome this limitation, by applying multiple low-dimension bias potentials in parallel [95, 96]. Nonetheless, the efficacy of PBMetD for sampling complex (disordered) protein conformational space is yet to be demonstrated. Another recent work presented a temperature accelerated sliced sampling method to explore the high dimensional free energy landscape by combining Temperature-accelerated MD/driven-adiabatic free energy dynamics (TAMD/d-AFED), MTD and US methods to sample many CVs simultaneously [97]. However, the approach shares the limitation of PBMetD where the underlying bias potentials remain low dimensional in nature. To address the problem of determining the best CVs for a particular problem of interest, machine learning algorithms and deep learning network have been recently proposed to analyze

information from many candidate CVs and construct the free energy landscape using low-dimensional representations [79, 80]. On-the-fly discovery of optimal CV was also demonstrated using the artificial neural networks that have a strong capacity of learning and optimization for given linear or nonlinear CVs [98]. In another recent study, an 8-dimensional optimal biased potential was constructed and applied to the free energy calculations of polypeptides using two machine learning algorithms, namely, nearest neighbor density estimator and artificial neural network [99]. It has been shown that similar deep neural networks are capable of constructing nontrivial biased potentials, for deep enhanced sampling of protein conformational space and overcoming so-called hidden barriers [100, 101]. These are exciting developments that may greatly expand the applicability of MTD, US and other CV-based sampling techniques to problems of increasing complexity, including simulations of IDPs and their dynamic interactions, especially when combined with REX.

1.4.2 Collective variables-free sampling methods and optimization

CV-free sampling avoids the need to identify a set of optimal CVs and can be highly desirable for simulating high-dimensional conformational fluctuation of IDPs. Many CV-free sampling methods have also been developed, including the tempering-based and energy-scaled biased methods. Tempering-based sampling methods rely on increasing the effective simulation temperature (*e.g.*, tempering) to accelerate barrier crossing. Examples include the temperature cool walking [102], annealed importance sampling [103], simulated tempering [81], and temperature-based REMD (T-REMD) [36]. T-REMD, in particular, has proven highly effective for protein folding and studies

of IDP conformation ensembles, where multiple replicas are simulated at different temperatures in parallel to promote barrier crossing as the system undergo random walk in the temperature space (Figure 1.2). Nevertheless, one potential limitation is that the number of replicas required for T-REMD scales as the squared root of the number of degree of freedoms (DOFs) of whole system to maintain a reasonable exchange acceptance probability. This can dramatically increase the computational cost of the explicit solvent T-REMD simulations. Several methods have been proposed to overcome this limitation of T-REMD, such as adding energy-related terms (such as accelerated-MD or Gaussian accelerated MD, named GaMD) or scaling the potential energy function (including the scaled MD that scaled all energy terms and replica exchange solute tempering (REST) methods that scaled part of energy terms) [85, 90, 104, 105].

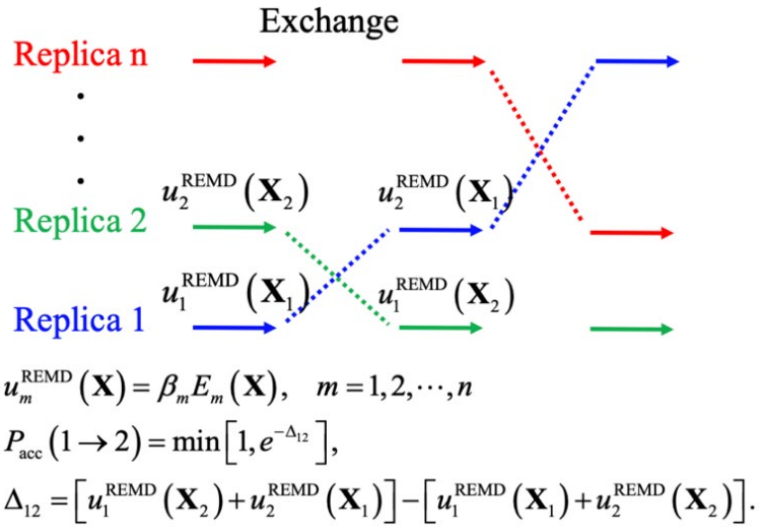

Figure 1.2 The generalized replica exchange molecular dynamics protocol based on unitless potentials, where the initial condition of each replica could have a varied temperature or scaled potential. β_m is the inverse of temperature, $E_m(\mathbf{X})$ is the potential energy of m^{th} condition for given a configuration X.

aMD adds boost potentials to reduce the energy barriers and accelerate sampling [72]. However, it suffers from a serious energetic noise when reweighting [106]. The GaMD has been thus developed to reduce noise by introducing a new harmonic boost potential, to allow a new reweighting technique that could accurately recover the free energy landscape using a cumulant expansion to the second order [83]. GaMD has achieved some success in studying protein folding, protein-ligand binding, and protein-protein interactions [106]. In particular, specifically developed Ligand GaMD [107] and Peptide GaMD [108] can capture the binding and dissociation of molecular ligands and highly disordered peptides within microsecond simulations. Recently, this GaMD method has also been combined with the REMD protocol, which can avoid the energy reweighting problem [105]. A combination of replica-exchange umbrella sampling (REUS) and GaMD has also been designed for the conformational sampling and free energy calculations [85]. It is noted that the CVs-free enhanced sampling methods are more generally more suitable for simulating IDP conformations and dynamics, because of the difficulty of identifying appropriate CVs for IDP simulations as discussed above.

REST is a special variant of T-REMD designed specifically to reduce the number of DOFs that contribute to the Metropolis criteria of replica exchange, such that smaller number of replicas are needed [37, 82]. The basic idea of REST is to separate the system into two ‘hot solute’ and ‘cold solvent’ regions. The ‘solvent’ could be actual water molecules but could also be any region of the system where no tempering is to be applied. This offers great flexibility in tailoring REST for a specific system of interest. Even more generally, the ‘solute’ region can be defined to include only a subset of interaction terms within the ‘solute’ region, such as dihedral-angle energy or Lennard-

Jones energy term in the generalized REST (gREST) method [109]. Temperature-dependent factors are used to scale the 'solute'-'solute' and 'solute'-'solvent' interactions, while keeping the 'solvent'-'solvent' interactions intact:

$$
\begin{aligned}
& u_m^{\mathrm{REST}}(\mathbf{X}) = \beta_0 \lambda_m^{\mathrm{pp}} E_{\mathrm{pp}}(\mathbf{X}) + \beta_0 \lambda_m^{\mathrm{pw}} E_{\mathrm{pw}}(\mathbf{X}) + \beta_0 E_{\mathrm{ww}}(\mathbf{X}), \\
& \text{REST1:} \quad \lambda_m^{\mathrm{pp}} = \frac{\beta_m}{\beta_0}, \quad \lambda_m^{\mathrm{pw}} = \frac{\beta_0 + \beta_m}{2\beta_0 \beta_m}, \\
& \text{REST2:} \quad \lambda_m^{\mathrm{pp}} = \frac{\beta_m}{\beta_0}, \quad \lambda_m^{\mathrm{pw}} = \sqrt{\frac{\beta_m}{\beta_0}},
\end{aligned}
\tag{1.1}
$$

where $\mathbf{X}$ is the conformational coordinates and β_{m} is the inverse of $k_{\mathrm{B}}T_{\mathrm{m}}$. The scaling of 'solute'-'solute' interactions allows the 'solute' to be simulated with an effective temperature of T_{m} while maintaining the 'solvent' temperature at T_0. As a result, the exchange acceptance probability will be independent of 'solvent'-'solvent' interactions, which reduces the effective system size and requires fewer replicas to cover the same temperature range. A key open choice in REST is how the 'solute'-'solvent' term is scaled . Different solute-solute and solute-solvent scaling factors can strongly affect the ability of driving conformational transitions of the selected 'solute' region. A strong solute-solute interaction favors the compact protein conformations, whereas a strong solute-solvent interaction prefers the disordered, solvent-exposed conformations. Different scaling schemes lead to very different characteristics of REST1 (original) and REST2 (revised) protocols (Equation 1.1). High temperature conditions favor the unfolded conformations in REST1, while both folded and unfolded conformations were observed in REST2 model for the condition with the same effective 'solute' temperature. The reason is that REST2 was designed to have a weaker solute-solvent interactions to promote the sampling of folded conformations even at high temperatures [82]. While this could allow the sampling of reversible folding transitions at all temperatures in REST2, it

could lead to conformational trapping and hampering the sampling of disordered conformations of IDPs. One important implication is that the performance of REST can be sensitive to the balance of protein-protein and protein-water interactions of a given protein force field. For example, Liu et al. showed that, while REST2 was highly effective in generating converged ensembles of 61-residue p53 N-terminal transactivation domain (TAD) using a99sb-disp, it completely failed to converge even with ~1 μs/replica in CHARMM36m and CHARMM36mw force fields [110]. Separate standard MD simulations reveal that p53-TAD can readily escape the apparent trapped conformations observed during REST2, suggesting that these traps arise due to the imbalance of scaled protein-protein, protein-water and water-water interactions [110].

REST has proven to be one of the most reliable choices for enhanced sampling of protein folding and particularly disordered conformational ensembles [111, 112]. Sugita and co-workers leveraged gREST to target the dihedral-angle energy term and successfully sampled folding transitions of beta-hairpins and Trp-cage in explicit water, using fewer replicas but cover wider conformational space compared to REST2 [109]. Walsh et al. applied REST to investigate n16N disordered peptide conformational ensembles [113]. The conformations obtained via REST methods showed a high consistency with NMR experimental data. Furthermore, REST are specifically appropriate in simulating IDRs as the disordered region can be targeted in REST without tempering the well-structured region (or water). Zhou and co-workers studied the disordered loop of Staphylococcus aureus sortase A (SrtA) to order transition upon binding to calcium [114]. Chen and Liu characterized Bcl-xL interfacial conformational dynamics in explicit solvent [115]. Both works directly showed that REST covered

broader conformational spaces for intrinsically disordered regions and led to faster convergence compared to either standard MD or T-REMD simulations. REST simulations have also been successfully integrated with experiment to study how cancer-associated mutations and drug molecules may modulate the disordered ensembles of p53-TAD and Aβ peptides in recent years [116-119].

Despite the success of REST for CV-free enhanced sampling, it does not benefit from targeted acceleration along specific CVs that are known to be rate limiting. For this, REST (or REX in general) has been combined with CV-based enhanced sampling to maximize the efficiency of sampling complex, high dimensional conformational space of proteins. Some of the examples are discussed in the sections above. Here we note a couple additional recent examples. By integrating free energy perturbation (FEP) and REST methods, Abel et al. obtained more thorough samplings of different ligand conformations around the active site and realized relative binding affinity predictions [120]. Okamoto and co-workers have applied REUS/REST two-dimensional replica-exchange method to predict two protein-ligand complex systems with the help of REST to weaken the solute-solvent interactions but improve the binding events and REUS to enhance the sampling along with the reaction coordinates [84].

Multiscale enhanced sampling (MSES) is yet another fascinating example of CV-free enhanced sampling strategy. Protein folding and other cooperative transitions such as self-assembly are known to be dominated by entropy barriers, which renders tempering ineffective for driving faster transitions. Coupled with a lack of obvious CVs, sampling complex conformational transitions of IDPs and their interactions is challenging for both CV-based and REX-based CV-free methods. For this, an effective solution is to couple

atomistic simulations with a coarse-grain (CG) model, such that one could benefit from both faster transitions of CG modeling and accuracy of atomistic force field [121]. A particularly attractive approach was first introduced by Kidera and coworkers, where restraint potentials were used to couple CG and atomistic conformational dynamics along "essential" DOFs shared by the two models [35]. The bias introduced by the coupling potential is removed using Hamiltonian REX (H-REX). Chen and coworkers further adapt the method to utilize topology-based CG models (see below), better coupling potential and advanced Hamiltonian/temperature REX (H/T REX) [34, 122, 123]. Coupling the CG and atomistic models using restraints is a key strength of these MSES protocols. It allows full control of the energetic impact of diverged structures at different resolutions, which improves exchange efficiency and provides superior scalability to large systems. MSES coupling also provides robust tolerance of CG defects by preventing the CG model from dictating the conformational dynamics. The efficacy of MSES has been illustrated using several systems. It was highly effective in simulate reversible transitions of small β-hairpins and helical IDPs [34, 122, 123] and proved instrumental in further refinement of a GBMV2 implicit solvent protein force field for both ordered and disordered peptides [124]. Very recently, MSES was also observed to be effective in sampling the cis–trans transitions of lutein by coupling the atomistic model with the Martini CG model [125]. Nonetheless, the application of MSES to larger and more complex proteins has proven more challenging than originally expected, apparently due to difficulty in effective coupling of CG and atomistic conformational fluctuations of a larger protein.

Other tempering methods including integrated tempering and simulated tempering have also been combined with different biased potentials to enhance sampling [86, 126]. For example, an integrated accelerated MD method has recently used to sample the conformations of pepX peptides, and it was shown that this method can improve the sampling efficiency and provide a good strategy for simulating IDPs [67, 86]. The combination with the metadynamics has also been presented to sample the conformational space of silica, and the acceleration was increased by over one order of magnitude [126]. One significant benefit is that only a single replica is required and could be suitable for Anton specialized hardware [48]. However, one drawback is that we have to estimate the relative free energies of all conditions (or equivalently the density of states), which requires recursive simulations and can be difficult to converge for complex systems such as large IDPs and complexes.

1.4.3 Reweighting techniques for generating unbiased ensembles

When bias potentials are used to enhance sampling, reweighting is often required to obtain the unbiased samples and construct statistically optimal unbiased free energy surfaces. Two reweighting methods are widely used for this, including the weighted histogram analysis method (WHAM) for the biased simulations with specific CVs and a more general multistate Bennett acceptance ratio (MBAR) approach [127, 128]. Stability of both WHAM and MBAR can be susceptible to large energetic fluctuations due to exponential dependence of weights on the value of the unitless potentials. Large energy fluctuations among sampled conformations can lead to large uncertainties during reweighting and thus final unbiased distributions. Another population based reweighting

method has been used for unbiasing the scaled MD simulations by making a multidimensional histogram of all sampled configurations [129]. However, the dimensionality of configurational space is usually very huge and thus be hardly completely described by some dimensionality reduction techniques (such as the principal component analysis). Recently, it was proposed that this energetic noise can be alleviated by truncating the cumulant expansion of the exponential average [83], which has been originally used in the accelerated molecular dynamics. It has shown that it can accurately recover the free energy profiles within an acceptable error (~k_BT), especially for the near-Gaussian biased unitless potentials [83]. This approximated reweighting methods have therefore been successfully used for reweighting several biased simulations [85]. It should be mentioned that those reweighting techniques can be used for reweighting any biased simulation, even for the REMD simulations. Nonetheless, all reweighting methods including MBAR relies on good overlap between the true conformational space and the region sampled by biased simulations. When the overlap is limited, the reweighted distributions will remain significantly different from the true result. Conformational space of even very short IDPs (*e.g.*, ~10 residues or longer) can be complex enough to present formidable challenges for recovering the true disordered ensemble from a biased trajectory, generated either at high temperatures or with modified Hamiltonian. Instead of analyzing self-convergence (as a function of simulation time), a more rigorous test of convergence is to analyze results obtained from simulations initiated from distinct and distal initial states (such as highly structured and fully disordered conformations [7]).

1.5 Multi-scale approaches for overcoming sampling problems of large systems

As discussed above, dramatic improvement in atomistic protein force fields coupled with enhanced samplings and GPU computing have now enabled us to generate the disordered conformational ensembles of increasingly complex IDPs in both bound and unbound states. Many important phenomena related to IDPs remain largely out of the reach of physics-based atomistic simulations, such as aggregation [130-132] and biological condensates [133-136]. Here, we review two of the key multi-scale approaches that allow one to simulate longer time-scale bioprocesses and more complex systems within current computational capability, namely, implicit solvent and coarse-grained (CG) models. Both approaches have been extensively studied and applied to globular proteins as well as IDPs.

1.5.1 Implicit solvent models for removing solvent DOFs

Implicit treatment of solvent is an effective approach to reduce the computational cost of atomistic IDP simulations. The basic idea is to directly estimate the solvation free energy to capture the mean effect of solvent on the thermodynamic properties of the solute [137]. Implicit solvent is essentially a multi-scale model, where the solvent is represented using certain physical models while keeping atomistic details of the solute. These models have emerged as attractive alternatives for simulations of IDPs and their interactions compared to the explicit solvent. In particular, many generalized Born (GB) based implicit solvent models have been developed, including the fast analytical continuum treatment of solvation (FACTS) [138], Amber GB models (such as GB-HCT[139], GB-OBC[140], and GB-Neck[141, 142]), analytical generalized Born plus

nonpolar (AGBNP) [143, 144], and GB models implemented in CHARMM program (such as GBSW [145] and GBMV[146, 147]). Several of these GB models can be optimized to provide a balance between computational efficiency and accuracy desired for IDP simulations [124, 148, 149], by systematic optimization of key physical parameters such as atomic radii to balance solvation and intramolecular interactions. Applied to various model IDPs with extensive experimental data, implicit solvent simulations have provided important insights on detailed conformational properties of the unbound state and how these properties may support function [32, 33, 150-152].

Despite many successes, implicit solvent models have not been widely tested and applied to the studies of larger IDPs. Several factors likely contribute to this. Most implicit solvent models are built upon existing protein force fields, which until recent years have significant limitations in describing disordered protein conformations. Implicit treatment of solvent also relies on various approximations for computational efficiency, such as treating water as a continuous dielectric medium in GB models, limiting the ability of implicit solvent to accurately capture the conformational dependence of solvation free energy. A particular limitation is the common use of surface area (SA)-based model for describing nonpolar solvation energy, which has known limitations in describing the length-scale dependence as well as solvent screening of dispersion interactions [149]. These limitations can result in a systematic bias towards overly compact conformational ensemble, which is more pronounced for larger IDPs.

Several recent efforts have been made to further improve implicit solvent models for IDP simulations. The GB-Neck2 model has been optimized to reproduce solvation energies for a variety of protein systems [142]. Recent benchmark studies have shown

that the GB-Neck2 model can reasonably discriminate folded and disordered peptides and could be used for quantitative protein folding simulations up to millisecond time scales [153-155]. Recently, the GBMV2 model, which includes an analytical approximation of molecular volume and is arguably one of the best GB models, has been implemented on the CUDA platform using the CHARMM/OpenMM interface [156]. The ~2 order of magnitude GPU acceleration greatly enables GBMV2 to simulate the conformation and interaction of larger IDPs. The ABSINTH implicit solvent model focuses on recapitulating the polymer properties of peptides and has been successfully used for a variety of IDP simulations, including Aβ peptides and aggregation of phenylalanine [157, 158] and sequence-conformation relationship of IDPs in general [8, 159]. Recently, an ABSINTH-C model was developed to address the problem of overly shallow Ramachandran distributions of ABSINTH, by adding residue-specific correction terms [160]. The new model not only has a capacity to maintain stable native structures of α-/β-folded proteins, but also increase the reversible folding of β-hairpin peptides.

1.5.2 Coarse-grain models for reducing the DOFs of proteins

Notwithstanding the ever-improving atomistic modeling, coarse-graining has remained an attractive and often effective strategy for extending the accessible time and length-scales of MD simulations. By grouping multiple (protein) atoms into CG beads and using simplified potential energy functions, CG modeling does not only reduce the system size, often by ~10-fold, but also allows much larger MD integration time steps up to 20 fs. Together, many CG models can be several orders of magnitude more efficient than atomistic ones. Numerous CG models have achieved varying levels of success in

studies of protein folding, binding, and assembly [43, 161]. Nonetheless, there are important distinctions between the conformational properties between globular proteins and IDPs, as well as the relative importance of electrostatic, hydrophobic, and hydrogen-bonding interactions in governing their conformational equilibria. Therefore, CG models optimized for the folded proteins are generally not suitable for the IDP simulations. It is often necessary to readjust the parameters of protein-protein and protein-solvent interactions or add new terms for a more accurate description of IDP conformations (Figure 1.3). Here, we summarize several of these refined CG models for more efficient sampling of IDP conformation and interactions as well as their successes and limitations.

Gō/Gō-like models, also known as topology-based models, are based on the funneled energy landscape theory [162] and have been highly successful in describing the folding mechanism and pathway of structured proteins [163]. Somewhat surprisingly, Gō-like models have also proven effective for determining the mechanism and kinetics of IDP interactions, particularly the coupled binding and folding process [110, 164-168]. The implication is that the binding and folding are governed by similar principles that require minimal frustration for efficiency. Note that Gō-like models generally require additional calibrations to provide a more quantitative description of the balance between intermolecular interactions and intrinsic conformational propensities [169]. A key limitation of topology-based modeling of IDPs is lack of the ability to capture the impacts of non-“native” structural features and nonspecific interactions, which could play important roles in IDP structure and function. This may be partially overcome by including new energy terms (Figure 1.3) such as explicit charge-charge interactions, inert crowder molecules and confinement potentials. A particularly interesting discovery from

these extended topology-based modeling of IDPs is the role of long-range electrostatic interactions in promoting efficient coupled binding and folding, allowing IDPs to fold at timescales beyond the μs "folding speed limit" to avoid a potential kinetic bottleneck in specific recognition [166, 167, 170]. IDP-binding proteins have evolved to contain charges near the binding interface to complement those highly conserved charges on IDPs. Long-range electrostatic interactions between these charges do not only accelerate the encountering of IDPs but also promote the efficiency of IDP folding upon nonspecific encounter.

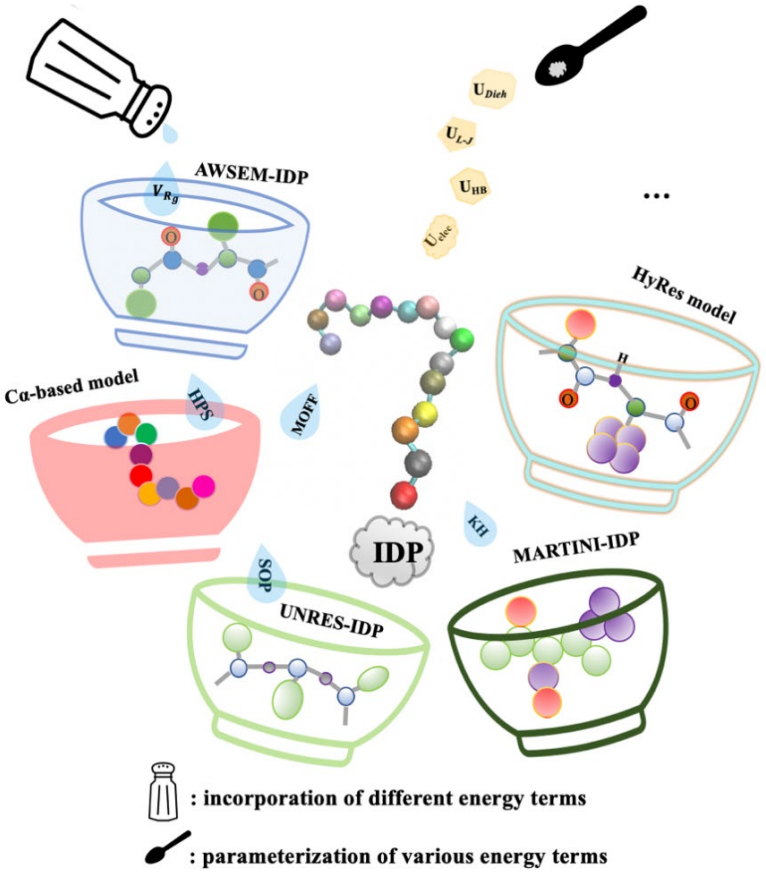

Figure 1.3 Coarse-grain modeling for addressing various IDPs-related challenges. These models can have a range of spatial resolutions and may be refined by introducing various effective potentials and/or re-calibrating the parameters of these energy terms.

Several higher resolution coarse-grained models have also been developed specifically for modeling IDPs. Thirumalai and co-workers reparametrize the two-bead self-organized polymer coarse-grained model (SOP-CG) to reproduce Rg values of a set of diverse IDPs with 20 to 441 residues [171]. The resulting SOP-IDP also accurately reproduces the small-angle X-ray scattering profiles for these IDPs. Nonetheless, SOP-IDP is designed for IDPs solely and lacks the transferability and compatibility in describing even small globular proteins under the physiological conditions. Recognizing the limitation of Cα-only backbone representation in capturing the intrinsic conformational propensities of IDPs, Chen and Liu developed a hybrid resolution (HyRes) model that contains an atomistic description of the backbone, to provide a semi-qualitative description of the secondary structure propensities, and intermediate resolution side chains, to allow qualitative description of the overall peptide chain dimension and transient long-range interactions [172]. While HyRes was originally designed for driving faster atomistic sampling for MSES simulations, applications to a set of small and large IDPs including p53-TAD suggest that HyRes may be appropriate for simulating IDP structure and interactions by itself [172]. Papoian and co-workers have developed the AWSEM-IDP model that can be used to efficiently sample the large conformational space of IDPs and at the same time can distinguish the levels of peptide chain expansion of globular proteins and IDPs [173]. AWSEM-IDP includes only C_α, C_β and O atoms, and has been reparametrized for IDPs by adjusting the secondary structure-related potential energy terms as well as introducing a new parameter, V_{Rg} term, for controlling the collapse and size fluctuation of the protein.

An important application for CG models is to study liquid-liquid phase transitions (LLPS) that are frequently mediated by IDPs [29, 45, 174, 175]. Dignon et al, proposed a residue-based C_α-only CG model to represent the disordered low complexity domain of the RNA-binding protein FUS-LCD and the DEAD-box helicase protein LAF-1 in the formation of LLPS [176]. The model uses the Debye-Hückel approximation for long-range electrostatic interactions and the hydrophobicity scale model [176] or the Kim-Hummer model [177] to short-range residue-residue interactions. The results indicated that both two approaches could reproduce the experimentally observed phase behaviors and changes in phase diagrams caused by mutation. Although they mentioned that the temperature-dependent phase behaviors were not compatible with the experimental absolute temperature and the ionic strength dependence was not fully tested due to the breakdown of the Debye-Hückel electrostatic energy potentials. The model could be further refined. For example, more residue-type parameters were considered to account for phosphorylation and acetylation effects [178], which allows in-depth investigation of how post-translational modifications may control LLPS behaviors. Recently, Latham and Zhang re-tuned Dignon et al's model to better reproduce the Rg distributions of a set of folded and disordered proteins [179]. The resulting Maximum entropy Optimized force field (MOFF) includes a new residue-residue interaction matrix and is more transferable for modeling both globular proteins and IDPs. Hummer and co-workers modified the MARTINI model via re-scaling the solute-solute non-bonded Lennard-Jones potentials to reproduce the experimental transfer free energies of phase separation among dilute and dense liquid phases and proposed a more general approach in tuning CG models with MD for LLPS related studies by optimizing and balancing the solute-solute and solute-solvent

interactions then matching the CG data to the atomistic simulation or experimental results [176]. The resulting MARTINI-IDP model was shown to successfully simulate the droplet formation and capture reversible phase transformations. These are exciting progresses that highlights the strong potential for simple Cα-only CG models in molecular simulations of LLPS involving IDPs. Nonetheless, difficulty in describing local structure propensities (such as transient helices) with the Cα-only representation may be an important limitation for studying certain specific effects of IDPs in LLPS.

1.6 Dissertation outline

Effective and reliable molecular simulations are crucial for characterizing the details of disordered conformational ensembles of IDPs in isolation, dynamic complexes or biological condensates. Although many advanced computational methods have been developed to generally simulate the IDPs, their current reliability is still inconclusive, even for the state-of-the-art accurate atomistic protein force fields. This dissertation has therefore investigated both explicit and implicit solvent atomistic force fields for simulating the folding and dynamics of IDPs. Chapter 2 includes the benchmark and optimization of CHARMM36m explicit solvent force fields for simulating both ordered and disordered proteins, which provides important insights on the capability of the state-of-the-art protein force fields and the potential limitations. Besides, this dissertation includes the development of implicit solvent force fields for accelerating conformational sampling of IDPs. For example, chapter 3 uses the graphics processing units to accelerate the GBMV2/SA implicit solvent simulations, chapter 4 further assesses its capability of simulating IDPs by using the multiscale enhanced sampling and gives its current

limitations to simulate both ordered and disordered proteins, and chapter 5 develops an improved GBMV2/NP model by introducing more physical treatment of nonpolar solvation free energies. The overall summary and future directions are also included in chapter 6.

CHAPTER 2

CHARMM36M EXPLICIT SOLVENT FORCE FIELDS FOR SIMULATING BOTH ORDERED AND DISORDERED PROTEINS: BENCHMARK AND OPTIMIZATION

2.1 Introduction

Many additive protein force fields have been successfully designed for biomolecular simulations, in particular for simulating the globular and small peptides [180]. However, recent studies have shown that they likely can give inaccurate descriptions in the solvation free energies of some amino acids in water [181] and provide compact ensembles of intrinsically disordered proteins (IDPs) [29, 182]. It has been suggested that these force fields could favor protein-protein interactions, leading to a preference for compact conformations and protein aggregates [58].

In recent years, several protein force fields and water models have been developed to address the issue of preference for compact conformations and improve the balance between protein-protein and protein-water interactions in both folded and unfolded proteins [52, 183, 184]. This has been achieved through modifications to the water model (*e.g.*, modified TIP3P, OPC, and TIP4P-D), torsion potentials (*e.g.*, backbone and sidechain dihedral angles), or nonbonded parameters (*e.g.*, scaling the Lennard-Jones parameters and charges). These balanced force fields have demonstrated improved agreement with experimental observations, particularly in terms of structural descriptions such as secondary structures and dimensions, for most disordered proteins. For instance, a simple scaling of protein-water Lennard-Jones parameters not only

improves the solvation free energy descriptions of amino acids and restores the dimensions of IDPs but also has minimal impact on the structural stability of well-structured proteins [183, 185]. Furthermore, enhancing the dispersion interactions of water models (*e.g.*, TIP4P-D) can be employed to modify both water-water and protein-water interactions, facilitating sampling of expanded disordered states and significantly improving the descriptions of folded and unfolded states [50]. Notably, representative force fields such as CHARMM36m and a99SB-disp have been specifically developed to mitigate the issue of over compactness observed in earlier force fields and achieve a more balanced protein-protein and protein-water interaction profile [48, 55].

Although many research studies have demonstrated impressive levels of accuracy in simulating many folded/unfolded proteins or peptides [180], some inconsistencies or failures were still observed in the CHARMM36m and a99SB-disp force fields. For example, many independent studies found that CHARMM36m force fields still favor the collapsed states, which were observed in simulating the nuclear coactivator binding domain [186], the phosphorylated disordered peptides [187, 188], and aggregates of ubiquitin proteins [58]. On another hand, we found that the CHARMM36m force fields, compared to the a99SB-disp force field, did not generate the converged ensembles of the p53 Transactivation Domain (p53-TAD), which could be attributed to a higher free energy barrier separating helical and unfolded coil states [110]. In addition, the a99SB-disp force field is still limited to accurately describe protein–protein complexes, although an improved force field (DES-Amber) was developed to reduce this effect [59]. These studies therefore suggest that these force fields are still inconsistent in simulating some disordered proteins or protein-protein complexes to a certain extent.

In this work, we propose to utilize small protein systems, specifically backbone/side chain pairs, as a means to investigate potential differences in describing protein-protein and protein-water interactions among various force fields. This approach allows us to avoid convergence issues and clearly identify potential problems. Notably, several studies have employed the potential of mean forces (PMFs) between amino acid side chains or backbone to benchmark protein force fields [185, 189, 190]. In our research, we focus on assessing the impact of state-of-the-art force fields, including CHARMM36m and a99SB-disp, on the free energy profiles of representative amino acid side chain pairs. Our hypothesis suggests that these force fields are likely to yield different stabilities for these pairs, with CHARMM36m favoring protein-protein interactions and potentially resulting in more compact conformations compared to a99SB-disp. By investigating these test systems, our study not only provides valuable insights for optimizing and benchmarking protein force fields but also sheds light on potential differences between CHARMM36m and a99SB-disp force fields in simulating intrinsically disordered proteins (IDPs).

2.2 Methods

2.2.1 Test systems and force fields

We first selected representative conformations of test monomers, including alanine dipeptides (original and modified), and nonpolar (aromatic and nonaromatic), polar, and charged (positive and negative) amino acids side chains (Figure 2.8). These conformations can form many representative dimers, including nonpolar-nonpolar,

nonpolar-polar/charge, and charge-charge pair (Figure 2.9 & Figure 2.10). It should be mentioned that they can potentially form hydrophobic, pi-pi stacking, backbone/sidechain hydrogen-bonds, and salt-bridge electrostatic interactions, which commonly exist in many protein-protein interactions. Besides these side chain interactions, we selected several important side chains of charged amino acids (Glu, Arg, and Lys) to investigate their interactions with different water models (TIP3P* and a99SB-disp), which enables us to identify the potential difference in describing protein-water interactions.

To quantify whether these force fields have different performance in simulating these dimer systems, we also calculated their free energy profiles and stabilities. To avoid potential convergence problems, we fixed the conformations of these side chains by selecting multiple representative conformations. Taking the Trp-Trp side chain dimer as an example, we selected the edge-edge, edge-face, face-face, parallel or antiparallel displaced conformations (Figure 2.9). We then individually calculated the PMFs along a distance between two atoms that could form representative interactions. We selected two types of representative protein force fields. The first one is the CHARMM36m force field with two different water models (the modified TIP3Pm and the recommended TIP3P*), named as c36m and c36mw, respectively. Another one is the a99SB-disp force field with the recommended water model, named as a99SB-disp. It is noticed that they have been optimized by their main developing group for both ordered and disordered proteins, and also, they have been widely used in protein simulations. Additionally, our previous simulations showed that they had different descriptions in secondary structures of a p53-TAD system [110]. This selection therefore provides useful information to answer why the c36m/c36mw still gave more compact conformations, compared to the a99SB-disp.

2.2.2 An on-the-fly approach to calculate the free energy difference between two overlap states

We assume that two states (*a* and *b*) are sufficiently overlap, then a first-order approximation can be applied in both f_{ab} and f_{bc} along an order parameter (ξ),

$$\begin{aligned} f_{ac}(x) &= f_a(x) - f_c(x) \approx \frac{\delta f}{\delta \xi}(x)\Big|_{\xi \to \xi_c} \xi_{ac} = g_c(x)\xi_{ac}, \\ f_{bc}(x) &= f_b(x) - f_c(x) \approx \frac{\delta f}{\delta \xi}(x)\Big|_{\xi \to \xi_c} \xi_{bc} = g_c(x)\xi_{bc}, \end{aligned} \tag{2.1}$$

when we have $|\xi_{ac}| = |\xi_{bc}|$, then we have,

$$\begin{aligned} &\left\langle \delta^2 f_{ac}(x) \right\rangle_c = (\xi_{ac})^2 \left\langle \delta^2 g_c(x) \right\rangle_c \\ &= (\xi_{bc})^2 \left\langle \delta^2 g_c(x) \right\rangle_c = \left\langle \delta^2 f_{bc}(x) \right\rangle_c . \end{aligned} \tag{2.2}$$

To calculate the reduced free energy, F_{ab}, we expand these two states at an intermediate state c,

$$\begin{aligned} F_{ac} &= -\log\left\langle e^{-f_{ac}(x)} \right\rangle_c \approx \left\langle f_{ac}(x) \right\rangle_c - \frac{1}{2}\left\langle \delta^2 f_{ac}(x) \right\rangle_c , \text{and} \\ F_{bc} &= -\log\left\langle e^{-f_{bc}(x)} \right\rangle_c \approx \left\langle f_{bc}(x) \right\rangle_c - \frac{1}{2}\left\langle \delta^2 f_{bc}(x) \right\rangle_c . \end{aligned} \tag{2.3}$$

Then, we have the following estimator of the F_{ab}, by using a second order cumulant expansion,

$$F_{ab} = F_{ac} - F_{bc} \approx \left\langle f_{ab}(x) \right\rangle_c \simeq \frac{1}{N}\sum_{n=1}^{N} f_{ab}(x_n), x_n \in \text{states}\{c\}. \tag{2.4}$$

It shows that the free energy difference between two sufficiently overlapping states can be calculated by using the trajectories sampled from an intermediate state, which provides an on-the-fly way to calculate the free energy difference.

2.2.3 Free energy profiles, stabilities, and decomposition

We then used the free energy perturbation (FEP) method to obtain all free energy profiles, where all protein conformations were fixed by using a massless strategy and the water molecules can be free to move in the simulation box. A simple approximation (Equation 2.4) was used to calculate the free energy difference between any two states (F_{ab}), where they can be written as the sum of individual reduced potential energies between two configurations (f_{ab}),

$$F_{ab} \approx \frac{1}{N}\sum_{n=1}^{N} f_{ab}\left(\mathbf{x}_n\right), \tag{2.5}$$

where the $f_{ab} = \beta\left(E_a - E_b\right)$, E_a and E_b are the potential energy of state a and b, respectively, β is the inverse of k_BT, $\mathbf{x}_n$ is one configuration from an immediate state $\xi_c = 1/2\left(\xi_a + \xi_b\right)$, and N is the number of samples collected.

We selected multiple windows that covered the distances ranging from 13.5 to 1.2 Å with a step of 0.01 Å, and then we ran 10 ns molecular dynamic (MD) simulations for each window, which can provide converged simulations (Figure 2.11). The PMF profile was then simply obtained by adding multiple states, while the standard errors of mean were calculated by dividing the 10 ns simulations into 10 blocks, where they are independent by looking at their time correlations. We defined the stability of one test system as the free energy difference between the first valley and the average PMF values along the distances more than 11.0 Å,

$$\text{Stability} = \text{Mean}\left[G\left(\xi \geq 11.0\text{Å}\right)\right] - G\left(\xi = \text{first min}\right), \tag{2.6}$$

where ξ is the distance between two representative atoms, and $G(\xi)$ is the free energy of the state ξ. Based on the Equation 2.1, we have the following free energy decomposition for a PMF profile,

$$f_{ab}(\mathbf{x}) = f_{ab}^{\mathrm{pp}}(\mathbf{x}) + f_{ab}^{\mathrm{pw}}(\mathbf{x}), \tag{2.7}$$

where the $f_{ab}^{\mathrm{ww}}(\mathbf{x})$ is zero, due to the use of the same configuration in both state *a* and state *b*. Obviously, we can have a further decomposition into the vdW and electrostatic contributions. Similarly, this can also be applied to the energy decomposition of monomer-water interactions.

2.2.4 Computational details

The atomic coordinates of each dimer were first obtained from a geometry optimization, and then fixed in all FEP simulations. We used a cubic simulation box, and the size was set to be large enough, so that they do not have interactions with their images. The topology and parameters of each side chain were determined by the corresponding residue in a given protein force field. They were then converted into the GROMACS topology file. The GROMACS program was used to generate the initial configurations and add the water models for all force fields [191]. The OpenMM program was used to equilibrate the simulation systems and run the production of FEP calculations [192]. The cutoff scheme (d_c = 12.0 Å) in the OpenMM program was used to calculate the nonbonded interactions, which considered a default reaction field approximation to cover the effect of atoms beyond the cutoff distance. The Langevin thermostat with a collision frequency of 1.0 ps^{-1} was used for the temperature control to give us the NVT ensembles, and the temperature was set to 300 K. The time step of 2 fs was used for all

productions. Other parameters were default values provided by the v7.4 OpenMM program. The VMD [193] was used to visualize the configurations and all analyses were written by the in-house Python scripts.

2.3 Results and Discussion

2.3.1 Free energy profiles of representative dimers

We calculated the PMF profiles of all amino acid side chain pairs, including the nonpolar-nonpolar, nonpolar-polar, polar-polar, polar-charged, charged-charged pairs. Six of them were then selected as representatives, to show the nonpolar-nonpolar, nonpolar-polar, π- π stacking, hydrogen bonding, and charge-charge interactions (Figure 2.1). It can be seen that these PMF curves have similar positions for their valleys and peaks, which suggests that they can give similar solvation shells, where the displacements were less than 0.5 Å mostly, especially for the minimum positions. However, the a99SB-disp force field showed different amplitude of free energy differences between those valleys or peaks, compared to the PMF profiles of both c36m and c36mw force fields. The representative examples include the wa_pd, ww_p, and wbco pairs. It is fair to state that three force fields have similar PMF profiles for most nonpolar pairs (*e.g.*, aa_h), which could be attributed to their similar parametrization strategies in describing vdW interactions. However, it was noted that most polar pairs were even worse than those nonpolar pairs, which can be also observed in the following stability analysis.

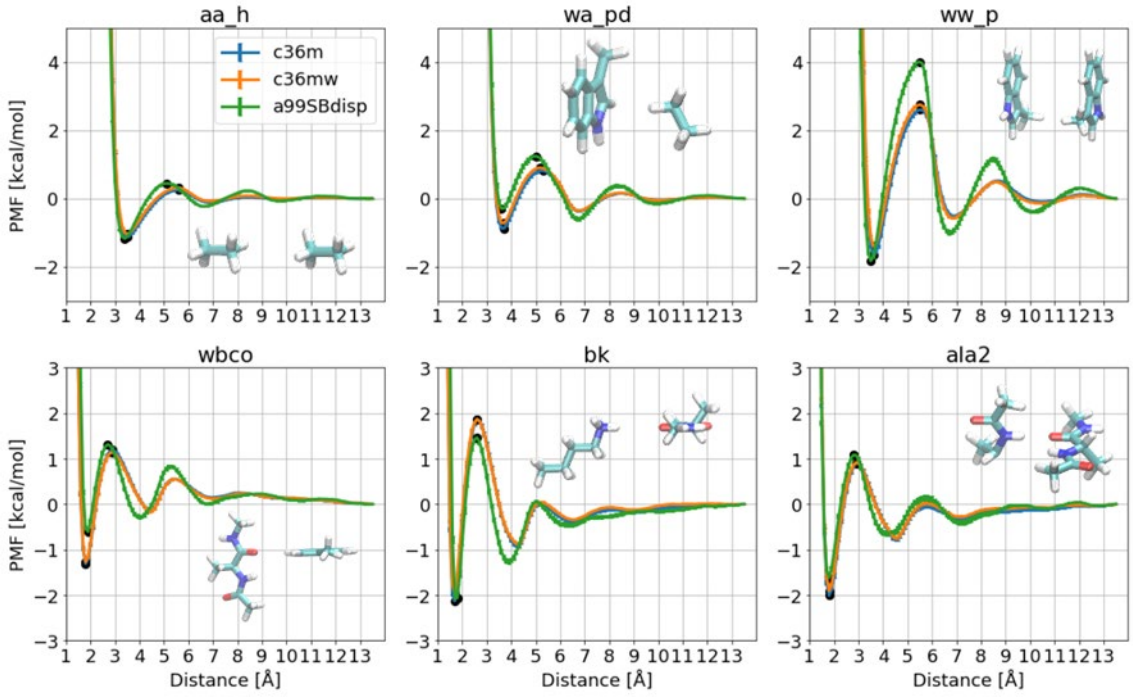

Figure 2.1 PMF profiles of representative dimers for three force fields (c36m, c36mw, and a99SB-disp) and their configurations were inserted. The selected six dimers represent basic nonpolar-nonpolar/polar, pi-pi stacking, hydrogen bonding, and charge-charge interactions.

2.3.2 Stabilities of nonpolar, polar, and charged pairs

It showed that both c36m and c36mw force fields gave quite similar PMF profiles, but they were different from the a99SB-disp force field. To provide detailed comparisons, we calculated their absolute stabilities and relative stabilities for all amino acid side chain pairs (Figure 2.2). The stability can be utilized to determine the strength of protein-protein interactions in water: the larger stability of one pair likely results in a more favorable protein-protein interaction, which could give more compact protein conformations. It can be observed that all pairs have a positive stability, except the “ek” pair. Two different configures of Glu-Lys side chain pair (“ek” and “eks”) were used in our calculations, where the ek pair did not have a formation of hydrogen-bond, so it has a

negative stability, while the configure of the “eks” pair favors the formation of a strong hydrogen bond. It also shows that the force fields have an important impact on the stabilities of both nonpolar and polar pairs. For example, a significant difference is observed in the polar pairs, especially for the charged pairs, such as “rrsa”, “eks”, and “hpe” pairs.

Figure 2.2 also displayed the relative difference between two protein force fields. We first compared the c36m and c36mw force fields because they used the same parameters of proteins but had different protein-water interactions. It clearly showed that the c36mw force field provided less stabilities for all pairs except the “he” pair, meaning that the c36mw force field could give less compact protein conformations. This is consistent with the previous observations, where the c36mw force field provides less compact protein conformations [55]. However, compared to the a99SB-disp force field, the c36mw still provides higher stabilities of many polar/charged pairs, although less significant for most nonpolar pairs. The distinct pairs are highlighted in Figure 2.2, where the relative stabilities are more than 1 kcal/mol. For example, the “rrsa” pair is much more stable in the c36mw (5.2 kcal/mol) than the a99SB-disp force field (2.4 kcal/mol). Surprisingly, most pairs with the Glu side chain gave a different stability in both force fields, particularly for “eks” pair. The corresponding configurations showed that they formed a strong hydrogen bond (Figure 2.10), so it suggested that both c36mw and a99SB-disp force fields had different descriptions of this hydrogen bond interaction.

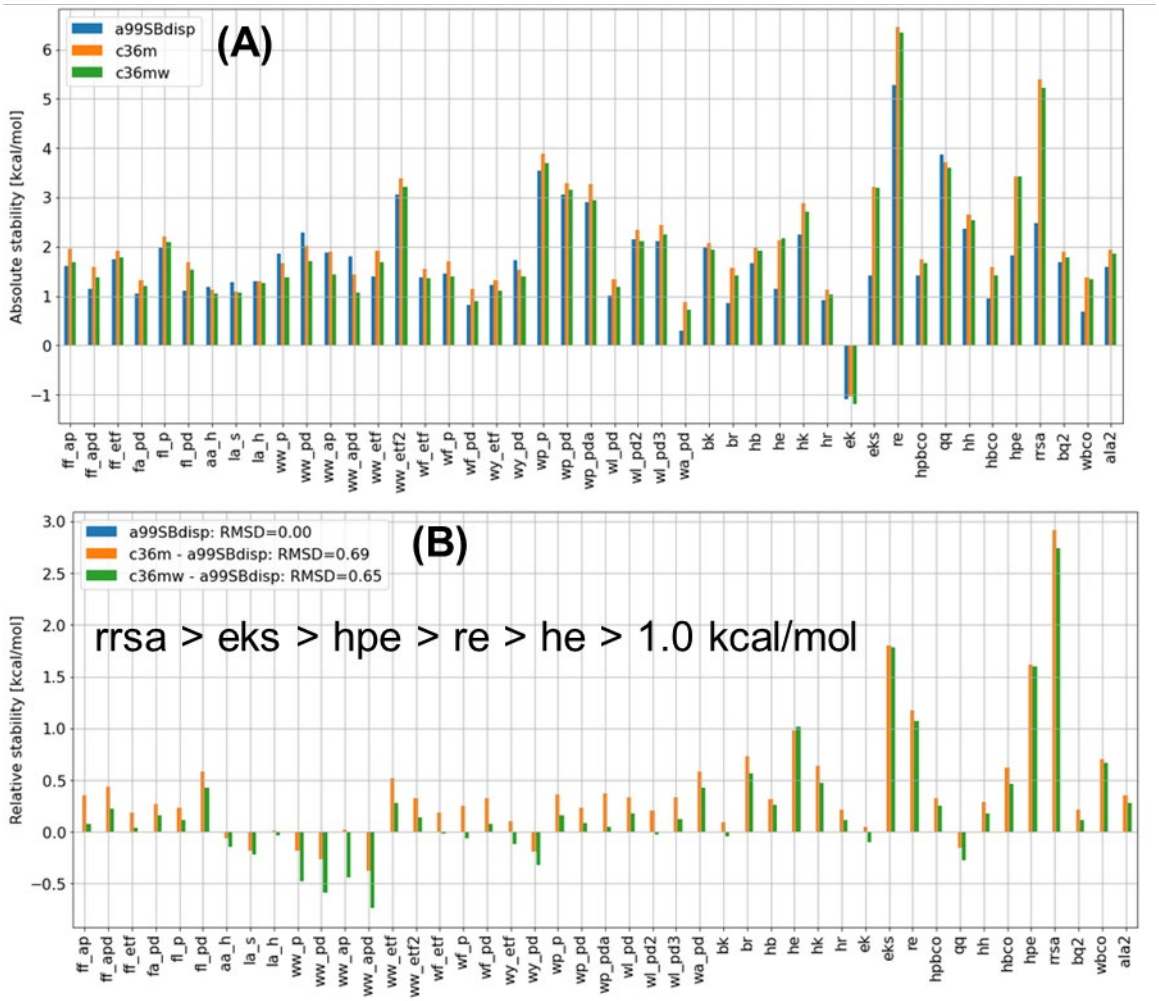

Figure 2.2 (A) The stabilities of all amino acid side chain pairs for three protein force fields. (B) The relative stabilities between force fields, and the pairs where their relative stabilities are more than 1 kcal/mol are labeled.

It can be concluded that the charged residual side chains (*e.g.*, Args and Glus) performed differently in both c36m/c36mw and a99SB-disp protein force fields, where the c36m/c36mw force fields gave larger stabilities of these charged pairs, especially for the pairs with Arg and Glu side chains. Our previous p53-TAD protein simulation suggested that the c36m gave more compact conformations [110]. We therefore hypothesized that the observed compactness could be due to the imbalanced descriptions of nonbonded parameters of charged residues. It is noted that we are not providing a

comprehensive optimization of the c36m/c36mw force fields, instead our target is to understand why these force fields performed differently in simulating some disordered proteins. To reveal the underlying reason on why those charged pairs performed so differently in both c36m/c36mw and a99SB-disp force fields, we therefore took the calculations from the a99SB-disp force field as the reference, and then did the free energy decomposition of particular pairs to reveal what contributions dominated the difference.

2.3.3 Glu side chain-involved pairs: imbalanced protein-protein and -water electrostatic interactions

2.3.3.1 Free energy decomposition

Similarly, we decomposed the free energy difference of the Glus-involved pairs into the protein-protein and protein-water contributions. Taking the "eks" as an example (Figure 2.3), it showed that the minimum positions were slightly different and around 1.8 Å, which can be also found in the pairs of "hpe" and "re". Nevertheless, it had significant differences in both protein-protein and protein-water contributions, although these different contributions compensated each other to achieve a less significant difference in total free energy. These results showed that tuning either protein-protein or protein-water interactions could help to achieve a balance. Similarly, we further decomposed the free energies by selecting two Glus-Lyss and Glus-H_2O systems, into the vdW and electrostatic interactions (Figure 2.4). It clearly showed that both a99SB-disp and c36m force fields gave distinct descriptions of the electrostatic interactions. It was noted that those two force fields gave similar descriptions of the Lyss-H_2O system for both vdW

and electrostatic interactions. This suggested that the free energy difference was attributed to the imbalanced electrostatic contributions.

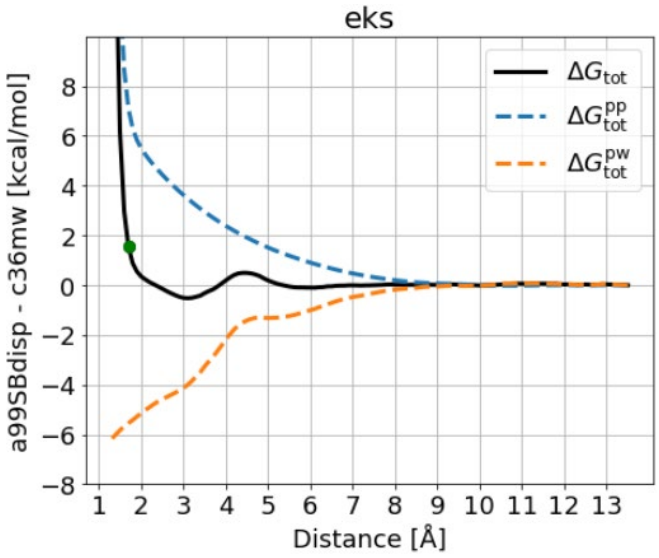

Figure 2.3 The free energy difference of "eks" pair between the a99SB-disp and c36m and its decomposition into the protein-protein and protein-water components. The minimum distance of the PMF profile for the a99SB-disp and c36m force fields was plotted as a dot, respectively, and the corresponding difference represented their stabilities.

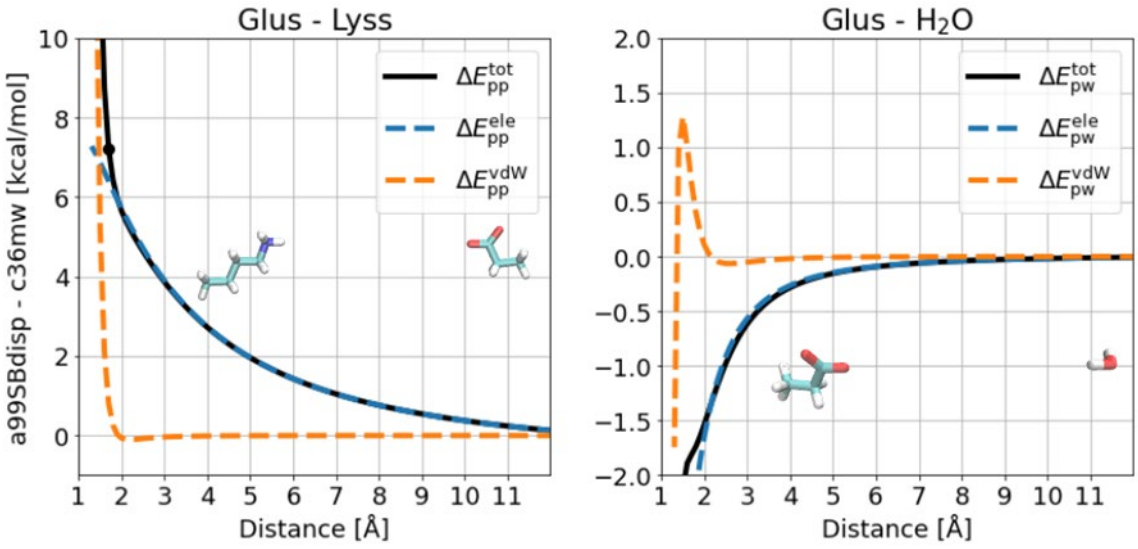

Figure 2.4 The energy difference between a99SB-disp and c36m force fields and its energy decomposition into the vdW and electrostatic contributions for the Glus-Lyss and Glus-H_2O systems.

2.3.3.2 Using a99SB-disp water model can decrease the free energy difference

We first increased the polarity of the Glu side chain by tuning the charge distribution and found that this strategy was insensitive to reduce the free energy difference. The reason was that increasing the polarity can increase the interactions with both protein and water, so that the total interaction changed slowly. Another strategy was to change the protein-water interaction. We then used the a99SB-disp water model in the c36m protein force field to calculate the stabilities of Glus-involved pairs. It was observed that it, compared to the c36m or c36mw force fields, can reduce their free energy differences, which were within 0.5 kcal/mol (Figure 2.5). It should be mentioned that the a99SB-disp water model has different nonbonded parameter values in both vdW and charges from the TIP3P* water model, so it suggests that we likely need to rebalance the electrostatic interactions between these charged side chains (*e.g.*, Glu and Asp) and water models.

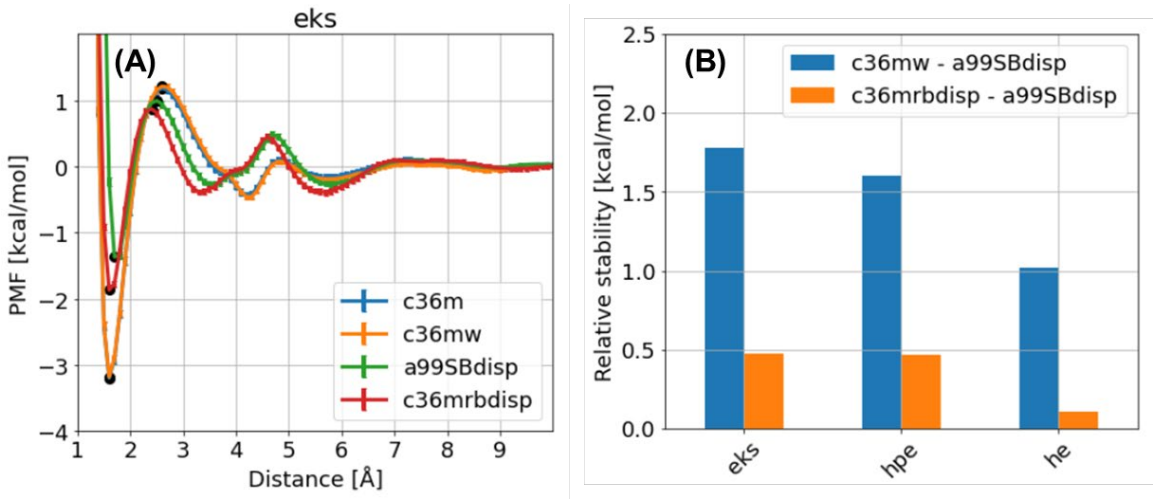

Figure 2.5 (A) The PMF profiles of "eks" pair for four force fields. (B) The free energy difference between two force fields. The c36mrbdisp force field used the c36m force field, but the a99SB-disp water model.

2.3.4 Arg-Arg side chain pair: protein-protein electrostatic interaction dominates the difference

2.3.4.1 Free energy decomposition

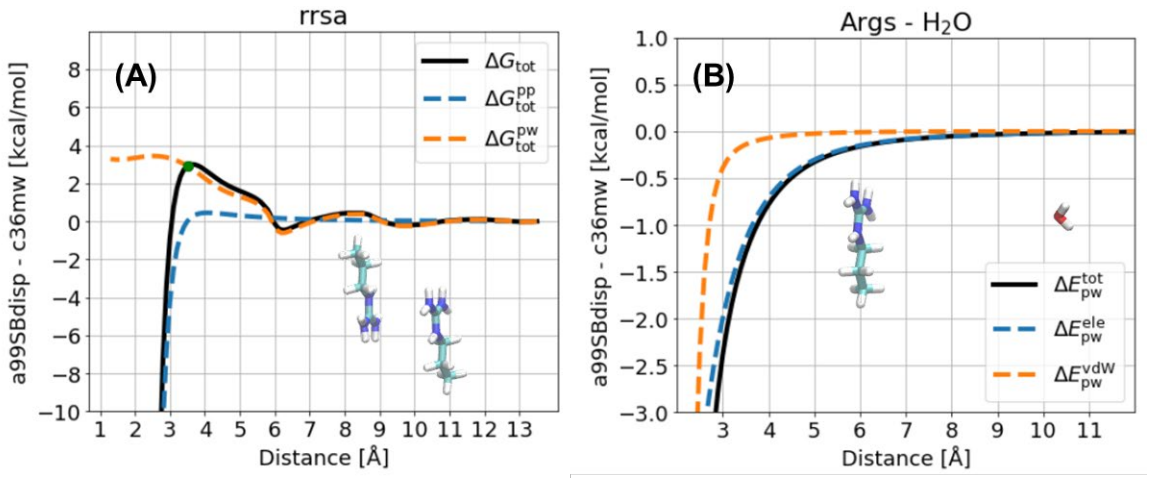

Figure 2.6 (A) The free energy difference of "rrsa" pair between the a99SB-disp and c36m and its decomposition into the protein-protein and protein-water components. The minimum distance of the PMF profile for the "rrsa" pair was plotted as a green dot, and the corresponding difference was its stability. (B) The energy difference between a99SB-disp and c36mw force fields and their decomposition into the electrostatic and vdW protein-water contributions of Args-H_2O system.

Figure 2.6 showed the free energy difference between a99SB-disp and c36m force fields for Arg-Arg side chain pair that is the "rrsa" pair. It was clearly observed that the difference in the stability was mainly contributed by the protein-water free energy component, rather than the protein-protein interactions. We therefore selected an Args-H_2O model as a quick test system that can be formed when the "rrsa" pair has a first solvation shell. To see which nonbonded component was more dominated in this protein-water interaction, we decomposed them into the vdW and electrostatic interactions (Figure 2.6). The energy decomposition demonstrated that the protein-water electrostatic

interaction made a dominant contribution to this energy difference, rather than the vdW interaction.

2.3.4.2 Tuning the charges of Args can decrease the free energy difference

To reduce the difference in the protein-water electrostatic contributions, we first tentatively tuned the charges of the Arg side chain. Given our previous observations that the c36m force field gave more compact conformation, we decided to tune the c36m force field. To keep the convention of the c36m force field, we kept the aliphatic part the same as the c36m force field, but the remaining part the same as the a99SB-disp force field. We finally tuned the charge of the "CZ" atom type, to keep the whole charge of the Arg side chain unchanged. Figure 2.7 showed the recalculated free energy difference of some pairs by using the modified c36mrbdisp force field. Obviously, its difference with the a99SB-disp force field was significantly reduced. For example, the difference was reduced from 2.9 to 0.5 kcal/mol, and the absolute difference of all Args-involved pairs were within 0.6 kcal/mol. These results showed that tuning the charges of Args was likely a simple but effective strategy.

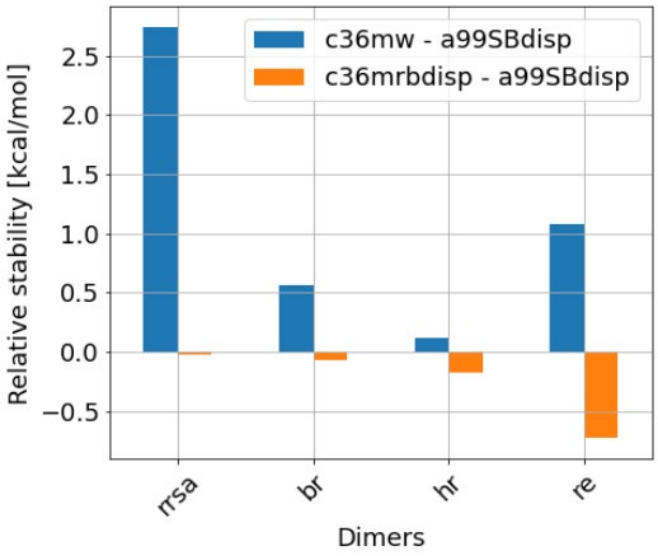

Figure 2.7 The free energy difference between two force fields. The c36mrbdisp includes the a99SB-disp water model and a modified c36m protein force field that changed the charges of the Arg side chain.

2.4 Conclusions

We tested the effect of three force fields (c36m, c36mw, and a99SB-disp) on free energy profiles of a set of representative amino acid side chain pairs. Both c36m and c36mw force fields gave quite similar PMF profiles, but the c36mw had less stability for all pairs except “he” pair. However, compared to the a99SB-disp force field, the c36 or c36mw had higher stabilities for most pairs, in particular for the polar and charged pairs (such as the pairs with the Arg and Glu side chains). These observations suggested that the c36 or c36mw could give more compact conformations of disordered proteins, which was consistent with previous p53-TAD protein simulations. The further free energy decomposition showed that the free energy difference between the a99SB-disp and c36m force fields was likely attributed to the imbalanced electrostatic interactions of protein-protein and protein-water, rather than the vdW interactions. Tuning the charges of the Arg side chain can reduce the free energy difference of the Args-involved pairs, but this

did not work for the Glus-involved pairs. However, a combination of c36m protein force fields and a99SB-disp water model can reduce the free energy difference of the Glus-involved pairs. These findings showed that balanced electrostatic interactions need to be considered carefully for further optimization of force fields.

2.5 Supporting material

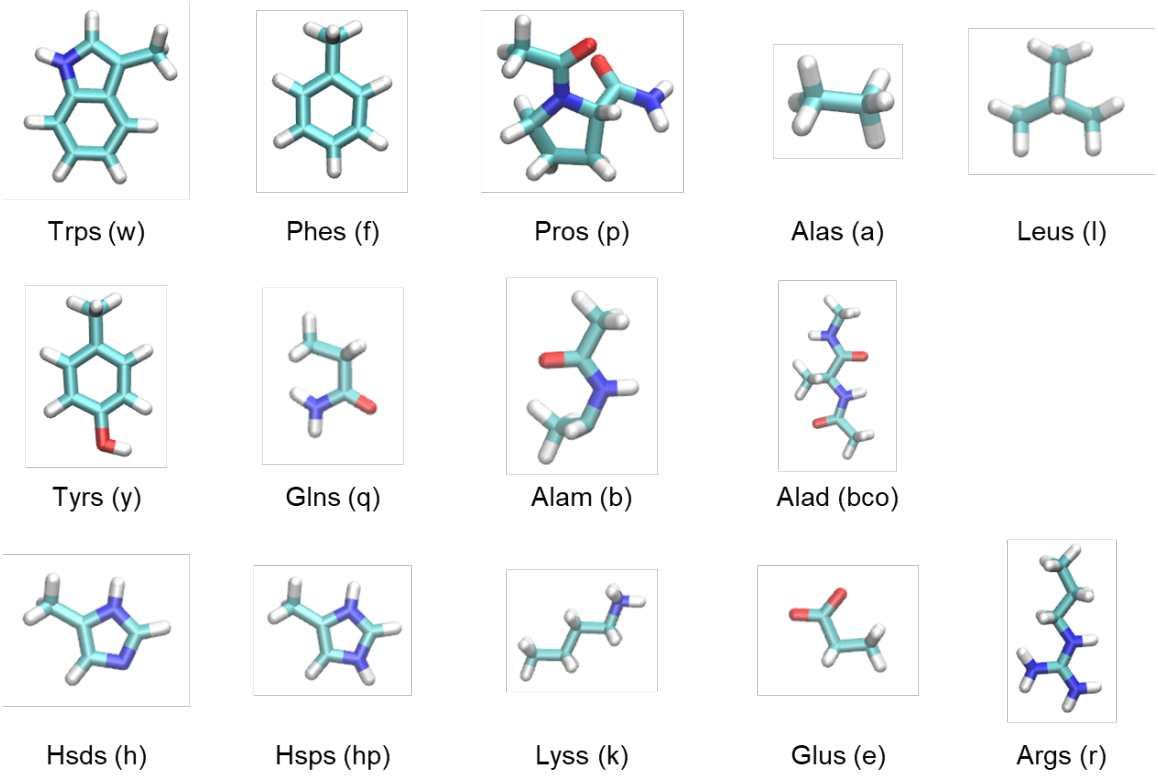

Figure 2.8 All backbone and side chain models. Their initial structures were optimized by the c36m protein force field.

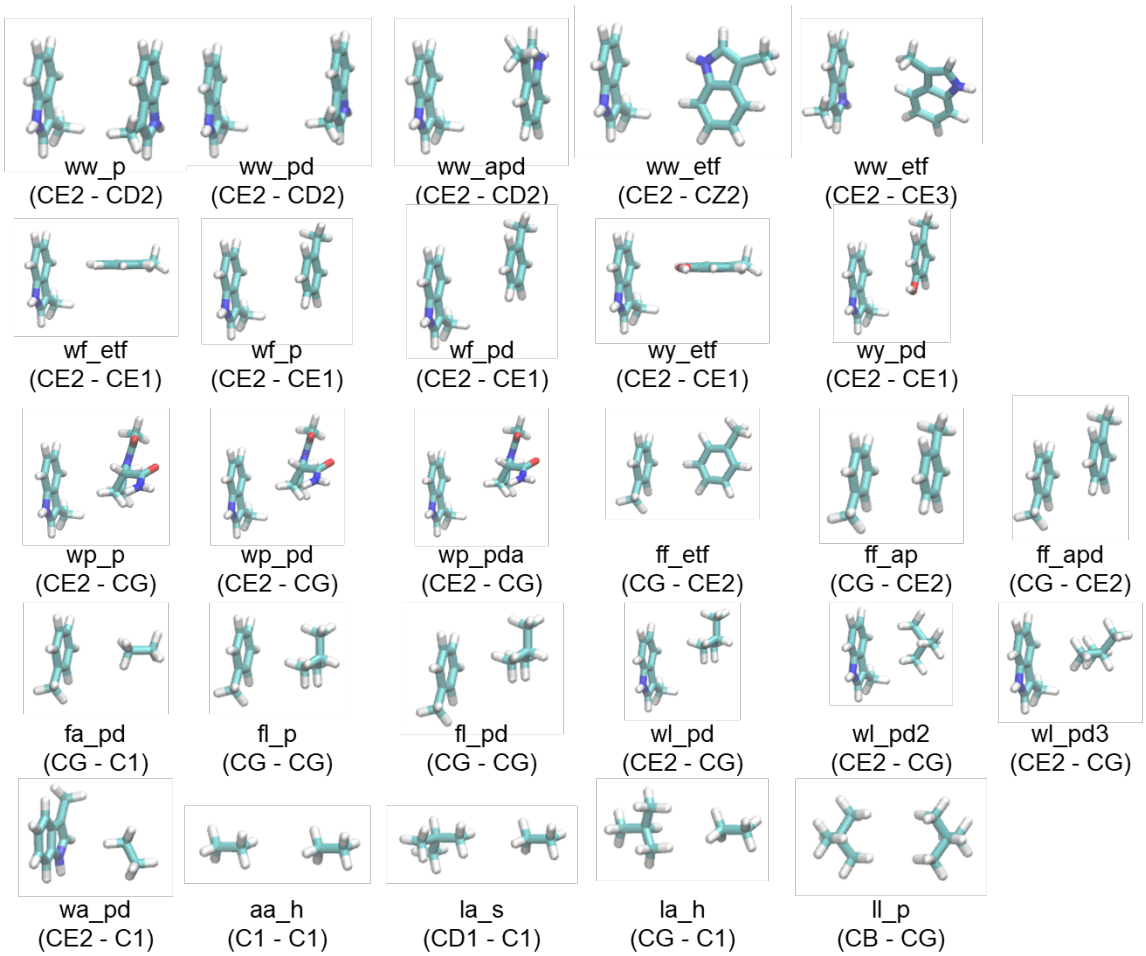

Figure 2.9 The configurations of selected nonpolar pairs.

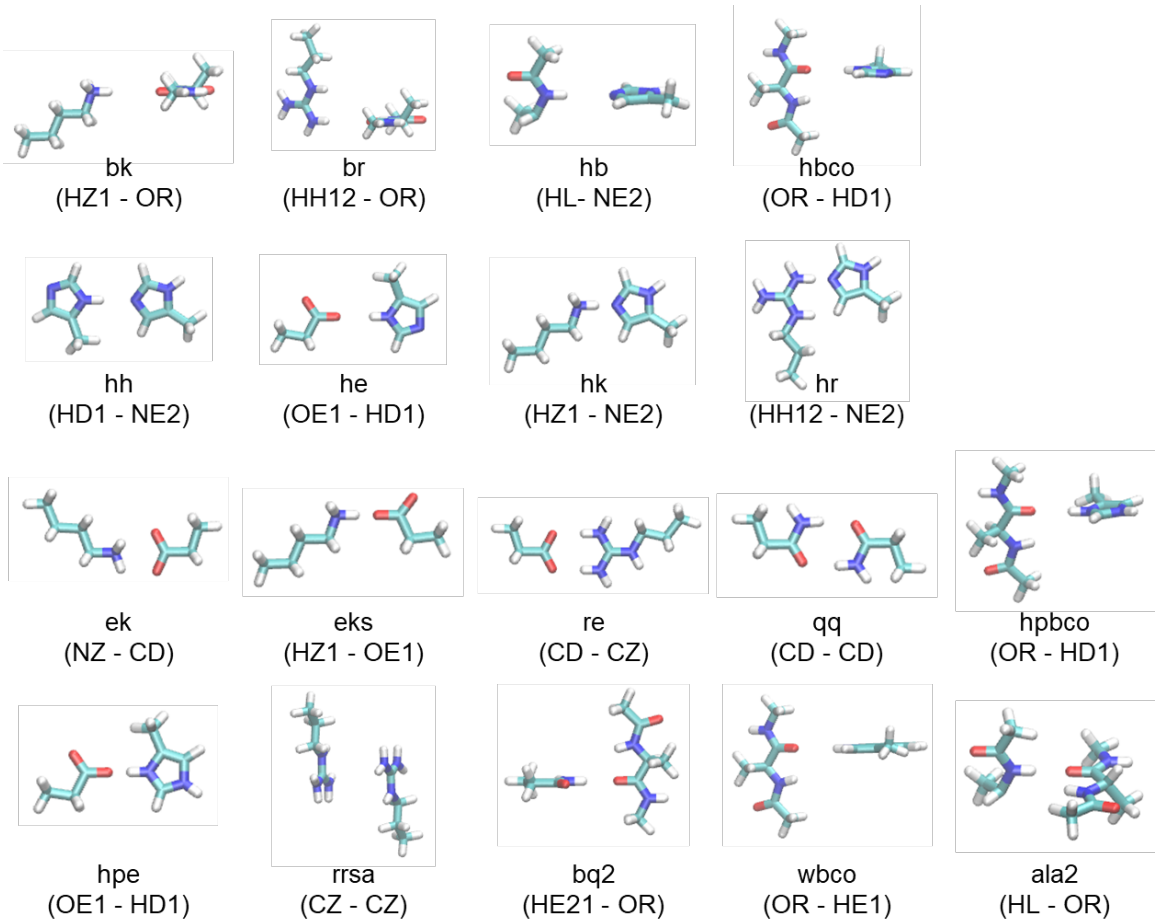

Figure 2.10 The configurations of selected polar pairs.

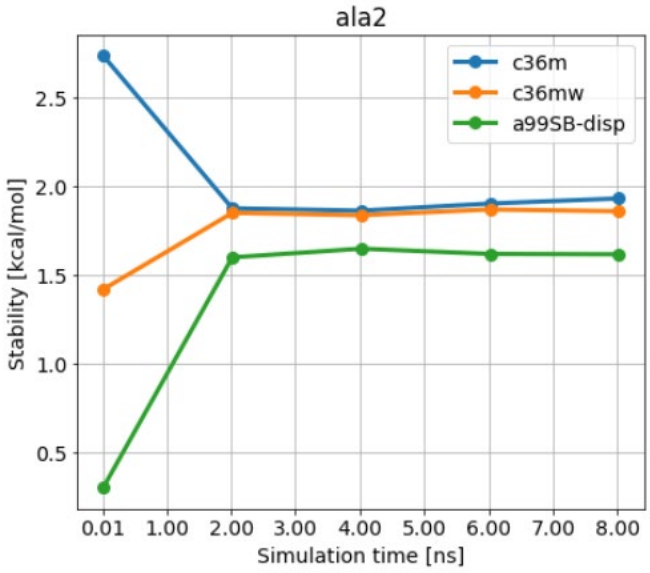

Figure 2.11 The convergence analysis by comparing stabilities with the increase of simulation time for modified di-alanine dipeptide.

CHAPTER 3

ACCELERATING THE GENERALIZED BORN WITH MOLECULAR VOLUME AND SOLVENT ACCESSIBLE SURFACE ARAE IMPLICIT SOLVENT MODEL USING GRAPHICS PROCESSING UNITS [2]

3.1 Introduction

It is crucial to provide an accurate description of the solvent environment during biomolecular simulations, where the solvent plays a vital role in governing the conformational fluctuations and transitions [194-196]. Conventionally, explicit solvent models provide a relatively detailed and accurate description on interactions between the solvent molecules and solutes, and are regarded as standard approaches to explore the influence of solvent on the solute molecule [197]. However, it dramatically increases the computational cost of a simulation, and the solvent friction further adds to the difficulty of sampling the solute conformations. Implicit solvent is a viable alternative that captures the effective influence of solvent on the solute by direct estimation of the solvation free energy as a function of the solute coordinates [198]. Implicit treatment of solvent substantially reduces the system size, thus allowing significant reduction of computational cost and faster sampling of solute conformations [137, 149, 199-201].

There are many approaches for estimating the solvation free energy in implicit solvent treatment, including the Poisson-Boltzmann (PB) and generalized Born (GB)

[2] Gong, X., et al., Accelerating the Generalized Born with Molecular Volume and Solvent Accessible Surface Area Implicit Solvent Model Using Graphics Processing Units. Journal of Computational Chemistry, **2020**. 41(8): p. 830-838.

models. Both PB and GB are based on continuum electrostatics treatment of solvent environment [202-206]. Compared with the PB model, the GB approximation allows the analytical evaluation of molecular forces and is more suitable for molecular dynamics (MD) simulations. The most important task in GB models is to evaluate the effective Born radius of each atom, which is dependent on all solute coordinates. GB models can be numerically equivalent to the underlying PB calculations, given accurate effective Born radii [198, 205]. Numerous approaches have been developed for efficient calculations of effective Born radii, including the Fast Analytical Continuum Treatment of Solvation (FACTS) [138], the Generalized Born Surface Area from Onufriev, Bashford, and Case (GBSA/OBC) [140], Analytical Generalized Born plus NonPolar 2 (AGBNP2) [144], and numerical integration-based ones such as the Generalized Born with Simple Smoothing function (GBSW) [145, 148, 207] and Generalized Born with molecular volume [124, 141, 142, 146, 147, 208-210] models. The GBMV2 model, in particular, contains an analytical approximation of the Lee-Richards molecular volume and reduces unphysical solvent-inaccessible high dielectric protein interior regions [124, 146, 147, 211]. It can reproduce the first solvent peak in the potentials of mean force (PMFs) of interactions between polar chemical groups [207]. A comparison of several implicit solvent models has also suggested that the GBMV2 model provides the best agreement with the experimental data, such as hydration free energies of small molecules [212, 213]. Recently, it was demonstrated that an optimized GBMV2 model could provide a reliable description of both folded and unfolded protein conformations.[29] In particular, it shows minimal over-compaction bias in simulation of disordered proteins frequently associated with many implicit and explicit solvent protein force fields [50,

110, 124, 183, 214]. A key limitation to broader application of GBMV2, however, is that it is ~10 times slower than vacuum calculations and scales poorly to parallel multi-core executions.

One powerful technique to improve efficiency is the use of graphics processing units (GPUs) that can have thousands of parallel processing cores. GPU-accelerated algorithms available in many MD engines, such as CHARMM [215], AMBER [216, 217], GROMACS [218], NAMD [219, 220], and OpenMM [221], have offered up to two orders of magnitude speedup over traditional CPU-based codes. Some efforts have also been made on the GPU acceleration of GB implicit solvent models. The GB/OBC model in Amber has been implemented and achieved routine microsecond molecular dynamics simulations [222]. The GBSW model has also been implemented in a CHARMM/OpenMM module that displays around 100-fold improvement on the efficiency while maintaining similar numerical accuracy [223]. Notably, these early implementations only include the electrostatic solvation energy and thus might not be directly deployed for biomolecular simulations without the contribution of nonpolar solvation energy. Recently, an efficient pair-wise approximation of the solvent accessible surface area (SASA) was added into the GBSA/OBC GPU model, albeit with limited accuracy [224]. The correlation between atomic SASAs calculated by the GPU model and exact numerical results varies significantly from 0.54 to 0.91 for a number of test proteins.

Here, we report the implementation of an efficient GPU-accelerated GBMV2/SA algorithm in a CHARMM/OpenMM module. The implementation takes advantage of the similarities between GBMV2 and GBSW algorithms and builds on several existing

kernels of the GPU-GBSW module. The numerical scheme for computing the Born radius of each atom also allows for implementation of an efficient algorithm for calculating atomic surface areas. Together, the current implementation provides a complete realization of the GBMV2/SA model on GPUs, making it appropriate for general MD simulations of biomolecules. In the below sections of this paper, the detailed methodologies of GPU-GBMV2/SA algorithm are discussed, including the treatment of electrostatic and nonpolar solvation contributions, the lookup table algorithm for efficient volume integration, and the scheme of GPU implementations. Key points of the original GBMV2 model are highlighted. Furthermore, the accuracy and efficiency of GPU-GBMV2/SA are benchmarked against the CPU-GBMV2/SA implementation, and the remaining computational bottlenecks are also discussed. Finally, the conclusions and an outlook towards future work are given.

3.2 Method

3.2.1 Rigorous formulation

The implicit solvent model can be, in principle, derived rigorously from the explicit solvent model characterized by a probability function $P(\mathbf{X},\mathbf{Y})$ [200],

$$P(\mathbf{X},\mathbf{Y}) = \frac{e^{-\beta U(\mathbf{X},\mathbf{Y})}}{\int d\mathbf{X} d\mathbf{Y} e^{-\beta U(\mathbf{X},\mathbf{Y})}}. \quad \textbf{(3.1)}$$

Here, the coordinates **X** and **Y** represent the complete configuration of solute (*e.g.*, proteins) and solvent (*e.g.*, water), respectively. The $U(\mathbf{X},\mathbf{Y})$ is the potential of explicit solvent system, which can be usually decomposed into three terms,

$U(\mathbf{X},\mathbf{Y}) = U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})$, where the $U_{\text{p-p}}(\mathbf{X})$ is the intramolecular protein potential, $U_{\text{p-w}}(\mathbf{X},\mathbf{Y})$ is the protein-water interactions, and $U_{\text{w-w}}(\mathbf{Y})$ is the water-water interactions. We then can define a reduced probability function, $\overline{P}(\mathbf{X}) = \int d\mathbf{Y} P(\mathbf{X},\mathbf{Y})$, by integrating out the solvent degrees of freedom **Y**. It is observed that this reduced probability function is not dependent explicitly on the solvent degrees of freedom by taking an average influence of the solvent into consideration. In a canonical system at temperature *T*, this reduced probability function can be further written as follows,

$$\begin{aligned}
&\overline{P}(\mathbf{X}) = \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}}{\int d\mathbf{X} d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}} = \frac{e^{-\beta W(\mathbf{X})}}{\int d\mathbf{X} e^{-\beta W(\mathbf{X})}}, \\
&W(\mathbf{X}) = U_{\text{p-p}}(\mathbf{X}) + \Delta G_{\text{solv}}(\mathbf{X}), \qquad \textbf{(3.2)} \\
&\Delta G_{\text{solv}}(\mathbf{X}) = -\frac{1}{\beta} \ln\left(\frac{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-w}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta U_{\text{w-w}}(\mathbf{Y})}} \right),
\end{aligned}$$

where $\Delta G_{\text{solv}}(\mathbf{X})$ is defined as the solvation free energy for transferring a solute from the gas phase to a solvent phase, which includes the solvent-induced influence but is explicitly unknown. It is therefore critical to formulate an accurate reduced protein potential $W(\mathbf{X})$ or probability $\overline{P}(\mathbf{X})$, because each thermodynamic property (*A*) of solute system is then fundamentally calculated by an expectation $\langle A \rangle = \int d\mathbf{X} A(\mathbf{X}) \overline{P}(\mathbf{X})$.

3.2.2 Solvation free energy decomposition

In conventional implicit solvent models, the total solvation free energy is generally decomposed into electrostatic and nonpolar terms by designing a thermodynamic cycle (Figure 3.1),

$$\Delta G_{\text{solv}}(\mathbf{X}) = \Delta G_{\text{np}}(\mathbf{X}) + \Delta G_{\text{elec}}(\mathbf{X}), \tag{3.3}$$

where the nonpolar component involves the free energy cost to create the solute cavity in the solvent and turn on the nonpolar solute-solvent van der Waals (vdW) interaction, and the electrostatic solvation free energy is the cost of charging up the solute in the solvent.

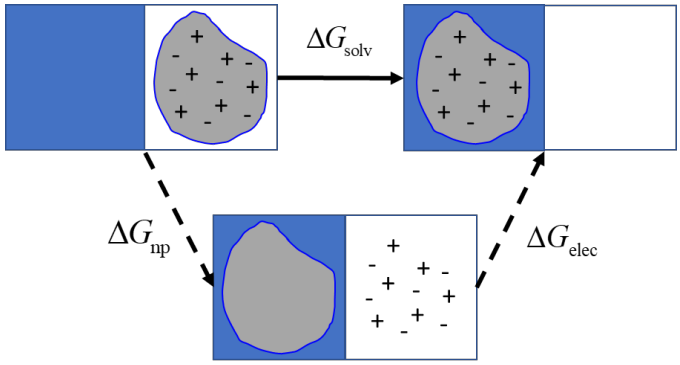

Figure 3.1 Thermodynamic cycle decomposes the solvation free energy into electrostatic (polar) and nonpolar components.

To obtain the explicit expression of each term, it is straightforward to write them into the following expressions,

$$e^{-\beta\Delta G_{\text{np}}(\mathbf{X})} = \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}^{\text{vdW}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{w-w}}(\mathbf{Y})\right]}}, \text{ and} \tag{3.4}$$

$$e^{-\beta\Delta G_{\text{elec}}(\mathbf{X})} = \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta\left[U_{\text{p-p}}(\mathbf{X}) + U_{\text{p-w}}^{\text{vdW}}(\mathbf{X},\mathbf{Y}) + U_{\text{w-w}}(\mathbf{Y})\right]}}. \tag{3.5}$$

Apparently, they can be calculated by a general potential energy function that decomposes the solvation free energy into the nonpolar and electrostatic components as below,

$$U(\mathbf{X},\mathbf{Y}) = U_{\text{p-p}}(\mathbf{X}) + U_{\text{w-w}}(\mathbf{Y}) + \lambda_{\text{vdW}} U_{\text{p-w}}^{\text{vdW}}(\mathbf{X},\mathbf{Y}) + \lambda_{\text{elec}} U_{\text{p-w}}^{\text{elec}}(\mathbf{X},\mathbf{Y}), \tag{3.6}$$

where the parameters λ_{vdW} and λ_{elec} can be turned on or off to calculate their solvation free energies explicitly.

3.2.3 Electrostatic solvation free energy and forces

The GB approximation developed by Still and coworkers [225] allows the electrostatic energy to be written as a pairwise summation,

$$\begin{aligned} \Delta G^{\text{elec}} &= -\frac{1}{2}\sum_{i,j} \tau_{ij} \frac{q_i q_j}{f_{ij}^{\text{GB}}}, \\ f_{ij}^{\text{GB}} &= \sqrt{\mathbf{R}_{ij}^2 + R_i^{\text{GB}} R_j^{\text{GB}} \exp\left(-\mathbf{R}_{ij}^2 \big/ K_s R_i^{\text{GB}} R_j^{\text{GB}}\right)}. \end{aligned} \tag{3.7}$$

where $\tau_{ij} = \left(1/\varepsilon_{\text{solute}} - \exp\left(-\kappa f_{ij}^{\text{GB}}\right)\big/\varepsilon_{\text{solvent}}\right)$, and q_i and R_i^{GB} are the atomic charge and effective Born radius of the atom *i*, respectively, κ is a Debye-Hückel screening parameter, and K_s is an empirical constant that is set to 8 in the GBMV2 model [146, 147]. The effective Born radius is defined as the radius of an equivalent spherical cavity that yields the same atomic self-polarization free energy. It is thus a function of the positions of all solute atoms. The pairwise GB expression allows analytical evaluation of atomic forces and is thus particularly suitable for MD simulations.

The GB forces with respect to the atomic positions include two terms,

$$F_a^{\text{elec}} = -\frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_a} = -\left(\sum_{ij} \frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_{ij}} \frac{\partial \mathbf{R}_{ij}}{\partial \mathbf{R}_a} + \sum_i \frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}} \frac{\partial R_i^{\text{GB}}}{\partial \mathbf{R}_a}\right), \tag{3.8}$$

where

$$\sum_{ij}\frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_{ij}}\frac{\partial \mathbf{R}_{ij}}{\partial \mathbf{R}_a}=$$
$$-\sum_i\left(\tau_{ia}-\frac{\kappa\exp\left(-\kappa f_{ia}^{\text{GB}}\right)f_{ia}^{\text{GB}}}{\varepsilon_{\text{solvent}}}\right)\frac{q_iq_a\left[1-\exp\left(-\mathbf{R}_{ia}^2/K_sR_i^{\text{GB}}R_a^{\text{GB}}\right)/K_s\right]}{\left(f_{ia}^{\text{GB}}\right)^3}\left(\mathbf{R}_i-\mathbf{R}_a\right),\quad\text{and}\qquad\textbf{(3.9)}$$

$$\sum_i\frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}}\frac{\partial R_i^{\text{GB}}}{\partial \mathbf{R}_a}=$$
$$\frac{1}{2}\sum_i\left[\sum_j\left(\tau_{ij}-\frac{\kappa\exp\left(-\kappa f_{ij}^{\text{GB}}\right)f_{ij}^{\text{GB}}}{\varepsilon_{\text{solvent}}}\right)\frac{q_iq_j\exp\left(-\mathbf{R}_{ij}^2/K_sR_i^{\text{GB}}R_j^{\text{GB}}\right)}{\left(f_{ij}^{\text{GB}}\right)^3}\left(R_j^{\text{GB}}+\frac{\mathbf{R}_{ij}^2}{K_sR_i^{\text{GB}}}\right)\right]\frac{\partial R_i^{\text{GB}}}{\partial \mathbf{R}_a}.\qquad\textbf{(3.10)}$$

It can be seen that the GB energy and forces depend on the effective Born radii, R_i^{GB}, and their derivatives with respect to atomic positions, $\partial R_i^{\text{GB}}/\partial \mathbf{R}_a$.

3.2.3.1 Born radii and their derivatives

Computing the effective Born radius of each solute atom is a key step for calculating the GB electrostatic solvation free energy. In GBMV2 model, the calculation of a given Born radius considers the contributions from the Coulomb field approximation and an empirical high-order correction term:

$$R_i^{GB}=\frac{P_1}{G_i}+P_2,\ \text{and}\qquad\textbf{(3.11)}$$

$$G_i=\left(1-\frac{1}{\sqrt{2}}\right)\frac{1}{4\pi}\int\frac{V(\mathbf{r})}{\left|\mathbf{r}-\mathbf{R}_i\right|^4}d\mathbf{r}+\left(\frac{1}{4\pi}\int\frac{V(\mathbf{r})}{\left|\mathbf{r}-\mathbf{R}_i\right|^7}d\mathbf{r}\right)^{1/4},\qquad\textbf{(3.12)}$$

where P_1 and P_2 are empirical fitting coefficients, $\mathbf{R}_i$ are atomic coordinates, and $V(\mathbf{r})$ is the molecular volume function [146, 147]. Optimal values of P_1 and P_2 are obtained by linear regression fitting of atomic GB radii of model proteins to the reference values obtained from high-resolution PB calculations [124, 148, 207, 226]. Detailed

expressions for derivatives of Born radii are given in the section 3.5.1 of supporting information.

3.2.3.2 Analytical approximation of the molecular volume

The molecular volume (MV) is defined as the solute volume that is formed by rolling a water probe on the solute [211]. Two methods have been previously implemented in the CPU version of GBMV2 [146, 147]. One is to use arbitrarily precise numerical grids for a highly accurate calculation of Born radii; but this method is computationally expensive, does not provide an analytical gradient, and thus is not suitable for efficient MD simulations. The other method introduces an efficient analytical approximation to the MV with comparable precision of calculating Born radii, which is also suitable for GPU acceleration. The molecular volume is given by a Fermi-Dirac switching function from a preprocessed "raw" molecular volume, $S(r)$,

$$V\left(\mathbf{r}_{mn}+\mathbf{R}_i\right)=\frac{1}{1+\exp\left[\beta\left(S\left(\mathbf{r}_{mn}+\mathbf{R}_i\right)-\lambda\right)\right]}, \tag{3.13}$$

where β and λ are the parameters that represent the width and midpoint of the switching function, respectively.

The expression of $S(\mathbf{r})$ in the GBMV2 model involved two terms,

$$\begin{aligned} &S\left(\mathbf{r}\right)=S_{\mathrm{vdW}}\left(\mathbf{r}\right)+S_{\mathrm{MV2}}\left(\mathbf{r}\right),\quad S_{\mathrm{vdW}}\left(\mathbf{r}\right)=2\sum_j F_{\mathrm{vdW}}\left(\mathbf{r}\right),\\ &S_{\mathrm{MV2}}\left(\mathbf{r}\right)=S_0\left[\sum_j F_{\mathrm{MV2}}\left(\left|\mathbf{r}-\mathbf{R}_j\right|\right)\right]\frac{\sum_j\left|\mathbf{r}-\mathbf{R}_j\right|^2 F_{\mathrm{MV2}}^2\left(\left|\mathbf{r}-\mathbf{R}_j\right|\right)}{\left|\sum_j\left(\mathbf{r}-\mathbf{R}_j\right)F_{\mathrm{MV2}}\left(\left|\mathbf{r}-\mathbf{R}_j\right|\right)\right|^2}, \end{aligned} \tag{3.14}$$

where $S_{\mathrm{vdW}}(\mathbf{r})$ is the vdW volume contribution and $S_{\mathrm{MV2}}(\mathbf{r})$ includes a vector-based scaling term to account for the discrepancy between vdW and MV volumes. There

are two significant points: One is that the atomic volume function, F_{MV2} (**r**), has a longer tail compared to the F_{vdW} (**r**), in order to probe more overlapping regions between atoms. The other is that the representation of MV. For the vdW volume, because S_{vdW} (**r**) is a monotonic function with the number of atoms, the summation can be immediately terminated when its value exceeds a certain cutoff. S_{MV2}(**r**), however, contains vector-based scaling approximation (VSA) term that helps to distinguish the "gap" (between atoms) and "open" (otherwise) regions, which is required to consider all atoms in proximity. As such, GBMV2 is considerably more expensive than GBSW, especially for small systems.

Additional details of the GBMV2 algorithms can be found in the section 3.5.1 of supporting information and the original paper [147]. Importantly, it can be seen that the next step is to calculate the S(**r**) at each numerical integration grid point, which can be accelerated by a lookup table algorithm (see section 3.2.5).

3.2.4 Solvent accessible surface area nonpolar solvation free energy and forces

The nonpolar energy can be decomposed into a short-range repulsive energy and long-range solute-solvent dispersion energy, and is, in the first order approximation, proportional to SASA [147, 149]. Thus, the nonpolar energy in the GBMV2 model is estimated as,

$$\Delta G_{\text{np}} = \sum_i \gamma_i A_i, \tag{3.15}$$

where the γ_i and A_i is the effective surface tension coefficient and SASA of each atom, respectively. The surface coefficient is often assumed to be same for all atom

types, reducing Equation 3.15 to $\Delta G_{\text{np}} = \gamma \sum_i A_i$. This linear approximation has been shown to provide an adequate description of nonpolar solvation energy for many biomolecular applications [137, 147].

3.2.4.1 Atomic SASA and vdW volume

The atomic SASA can be expressed as:

$$A_i = \int_{|\mathbf{r}-\mathbf{R}_i|=R_i^{\text{vdW}}+R_{\text{w}}} f\left(\overline{V}_i(\mathbf{r})\right) d\mathbf{r}, \tag{3.16}$$

where the excluded volume $\overline{V}_i(\mathbf{r}) = \sum_{j\neq i} V_j(\mathbf{r})$ involves volumes for all atoms except for atom *i*, and the smooth function *f* represents the exposed rate at **r** point, which should be one if the excluded volume is zero, and it should be zero if the sum of excluded volume is one. In the GBMV2/SA model, an analytic expression of the vdW volume is used,

$$\overline{V}_i(\mathbf{r}) = \sum_{j\neq i} 2f(u_j), \tag{3.17}$$

the u_j and exposed function *f* are written as,

$$f(u) = \begin{bmatrix} 1 & u \le 0 \\ 1-10u^3+15u^4-6u^5 & 1>u>0 \\ 0 & u \ge 1 \end{bmatrix}, \text{and} \tag{3.18}$$

$$u_j = \frac{\left|\mathbf{r}-\mathbf{R}_j\right|^2 - \left(R_j^{\text{vdW}}+t_-^{\text{vdW}}\right)^2}{\left(R_j^{\text{vdW}}+t_+^{\text{vdW}}\right)^2 - \left(R_j^{\text{vdW}}+t_-^{\text{vdW}}\right)^2}, \tag{3.19}$$

where R_j^{vdW} is the vdW radius of *j* atom and R_{w} is the radius of solvent molecule, for the water molecule, which is 1.4 Å. The switching widths, t_+^{vdW} and t_-^{vdW}, have been optimized to 1.2 and 1.5 Å, respectively, for R_{w} = 1.4 Å.

The general integral of Equation 3.16 cannot be solved analytically. However, a straightforward numerical expression is given as follows,

$$A_i \approx 4\pi\left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right)^2 \sum_m w_m f\left(\overline{V}_i\left(\mathbf{r}_m + \mathbf{R}_i\right)\right), \tag{3.20}$$

where the excluded molecular volume at each grid point is determined quickly by the lookup table algorithm described below. Detailed derivations of the nonpolar energy and forces term can be found in the section 3.5.2 of supporting information.

3.2.5 Implementation algorithms and parallelization

3.2.5.1 Numerical integration

The important component of computing Born radius is to evaluate the 3-dimension integrals shown in the Equation 3.12. In the GBMV2 and GBSW models, the integrals are evaluated using numerical quadrature, where they are split up into radial and angular components [146, 147, 226]. The radial integral is approximated by Riemann-Stieltjes summation with the standard set of radial grid points, while the angular integral is calculated by the Lebedev quadrature.

$$\begin{aligned}\frac{1}{4\pi}\int\frac{V(\mathbf{r})}{\left|\mathbf{r}-\mathbf{R}_i\right|^k}d\mathbf{r} &= \frac{1}{4\pi}\int\frac{V(\mathbf{r}+\mathbf{R}_i)}{\left|\mathbf{r}\right|^k}d\mathbf{r} \\ &\approx \frac{1}{(k-3)\left(R_i^{\mathrm{eff}}\right)^{k-3}} - \sum_{m\in\mathrm{rad}}\sum_{n\in\mathrm{ang}} w_{mn}^k V\left(\mathbf{r}_{mn}+\mathbf{R}_i\right),\end{aligned} \tag{3.21}$$

where w_{mn}^{k} is the weight of each grid point r_{mn} and R_i^{eff} is an effective integration starting point less than the vdW radius of each atom, in order to avoid the singularity of integrals. It is noted that the precise definition of the (solute) molecular volume in Equation 3.21 is a key quantity in determining the Born radii. The vdW-like volume employed in GBSW is simple and efficient to evaluate, and it provides stable forces [145]. However, it generates small and unphysical solvent-inaccessible high dielectric regions inside the solute, leading to an over-estimation of solvation free energy and a systematic over-stabilization of nonspecific compact conformations [148, 207]. This critical shortcoming is effectively solved by adopting an approximate Lee-Richards molecular volume in GBMV2.

3.2.5.2 Lookup table algorithm

The numerical volume integrations in GBMV2 (and GBSW) require quick access of all atoms within a certain distance that could contribute to the volume function. This is enabled by constructing a lookup table [145-147]. Specifically, the lookup table contains a spatially uniform cubic grid enclosing all solute atoms. At each grid point, all the atoms that are less than a certain distance, $R_{\max}$, are stored in a lookup table array,

$$\left|\mathbf{r}-\mathbf{R}_i\right| \le R_{\max} = \max\left[R_i^{\text{vdW}}\right] + 2.1 + \frac{\sqrt{3}}{2}c + R_{\text{buffer}}, \qquad \textbf{(3.22)}$$

where c is the width of the grid cell, the value 2.1 Å is the length of the tail of the atomic function $F_{\text{MV2}}(\mathbf{r})$, and R_{buffer} is an adjustable length that determines how far any atom can move before rebuilding the lookup table. The default value of R_{buffer} is zero, meaning the lookup table will be updated at each simulation step. By using the lookup

procedure, the cost of computing the molecular volumes is reduced to linear scaling with the number of grid points. It is noted that the number of neighbor atoms at each grid point is much larger in GBMV2 than GBSW due to the longer tail of atomic function, which contributes to a two to three-fold computational cost increase.

3.2.5.3 Parallelization and CUDA implementation

The existing GBSW kernels were adapted for the implementation of GPU-GBMV2/SA. As a plugin of CHARMM/OpenMM program, the overall design of the GPU-GBSW model is considered as a stand-alone solvent model in the OpenMM library [223]. It contains eight kernels, four of which are used to implement the lookup table, and the other four are used to calculate the electrostatic solvation energies and forces of hydrogen and non-hydrogen atoms. Kernels to support the lookup table were directly modified to support a larger value of R_{max} and the greater table depth required for GBMV2. In GBMV2, hydrogens have non-zero input radii and do not need to be treated separately. As such, the GBMV2 electrostatic term only requires three kernels (see Table 3.1). A new kernel, calcSASA, was developed to calculate atomic SASA and forces. The GPU algorithm for computing SASA terms is similar, where the number of blocks is equal to the number of atoms and threads loop over all quadrature integration grip points. Note that the calcSASA kernel is an independent kernel that can be used for both GPU-GBMV2 and GPU-GBSW models.

The CUDA implementation has similar algorithmic construction to the CUDA-GBSW model, and thus can be directly based on the existing CUDA platform [223]. Detailed algorithms for computing the electrostatic solvation energies and forces can be

found in the section 3.5.3 of supporting information, and the description of nonpolar term is briefly outlined because of the similar algorithms. Overall, these algorithms are analogous to the multi-processor tasks with CPU languages, as what previous GBSW paper stated. [223] To minimize the modification of GBSW CUDA implementation, each block is designed for each atom likewise. However, each thread is not used to each quadrature point, and instead we used the optimal 256 threads to loop all quadrature points per block. Meanwhile, high-speed shared memory was used for the sum reduction of quadrature points and neighbor atoms. Different from the CUDA-GBSW, the Born radii gradients are not saved, because its length is much larger resulted from the longer tail of atomic volume function. Instead, some shorter intermediate arrays (see section 3.5.3 of supporting information) are saved into the global memory in order to reduce the computational complexity of electrostatic solvation forces.

Table 3.1 Layout of key kernels for GPU-GBMV2/SA. Kernels for creating a lookup table array are similar to those used in GPU-GBSW.

Kernels	Description
calcBornR	To calculate the Born radius of each atom and save the temporary variables for the rapid calculations of the electrostatic forces. Each block is assigned to one atom, and 256 threads are used to loop over all the grid points. The equations can be found in the electrostatic energies part of section 3.5.1 of supporting information.
computeGBMVForce	To calculate GB electrostatic energies and the derivatives with respect to atomic coordinates.
reduceGBMVForce	To calculate the electrostatic forces. Each block is assigned to one atom, and 256 threads are used to loop all the grid points. The equations can be found in the electrostatic forces part of section 3.5.1 of supporting information.
calcSASA	To calculate the nonpolar energies and forces. Each block is assigned to one atom, and 256 threads are used to loop all the grid points. The equations can be found in the section 3.5.2 of supporting information.

3.2.6 Computational details

The correctness and accuracy of GPU-GBMV2/SA were mainly assessed by its ability to reproduce atomic energies and forces of the original CPU-GBMV2/SA implementation in CHARMM as well as PB-derived atomic self-solvation free energies. The model systems include the set of 22 small proteins previously used for the numerical parametrization of the original GBSW and GBMV models [147, 207]. The accuracy of GPU-GBMV2/SA was also validated by examining the interaction energy profiles between selected sidechains, in comparison to explicit solvent results from previous works [148, 207]. The numerical stability of the GPU-GBMV2/SA model was assessed by examining the energy conservation properties under different configurations. Furthermore, a small helical model peptide, $(AAQAA)_5$, was used to examine the stability of GPU-GBMV2/SA in long-time MD simulations and its ability to recapitulate the peptide conformational equilibrium. For this purpose, two distinct initial structures, an ideal helix and a fully extended conformation, were used to initiate independent control and folding simulations, allowing a rigorous diagnosis of convergence. A time step of 2 fs was used. The previously optimized GBMV2/SA protein force field was used, and the results were directly compared with those from CPU simulations [124].

The efficiency of the GPU versus CPU versions of GBMV2/SA was benchmarked using five folded proteins ranging from 856 to 77,304 atoms as well as an intrinsically disordered protein, the N-terminal transactivation domain (TAD) of p53 (926 atoms). The initial structures of folded proteins were downloaded from the Protein Data Bank (PDB) and then energy minimized followed by 5,000 steps of NVT equilibration.

The initial structure of p53-TAD was taken from a previous study [227]. Default GBMV2/SA parameters were used in all calculations, except for three keywords, beta = -12, P3 = 0.65, P6 = 8, which correspond to β, S_0, and K_s, in Equations 3.13 and 3.7, respectively. The input radius of each atom is obtained from the previously optimized GBMV2 force field [124]. The cutoff distance for nonbonded interactions was set at 20 Å and a time step of 2 fs was used. All GPU simulations were done on an NVIDIA TITAN X (Pascal) graphics card, and CPU calculations were carried out on an Intel Xeon E5-2620 v4 2.10GHz CPU. For CUDA calculations, the performance analysis of important kernels was also reported in Figure 3.9, including threads per block, registers per thread, and theoretical vs. achieved occupancy etc.

3.3 Results and Discussion

3.3.1 Electrostatic solvation energies and forces

Proper GPU implementation of the GBMV2 is first assessed by its ability to reproduce the atomic electrostatic self-solvation energies and forces. As summarized in Figure 3.2, atomic self-solvation energies and forces of all 22 small proteins are essentially identical between the GPU and original CPU implementations. The numerical differences between CPU and GPU results (see inserts) are extremely small, completely negligible compared to the absolute GB electrostatic energies and forces. This demonstrates that the electrostatic solvation term of GPU-GBMV2/SA has been implemented correctly in the CUDA platform.

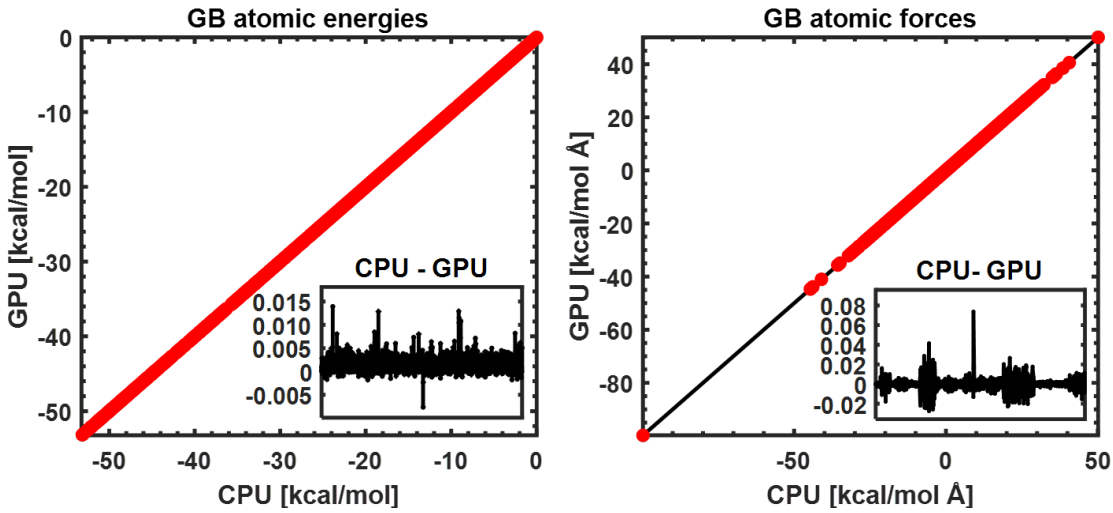

Figure 3.2 Accuracy of GPU-GBMV2/SA atomic electrostatic self-solvation energies (left) and forces (right), compared with those of CPU-GBMV2. The diagonal line (y=x) is shown for reference. All atoms of 22 small proteins are included in this comparison. The inserted panels show the difference between CPU and GPU results (in the same unit, kcal/mol or kcal/mol Å for each of all atoms from the protein test set.

We also validated that atomic self-solvation energies provided by GPU-GBMV2 are consistent with PB-derived results, which is a key indicator of the quality of a GB implicit solvent model. Given the numerical equivalence of GPU- and CPU-GBMV2 models, GPU-GBMV2 should achieve a similar correlation with PB. Indeed, as summarized in Figure 3.3, the correlation coefficient between effective Born radii derived from PB and GPU-GBMV2 is 0.9985, consistent with the results of CPU-GBMV2 [147]. We note that the superb ability of GBMV2 to reproduce PB is attributed to both the higher order correction to the Coulomb field approximation (Equation 3.12) and effective approximation of SMV (Equation 3.14) [147].

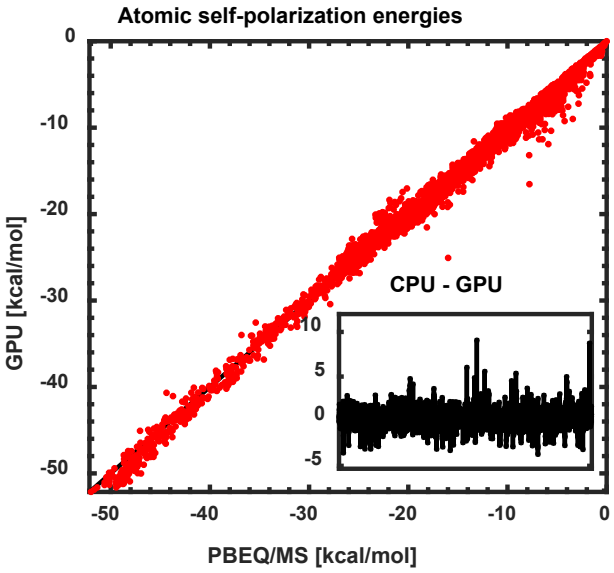

Figure 3.3 Atomic electrostatic self-solvation energies derived from GPU-GBMV2 versus PB. All atoms from 22 small proteins are included. The insert shows the difference for each atom.

3.3.2 Nonpolar solvation energy and forces

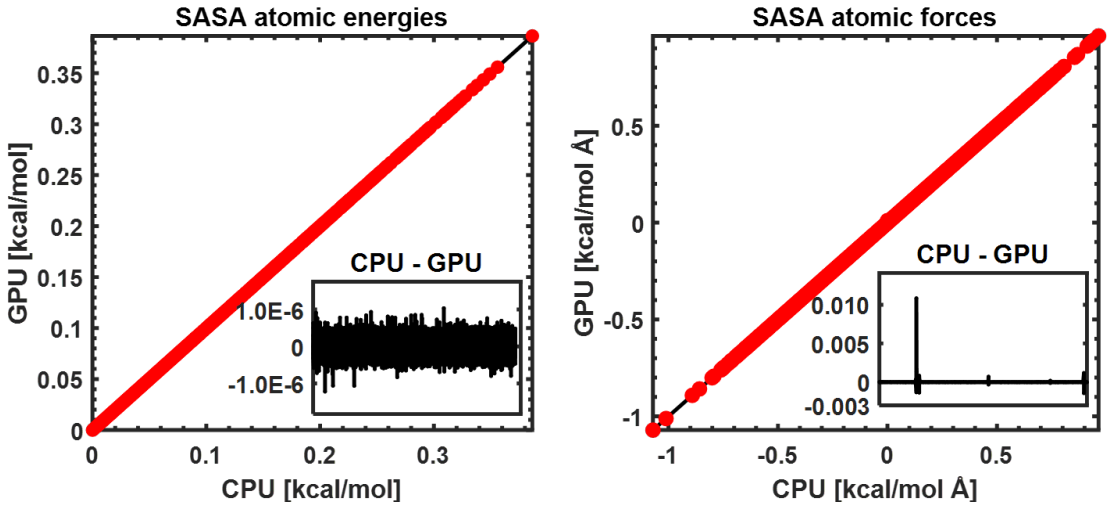

Figure 3.4 The accuracy of GPU and CPU-GBMV2/SA in calculating atomic SASA energies (left) and forces (right). The surface tension coefficient is 5 cal/mol Å^2. All atoms from 22 small proteins are included. The inserted panels show the difference between CPU and GPU results (in the same unit, kcal/mol or kcal/mol/Å for each of all atoms from the protein test set.

Nonpolar solvation energy plays important roles in driving the conformations of proteins, although it makes smaller contributions to the total solvation energies compared to the GB term. Figure 3.4 shows that the nonpolar energies and forces of GPU-GBMV2/SA are also numerically equivalent to those calculated by the original CPU-GBMV2/SA, indicating that both SASA energies and forces have been implemented in the present CUDA platform correctly. As such, it can be expected that the errors of nonpolar energies are on the order of 1 - 2% compared with the exact SASA analytic model for proteins [147]. The successful implementation of the SASA term in the CUDA platform provides a complete GPU-GBMV2/SA implicit solvent model that can now be readily deployed for biomolecular simulations. In addition, it also paves the way for the future development of better nonpolar solvation models, such as by including the dispersion contribution [149].

3.3.3 Energy conservation and numerical stability

After establishing the correctness of the GPU implementation, we evaluated the numerical stability of GBMV2/SA by examining the energy conservation properties in NVE simulations with three different surface tension parameters (γ). As summarized in Figure 3.5, the energies from CPU and GPU calculations display similar trends for all three cases, suggesting that the GPU version has similar numerical stability compared to the CPU version. The energy drifts over 300 ps are significant, but in line with a previous analysis of the numerical stability of GBMV2 on CPU [209]. The energy fluctuations in GPU calculations (after removing the linear drift) are slightly higher than those in CPU runs, likely due to the use of mixed single/double precisions. Comparison of the energy

conservation properties from simulations with different γ show that SASA as implemented is numerically highly stable. We note that GBMV2 is numerically less stable compared to GBSW because of the sharp molecular surface definition as well as the VSA term. Nonetheless, peptide simulations suggest that GBMV2 can be reliable even with a 2-fs time step with a proper thermostat in NVT simulations, showing no sign of numerical instabilities or any significant artifacts in the resulting trajectories [124].

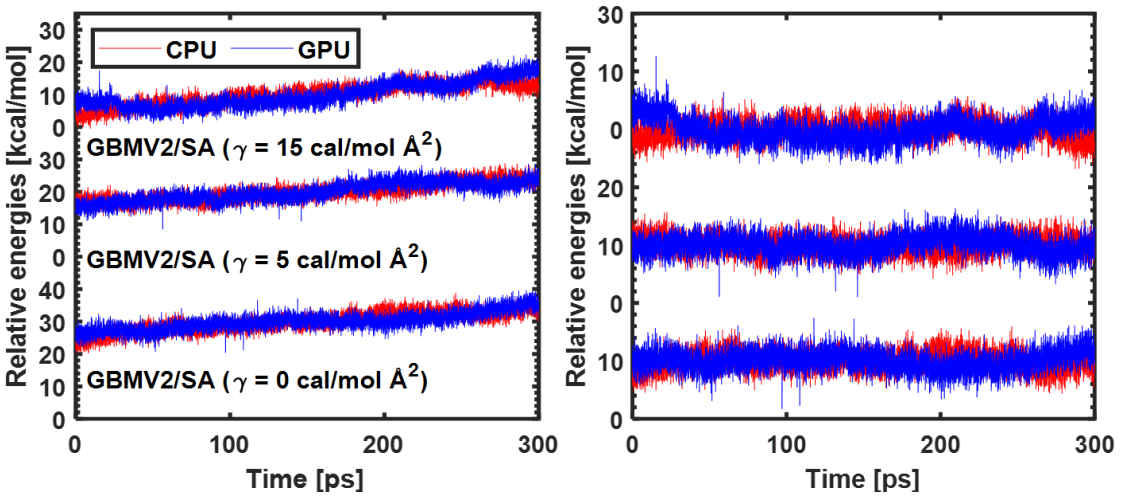

Figure 3.5 Energy conservation of MD simulations for a small protein (PDB: 1BDC) in CPU- and GPU-GBMV2/SA. Energies versus simulation time before (left) and after (right) removing the linear drift. The time step was set to 1 fs. The relative CPU/GPU energy drift rates are 0.0072/0.0085, 0.0048/0.0068 and 0.0071/0.0110 (unit: % / ps) for three cases (γ = 0, 5, 15 cal / mol Å²), respectively. The standard fluctuations of CPU/GPU energies (after removing the linear drift) are 1.5434/1.5942, 1.4566/1.5963, and 1.5934/2.0047 kcal/mol), for three cases, respectively. Only the last 100 ps trajectories were included in the energy drift analysis.

3.3.4 Sidechain interaction and peptide folding simulations

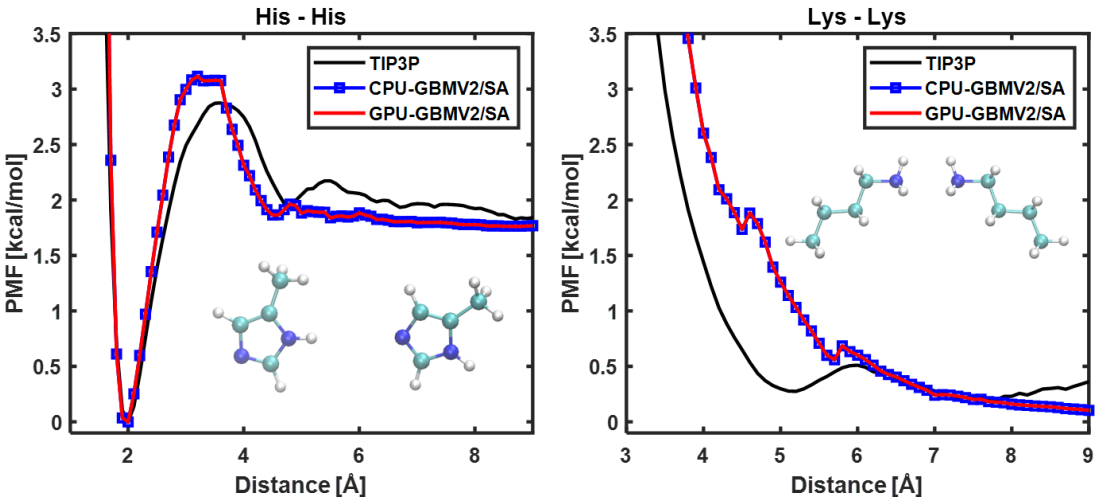

Figure 3.6 Free energy profiles of interactions for two sidechain pairs, (left) His – His and (right) Lys – Lys, in TIP3P, CPU- and GPU-GBMV2/SA solvent. γ = 5 cal/mol Å^2 was used.

Before applying GPU-GBMV2/SA to protein simulations, we first validated its ability to accurately describe interactions between various backbone and side chain chemical groups. The balance of these interactions governs the ability of a force field to properly capture the protein conformational equilibria. Figure 3.6 compares the free energy profiles of two representative sidechains pairs. It demonstrates that GPU-GBMV2/SA exactly reproduces CPU-GBMV2/SA as expected, and the implicit solvent results also closely match the profiles derived from free energy calculations in TIP3P explicit solvent [148].

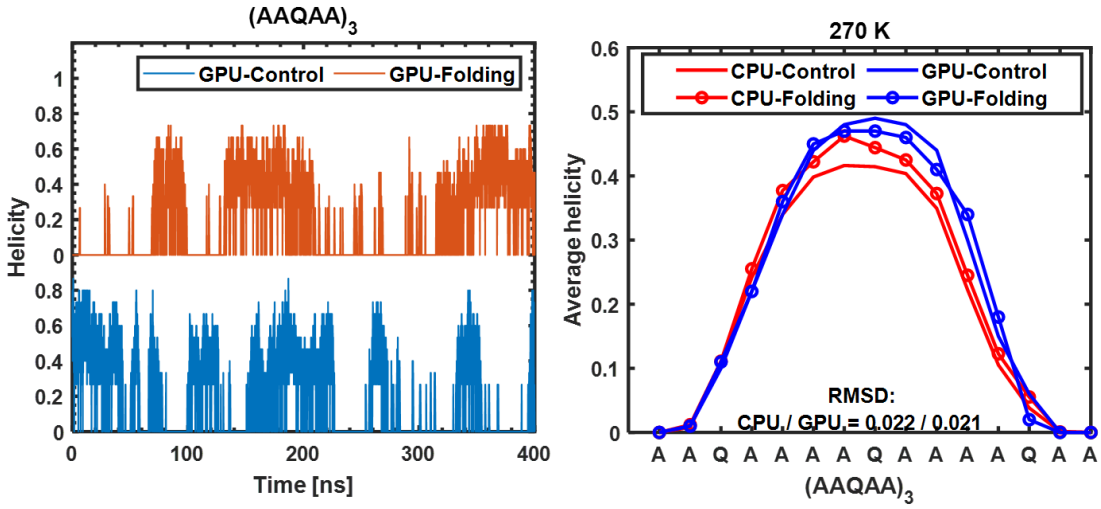

Figure 3.7 Left: Helicity of $(AAQAA)_3$ during folding and control GPU-GBMV2/SA simulations at 270 K. Right: Average residue helicity profiles calculated from GPU simulations in comparison with previous results derived from CPU simulations.[26] The RMSD values shown are the root-mean-square differences between profiles derived from control and folding simulations.

The peptide $(AAQAA)_3$ has been widely used as a model flexible peptide for force field evaluation and calibration [124, 148, 207]. Figure 3.7 shows the time evolution of helicity of $(AAQAA)_3$ during two independent control and folding simulations at 270 K in GPU-GBMV2/SA. It can be observed that several reversible conformational transitions between the (partial) helices and unfolded structures were sampled in both simulations within 200 ns, indicating that the implicit treatment of solvent using the GBMV2/SA model greatly facilitates protein conformational sampling without the friction from explicit solvent molecules. The resulting average residue helicity profiles are well- converged; the RMSD value between results from control and folding GPU runs is only 0.021. These results are comparable to results derived from previous replica exchange simulations on a CPU platform [124].

3.3.5 Computational efficiency

Figure 3.8 summarizes the performance of GPU-GBMV2/SA in comparison to the CPU version for six folded and unfolded proteins of various sizes and topologies. It shows that the GPU version offers ~ 60 to 70-fold speed up, with the larger systems exhibiting slightly superior efficiency. We note that a faster version of CPU-GBMV2/SA has been previously developed, which extensively utilizes pre-calculated data arrays to speed up the evaluation of Born radii and derivatives [209]. Our current testing shows that the fast CPU version is ~50% more efficient than the standard one. Additionally, the performance with multicores seems to be less powerful compared to that using one GPU and has a poor scaling with the number of cores. For example, the 12-core multiprocessing calculations only gain around 6-fold speed up (see Table 3.4). One possible reason is that each of parallel tasks is only used to loop the number of atoms for current CPU parallelization, while these tasks to loop the quadrature points have been further distributed into branch of threads per block on a GPU. We have also profiled the timing distribution of each kernel in GPU-GBMV2/SA. The four kernels associated with the lookup table account for only ~5% of the time, although it is memory intensive. The calculations of electrostatic and nonpolar terms take up around 85% and 7% of the total time, respectively. Thus, the bottleneck of the GBMV2/SA algorithm is clearly the calculation of Born radii and their derivatives. The reason is that the calculation of the Born radius for each atom involves a complicated expression based on around 800 numerical quadrature points and 100 neighbor atoms for each grid point; the derivatives of Born radii involve even more extensive operations (see detailed expressions in the

section 3.5.1). Consequently, the GB force calculations are about three-fold slower than the GB energies calculations.

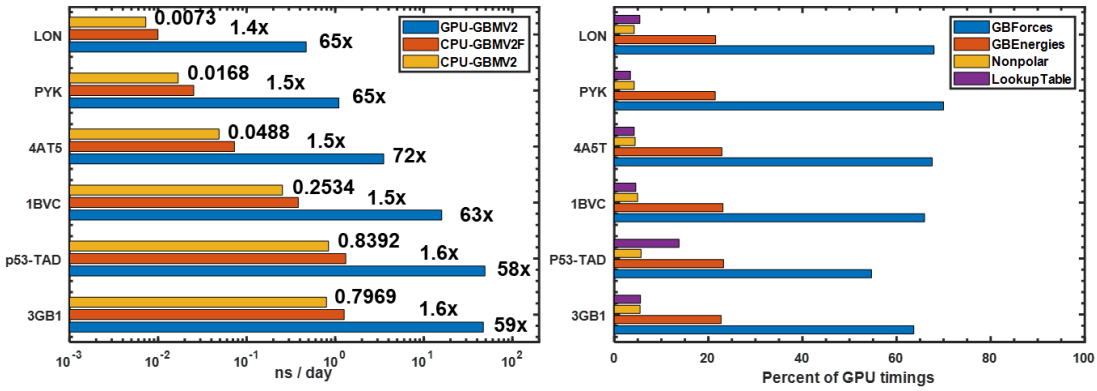

Figure 3.8 (Left) Timings of CPU- and GPU-GBMV2/SA simulations. The numbers next to the CPU-GBMV2/SA bars are the production time in ns/day, and the ratios next to the fast CPU-GBMV2/SA and GPU-GBMV2/SA are folds of speedup compared to CPU-GBMV2/SA. The production rates of GPU simulations are (in ns/day): 47.00 (3GB1), 48.96 (p53-TAD), 15.93 (1BVC), 3.52 (4AT5), 1.10 (PYK) and 0.47 (LON). (Right) Percentages of time spent in various parts of GPU-GBMV2/SA calculation, including constructing and updating the lookup table ("Lookup Table"), nonpolar energies and forces ("Nonpolar") and electrostatic energies and forces calculations ("GBEnergies" and "GBForces"). The GPU and CPU calculations were done on one NVIDIA TITAN X (Pascal) and one core of Intel Xeon E5-2620 v4 2.10GHz CPU, respectively.

3.4 Conclusions

A GPU-accelerated GBMV2/SA model has been implemented within the CHARMM/OpenMM interface, including both the GB electrostatic and SASA nonpolar solvation terms. The GB term has been implemented based on the existing CUDA kernels of the GPU-GBSW model [223]. Together with a SASA nonpolar term, it provides a complete and accurate GBMV2/SA implicit solvent model that is suitable for protein simulations. Results show that the GPU-GBMV2/SA solvation energies and forces are essentially the same as those in the original CPU-GBMV2/SA model with negligible errors, giving rise to similar energy conservation properties. Benchmarks based on a set

of folded and unfolded proteins show that the current implementation of GPU-GBMV2/SA offers about 60 to 70-fold speedup on a single NVIDIA TITAN X graphics card compared to a single core of an Intel Xeon E5-2620 v4 2.10GHz CPU. While the speedup is somewhat modest compared to those achieved by GBSW or GBSA/OBC in Amber, it is still quite substantial and will enable the application of GBMV2 for MD of larger systems and for longer timescales. for both folded and unfolded proteins.

We note that there is still room for further improvement of the computational efficiency of GPU-GBMV2/SA. For example, a key bottleneck is the large lookup table required for evaluating the volume integrals due to longer tails required for analytical approximation of MV. The numbers of atoms within the proximity of each grid point can be as high as ~100. It is likely that the list can be truncated without significant reduction to numerical accuracy. One can also optimize the usage of computational memory of lookup table array, *e.g.*, by using the flexible allocation or avoiding the allocation by looping neighbor grid boxes. Development of the GPU-GBMV2/SA algorithm will also allow one to perform extensive folding simulations of model proteins and peptides to critically evaluate the ability of the simple SASA nonpolar model for describing the conformation equilibria [149]. This will pave the way for further development of better treatments of the nonpolar solvation that can more accurately capture the conformational dependence of solvation free energies.

3.5 Supporting information

3.5.1 Electrostatic solvation energy and forces

The electrostatic solvation energies in a low concentration of salt are described as follows,

$$\begin{aligned} \Delta G^{\text{elec}} &= -\frac{1}{2}\sum_{i,j}\tau_{ij}\frac{q_i q_j}{f_{ij}^{\text{GB}}}, \\ \tau_{ij} &= \left(1/\varepsilon_{\text{solute}} - \exp\left(-\kappa f_{ij}^{\text{GB}}\right)\Big/\varepsilon_{\text{solvent}}\right), \\ f_{ij}^{\text{GB}} &= \sqrt{\mathbf{R}_{ij}^2 + R_i^{\text{GB}} R_j^{\text{GB}} \exp\left(-\mathbf{R}_{ij}^2 \big/ K_s R_i^{\text{GB}} R_j^{\text{GB}}\right)}, \end{aligned} \tag{3.23}$$

where q_i and R_i^{GB} the partial charge and Born radius of i^{th} atom, $\mathbf{R}_{ij}$ is a distance vector between two atoms, K_s is usually set to 8 for GPU-GBMV2/SA electrostatic calculations, $\varepsilon_{\text{solute}}$ and $\varepsilon_{\text{solvent}}$ are the dielectric constant of solute and solvent, respectively, and κ is a Debye-Hückel screening parameter.

In GBMV2/SA model, the Born radii are related to the molecular volume by considering the numerical Coulomb and high-order correction terms.

$$\begin{aligned} R_i^{GB} &= \frac{P_1}{a_0 G_i^0 + a_1 G_i^1} + P_2, \\ G_i^0 &= \frac{1}{R_i^{\text{eff}}} - \sum_n w_n^0 V\left(\mathbf{r}_n + \mathbf{R}_i\right),\ G_i^1 = \left(\frac{1}{4\left(R_i^{\text{eff}}\right)^4} - \sum_n w_n^1 V\left(\mathbf{r}_n + \mathbf{R}_i\right);\right)^{1/4}, \end{aligned} \tag{3.24}$$

where the parameters of Born radii are $P_1 = 0.9085$, $P_2 = -0.102$ Å, $a_0 = 1 - 1/\sqrt{2}$, and $a_1 = 1$, $\mathbf{r}_n$ are the coordinates of grid points, $\mathbf{R}_i$ are the atomic coordinates, w_n^0 are the grid weights of the CFA term, and w_n^1 are the grid weights of the correction term, and R_i^{eff} are the effective atomic radii used for the quadrature integrals.

The molecular volume has a complicated expression.

$$V\left(\mathbf{r}_n+\mathbf{R}_i\right)=\frac{1}{1+\exp\left[\beta\left(S\left(\mathbf{r}_n+\mathbf{R}_i\right)-\lambda\right)\right]},$$
$$S\left(\mathbf{r}_n+\mathbf{R}_i\right)=S_0X_1\frac{X_2}{\left(\mathbf{X}_3\right)^2}+2X_4,\ \mathbf{t}_{nij}=\mathbf{r}_n+\mathbf{R}_i-\mathbf{R}_j, \tag{3.25}$$

and four intermediate volumes are written as follows,

$$X_1\left(\mathbf{r}_n+\mathbf{R}_i\right)=\sum\nolimits_j F_{\mathrm{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right),$$
$$X_2\left(\mathbf{r}_n+\mathbf{R}_i\right)=\sum\nolimits_j \left|\mathbf{t}_{nij}\right|^2 F_{\mathrm{MV2}}^2\left(\left|\mathbf{t}_{nij}\right|\right),$$
$$\mathbf{X}_3\left(\mathbf{r}_n+\mathbf{R}_i\right)=\sum\nolimits_j \mathbf{t}_{nij}F_{\mathrm{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right),$$
$$X_4\left(\mathbf{r}_n+\mathbf{R}_i\right)=\sum\nolimits_j F_{\mathrm{vdW}}\left(u_{nij}\right),$$
$$F_{\mathrm{vdw}}\left(u_{nij}\right)=\begin{bmatrix}1 & u_{nij}\le 0\\ 1+u_{nij}^3\left[u_{nij}\left(15-6u_{nij}\right)-10\right] & 1>u_{nij}>0\\ 0 & u_{nij}\ge 1\end{bmatrix},\ u_{nij}=\frac{\left|\mathbf{t}_{nij}\right|^2-\left(R_j^{\mathrm{vdW}}+t_-^{\mathrm{vdW}}\right)^2}{\left(R_j^{\mathrm{vdW}}+t_+^{\mathrm{vdW}}\right)^2-\left(R_j^{\mathrm{vdW}}+t_-^{\mathrm{vdW}}\right)^2}$$
$$F_{\mathrm{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right)=\begin{bmatrix}F_{\mathrm{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right)\left(1-F_{\mathrm{vdW}}\left(\left|\mathbf{t}_{nij}\right|\right)\right) & R_j^{\mathrm{vdW}}+t_-^{\mathrm{vdW}}<\left|\mathbf{t}_{nij}\right|<=R_j^{\mathrm{vdW}}+t_+^{\mathrm{vdW}}\\ F_{\mathrm{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right) & R_j^{\mathrm{vdW}}+t_+^{\mathrm{vdW}}<\left|\mathbf{t}_{nij}\right|<=R_j^{\mathrm{vdW}}+t_-^{\mathrm{MV2}}\\ F_{\mathrm{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right)F_{\mathrm{vdw}}\left(\left|\mathbf{t}_{nij}\right|\right) & R_j^{\mathrm{vdW}}+t_-^{\mathrm{MV2}}<\left|\mathbf{t}_{nij}\right|<=R_j^{\mathrm{vdW}}+t_+^{\mathrm{MV2}}\end{bmatrix}, \tag{3.26}$$
$$F_{\mathrm{MV2}}^*\left(\mathbf{t}_{nij}\right)=\frac{\left(C_j\right)^2}{\left(C_j+\left|\mathbf{t}_{nij}\right|^2-\left(R_j^{\mathrm{vdW}}\right)^2\right)^2},C_j=c_1+c_2R_j^{\mathrm{vdW}},\ \text{or}\ F_{\mathrm{MV2}}^*\left(\mathbf{t}_{nij}\right)=\exp\left[\alpha\left(\left|\mathbf{t}_{nij}\right|-R_j^{\mathrm{vdW}}\right)\right].$$

where, the parameters of molecular volume are β = -12, λ = 0.5, S_0 = 0.65, and α = -1.98 1/Å; the parameters of approximated function are c_1 = 0.45 Å^2, and c_2 =1.25 Å; the parameters of atomic volume function are t_-^{vdW} = -0.125 Å, t_+^{vdW} = 0.25 Å, t_-^{MV2} = 1.90 Å, and t_+^{MV2} = 2.10 Å, respectively, R_i^{vdW} are the atomic vdW or input radii. The electrostatic solvation forces in terms of atomic positions are expressed as follows,

$$
\begin{aligned}
F_a^{\text{elec}} &= -\frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_a} = -\left(\sum_{ij} \frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_{ij}} \frac{\partial \mathbf{R}_{ij}}{\partial \mathbf{R}_a} + \sum_i \frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}} \frac{\partial R_i^{\text{GB}}}{\partial \mathbf{R}_a} \right), \\
F_a^{\text{elec,1}} &= -\sum_{ij} \frac{\partial \Delta G^{\text{elec}}}{\partial \mathbf{R}_{ij}} \frac{\partial \mathbf{R}_{ij}}{\partial \mathbf{R}_a} \\
&= \sum_i \left(\tau_{ia} - \frac{\kappa \exp\left(-\kappa f_{ia}^{\text{GB}}\right) f_{ia}^{\text{GB}}}{\varepsilon_{\text{solvent}}} \right) \frac{q_i q_a \left[1 - \exp\left(-\mathbf{R}_{ia}^2 / K_s R_i^{\text{GB}} R_a^{\text{GB}} \right) / K_s \right]}{\left(f_{ia}^{\text{GB}} \right)^3} \left(\mathbf{R}_i - \mathbf{R}_a \right), \\
F_a^{\text{elec,2}} &= -\sum_i \frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}} \frac{\partial R_i^{\text{GB}}}{\partial \mathbf{R}_a} = -\left(F_a^{\text{elec,2a}} + F_a^{\text{elec,2b}} \right), \\
F_a^{\text{elec,2a}} &= \frac{\partial \Delta G^{\text{elec}}}{\partial R_a^{\text{GB}}} \sum_j \sum_n \left(\frac{\partial R^{\text{GB}}}{\partial \mathbf{R}} \right)^{naj}, \quad F_a^{\text{elec,2b}} = \sum_i \sum_n \left(-\frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}} \right) \left(\frac{\partial R^{\text{GB}}}{\partial \mathbf{R}} \right)^{nia},
\end{aligned}
\tag{3.27}
$$

where,

$$\frac{\partial \Delta G^{\text{elec}}}{\partial R_i^{\text{GB}}} = \frac{1}{2}\sum_j \left(\tau_{ij} - \frac{\kappa \exp\left(-\kappa f_{ij}^{\text{GB}}\right) f_{ij}^{\text{GB}}}{\varepsilon_{\text{solvent}}} \right) \frac{q_i q_j \exp\left(-\mathbf{R}_{ij}^2 / K_s R_i^{\text{GB}} R_j^{\text{GB}}\right)}{\left(f_{ij}^{\text{GB}}\right)^3} \left(R_j^{\text{GB}} + \frac{\mathbf{R}_{ij}^2}{K_s R_i^{\text{GB}}} \right),$$

$$\left(\frac{\partial R^{\text{GB}}}{\partial \mathbf{R}}\right)^{nij} = \frac{P_1 \beta \left(a_0 w_n^0 + \dfrac{a_1 w_n^1}{4\left(G_i^1\right)^3} \right)}{\left(a_0 G_i^0 + a_1 G_i^1\right)^2} \frac{\exp\left[\beta\left(S_{ni} - \lambda\right)\right]}{\left(1 + \exp\left[\beta\left(S_{ni} - \lambda\right)\right]\right)^2}$$

$$\left(\frac{\partial X_{1,ni}}{\partial \mathbf{R}_j} \frac{S_0 X_{2,ni}}{\left(\mathbf{X}_{3,ni}\right)^2} + \frac{\partial X_{2,ni}}{\partial \mathbf{R}_j} \frac{S_0 X_{1,ni}}{\left(\mathbf{X}_{3,ni}\right)^2} - \frac{\partial \mathbf{X}_{3,ni}}{\partial \mathbf{R}_j} \frac{\mathbf{X}_{3,ni} 2 S_0 X_{1,ni} X_{2,ni}}{\left(\mathbf{X}_{3,ni}\right)^4} + 2 \frac{\partial X_{4,ni}}{\partial \mathbf{R}_j} \right).$$

$$\frac{\partial X_{1,ni}}{\partial \mathbf{R}_j} = \frac{\partial F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \mathbf{t}_{nij},$$

$$\frac{\partial X_{2,ni}}{\partial \mathbf{R}_j} = 2 F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right) \left(F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right) + \frac{\partial F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \left|\mathbf{t}_{nij}\right|^2 \right) \mathbf{t}_{nij},$$

$$\frac{\partial \mathbf{X}_{3,ni}}{\partial \mathbf{R}_j} = F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right) + \frac{\partial F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \left(\mathbf{t}_{nij}\right)^T \mathbf{t}_{nij},$$

$$\frac{\partial X_{4,ni}}{\partial \mathbf{R}_j} = \frac{\partial F_{\text{vdW}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \mathbf{t}_{nij},$$

$$\frac{\partial F_{\text{vdW}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} = \frac{60 u_{nij}^2 \left(u_{nij}\left(2 - u_{nij}\right) - 1\right)}{\left(R_j^{\text{vdW}} + t_+^{\text{vdw}}\right)^2 - \left(R_j^{\text{vdW}} + t_-^{\text{vdw}}\right)^2},$$

$$\frac{\partial F_{\text{MV2}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} = \begin{bmatrix} \dfrac{\partial F_{\text{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \left(1 - F_{\text{vdw}}\left(\left|\mathbf{t}_{nij}\right|\right)\right) - F_{\text{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right) \dfrac{\partial F_{\text{vdW}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \\ \dfrac{\partial F_{\text{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \\ \dfrac{\partial F_{\text{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} F_{\text{vdw}}\left(\left|\mathbf{t}_{nij}\right|\right) + F_{\text{MV2}}^*\left(\left|\mathbf{t}_{nij}\right|\right) \dfrac{\partial F_{\text{vdW}}\left(\left|\mathbf{t}_{nij}\right|\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} \end{bmatrix}, \quad \textbf{(3.28)}$$

$$\frac{\partial F_{\text{MV2}}^*\left(\mathbf{t}_{nij}\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} = \frac{-4\left(C_j\right)^2}{\left(C_j + \left|\mathbf{t}_{nij}\right|^2 - \left(R_j^{\text{vdW}}\right)^2\right)^3}, \text{ or } \frac{\partial F_{\text{MV2}}^*\left(\mathbf{t}_{nij}\right)}{\partial \left|\mathbf{t}_{nij}\right| \left|\mathbf{t}_{nij}\right|} = \frac{\alpha \exp\left[\alpha\left(\left|\mathbf{t}_{nij}\right| - R_j^{\text{vdW}}\right)\right]}{\left|\mathbf{t}_{nij}\right|}.$$

3.5.2 Nonpolar solvation energy and forces

Based on the expression of nonpolar energy, the forces, the derivatives in terms of the atomic position, are expressed as follows,

$$\begin{aligned}\Delta G_{\mathrm{np}} &= \sum_i \gamma_i A_i \\ &= \sum_i \gamma_i 4\pi \left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right)^2 \sum_m w_m f\left(\overline{V}_i\left(\hat{\mathbf{r}}_m\left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_i\right)\right),\end{aligned} \tag{3.29}$$

$$F_a^{\mathrm{np}} = -\frac{\partial \Delta G_{\mathrm{np}}}{\partial \mathbf{R}_a}, \text{ and}$$

$$\begin{aligned}\frac{\partial \Delta G_{\mathrm{np}}}{\partial \mathbf{R}_a} &= \sum_i \gamma_i 4\pi \left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right)^2 \sum_m w_m f'\left[\overline{V}_i\left(\hat{\mathbf{r}}_m\left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_i\right)\right] \\ &\sum_{j\neq i} 4 f'\left(u_{mij}\right) \frac{\left(\hat{\mathbf{r}}_m\left(R_i^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_i - \mathbf{R}_j\right)}{\left(R_j^{\mathrm{vdW}} + t_+^{\mathrm{SA}}\right)^2 - \left(R_j^{\mathrm{vdW}} + t_-^{\mathrm{SA}}\right)^2}\left(\delta_{ia} - \delta_{ja}\right).\end{aligned} \tag{3.30}$$

where the $f'\left(u\right)$ is the derivative of exposed function, and $\hat{\mathbf{r}}_m$ is the unit vector of grid points.

In order to implement the nonpolar energy and forces in one kernel, whose calculation was divided into two parts and then effectively avoid the conflicts of blocks.

$$\begin{aligned}&\frac{\partial \Delta G_{\mathrm{np}}}{\partial \mathbf{R}_a} = \sum_m \gamma_a 4\pi \left(R_a^{\mathrm{vdW}} + R_{\mathrm{w}}\right)^2 w_m f'\left[\overline{V}_a\left(\hat{\mathbf{r}}_m\left(R_a^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_a\right)\right] \\ &\left[\sum_{j\neq a} 4 f'\left(u_{maj}\right) \frac{\left(\hat{\mathbf{r}}_m\left(R_a^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_a - \mathbf{R}_j\right)}{\left(R_j^{\mathrm{vdW}} + t_+^{\mathrm{SA}}\right)^2 - \left(R_j^{\mathrm{vdW}} + t_-^{\mathrm{SA}}\right)^2}\right] + \\ &\sum_m \frac{4w_m}{\left(R_a^{\mathrm{vdW}} + t_+^{\mathrm{SA}}\right)^2 - \left(R_a^{\mathrm{vdW}} + t_-^{\mathrm{SA}}\right)^2}\left[\sum_{j\neq a} f'\left[\overline{V}_j\left(\hat{\mathbf{r}}_m\left(R_j^{\mathrm{vdW}} + R_{\mathrm{w}}\right) + \mathbf{R}_j\right)\right]\right. \\ &\left.\gamma_j 4\pi \left(R_j^{\mathrm{vdW}} + R_{\mathrm{w}}\right)^2 f'\left(u_{mja}\right)\left(\mathbf{R}_a - \hat{\mathbf{r}}_m\left(R_j^{\mathrm{vdW}} + R_{\mathrm{w}}\right) - \mathbf{R}_j\right)\right].\end{aligned} \tag{3.31}$$

3.5.3 CUDA algorithms for computing the electrostatic solvation energy and forces

Two important steps are used to calculate the electrostatic solvation energies. The first step is to calculate the Born radius of each atom. Besides looping over all atoms, it is necessary to loop each numerical integration grid and then all neighbor atoms at each grid point (as given by the lookup table). The major cost is to compute the molecular volume at each grid point (Equation 3.25), which has four intermediate volumes (Equation 3.26) that can be attributed to the neighbor atoms. The pseudocode is given in Table 3.2. After obtaining the Born radii, the existing kernel was used for computing the electrostatic solvation energies (Equation 3.23).

Computing the electrostatic solvation forces are much more complicated than that of energies. The forces in terms of the coordinates can be calculated using a similar algorithm as implemented in the GPU-GBSW plugin. For the forces in terms of the Born radii, the computation of atomic forces is divided into two parts, in order to avoid the conflict of blocks. The algorithm is summarized in the following pseudo code.

The computational bottleneck is to calculate the second part, because frequent access of the global arrays (S, X_1, X_2, X_3, and X_4) is expensive. Additionally, these global arrays plus the lookup table array takes up most of the global memory, which should be optimized by minimizing the effective size.

Table 3.2 The pseudocode of calculating the Born radius of each atom.

Each block loops the atoms (i)
    Assign the shared memory (size = # the numerical grids)
  Loop the numerical grids (n) using 256 threads (optimal)
    Initialize the V, S, X_1, X_2, X_3, and X_4
    Using the lookup table array to locate the neighbor atoms ($\mathbf{r}_n + \mathbf{R}_i \Rightarrow \mathbf{R}_j$)
    Loop the neighbor atoms
      Calculate the F_{vdW}, F_{MV2}, and X_1, X_2, X_3, and X_4 (Equation 3.26)
      Calculate the S and V (Equation 3.25), and save them into shared memory
      Save S, X_1, X_2, X_3, and X_4 into global memory for the calculations of forces
    Loop the numerical grids (n) using 1 thread
      Do the sum reduction (Equation 3.24) by extracting data from the shared memory
    Save the sum into the Born radius of each atom
END

Table 3.3 The pseudocode of calculating electrostatic solvation forces of each atom.

```
Each block loops the atoms (i)
    Assign the shared memory (size = # the numerical grids)
    # First part of Equation 3.27: F^elec, 2a
  Loop the numerical grids (n) using 256 threads (optimal)
    Using the lookup table array to locate the neighbor atoms (r_n + R_i => R_j)
     Access the global arrays (S, X1, X2, X3, and X4)
     Loop the neighbor atoms
          Calculate the derivatives of F_vdW, F_MV2, and X1, X2, X3, and X4 (Equation 3.28)
               Calculate derivatives of Born radii (Equation 3.28)
          Do the sum reduction and save into the shared memory
    Loop the numerical grids (n) using 1 thread
          Do the sum reduction (Equation 3.24) by extracting data from the shared memory
    Save the sum into the atomic forces array
    # Second part of Equation 3.27: F^elec, 2b
  Loop the numerical grids (n) using 256 threads (optimal)
    Using the lookup table array to locate the neighbor atoms (r_n + R_i => R_j)
     Loop the neighbor atoms
        Access the global arrays (S, X1, X2, X3, and X4)
          Calculate the derivatives of F_vdW, F_MV2, and X1, X2, X3, and X4 (Equation 3.28)
               Calculate derivatives of Born radii (Equation 3.28) using S, X1, X2, X3, and X4
          Do the sum reduction and save into the shared memory
    Loop the numerical grids (n) using 1 thread
          Do the sum reduction (Equation 3.24) by extracting data from the shared memory
    Save the sum into the atomic forces array
END
```

3.5.4 Structure analysis of key GPU-GBMV2/SA kernels

In the most time-consuming reduceGBMVForce kernel, each thread uses 54 registers (Figure 3.9), and each block uses 54 x 256 = 13,824 registers. Since each streaming multiprocessor (SM) provides 65,536 register on Titan X (Pascal), only 4 blocks (equivalently, 32 warps or 1024 threads) could run simultaneously on each SM. The analysis of the computeNonbonded kernel of OpenMM is also provided for reference. Even though the computeNonbonded kernel has a higher theoretical occupancy of 62.5%, the actual achieved occupancies are similar between these two kernels.

Variable	Achieved	Theoretical	Device Limit
Occupancy Per SM			
Active Blocks		4	32
Active Warps	29.14	32	64
Active Threads		1024	2048
Occupancy	45.5%	50%	100%
Warps			
Threads/Block		256	1024
Warps/Block		8	32
Block Limit		8	32
Registers			
Registers/Thread		54	65536
Registers/Block		14336	65536
Block Limit		4	32
Shared Memory			
Shared Memory/Block		3072	98304
Block Limit		32	32

Variable	Achieved	Theoretical	Device Limit
Occupancy Per SM			
Active Blocks		5	32
Active Warps	31.69	40	64
Active Threads		1280	2048
Occupancy	49.5%	62.5%	100%
Warps			
Threads/Block		256	1024
Warps/Block		8	32
Block Limit		8	32
Registers			
Registers/Thread		42	65536
Registers/Block		12288	65536
Block Limit		5	32
Shared Memory			
Shared Memory/Block		2048	98304
Block Limit		48	32

Figure 3.9 GPU utilization using the nvvp and nvprof tools for the reduceGBMVForce kernel in GBMV2/SA (left) and the computeNonbonded kernel of OpenMM (right). The profile results were obtained using protein 3GB1.

3.5.5 Multi-Core Performance of CPU-GBMV2/SA

Table 3.4 Benchmarks of GBMV2/SA for GPU vs. parallel CPU calculations with 1, 2, 4, 8, 12 and 16 cores. The time step was set to 2-fs. The GPU and CPU calculations were done on one NVIDIA TITAN X (Pascal) and the Intel Xeon E5-2620 v4 2.10GHz CPU, respectively.

PDBID (#Atoms)	3GB1 (855)	P53-TAD (926)	1BVC (2459)	4AT5 (11766)
CPU-GBMV2/SA (ns/day)	0.7969 (1x)	0.8392 (1x)	0.2534 (1x)	0.0488 (1x)
Fast CPU-GBMV2/SA (1-core)	1.2614 (1.6x)	1.3168 (1.6x)	0.3826 (1.5x)	0.0728 (1.5x)
2-core / 1-core	2.0x	2.0x	2.0x	2.0x
4-core / 1-core	4.0x	4.0x	3.9x	3.5x
8-core / 1-core	5.1x	5.1x	5.2x	4.6x
12-core / 1-core	5.6x	5.6x	5.7x	4.8x
16-core / 1-core	6.4x	6.4x	6.7x	5.5x
GPU-GBMV2/SA (1-GPU)	46.9974 (59.0x)	48.9630 (58.3x)	15.9292 (62.9x)	3.5294 (72.3x)

CHAPTER 4

ASSESSING GBMV2/SA IMPLICIT SOLVENT FORCE FIELD FOR SIMULATING INTRINSICALLY DISORDERED PROTEINS USING THE MULTISCALE ENHANCED SAMPLING

4.1 Introduction

Compared to well-structured proteins, intrinsically disordered proteins (IDPs) or regions (IDRs) lack stable tertiary structures under physiological conditions but exhibit sophisticated signaling and functions in multicellular organisms [228]. They constitute approximately one-third of eukaryotic proteins and are associated with around 25% of missense mutation-related diseases, including cancer [10, 12]. The inherent thermodynamic instability of IDPs allows for greater conformational flexibility, enabling them to respond sensitively to various stimuli such as binding, changes in cellular environments (*e.g.*, pH), and post-translational modifications [5, 229]. Experimental studies have shown that IDPs can adopt relatively stable structures and carry out their biological functions when bound to their partners [230, 231]. However, understanding the recognition of targets by IDPs and the mechanisms underlying their folding and binding processes is challenging due to their intrinsic flexibility.

Characterizing representative states and transitions of IDPs at the atomistic level is essential for understanding their kinetics of binding and folding. Experimental studies often lack the efficiency to provide comprehensive descriptions, whereas MD simulations offer the ability to calculate thermodynamic and kinetic properties for both flexible and well-structured proteins [7]. However, MD simulations for IDPs require highly accurate

force fields and sufficient sampling of relevant conformational ensembles, which pose additional challenges [42, 232]. Therefore, it is crucial to explore alternative approaches that strike a balance between accuracy and efficiency.

Explicit solvent protein force fields often lead to overly compact ensembles for IDPs due to limitations in describing protein-protein, protein-water, and water-water interactions, as well as computational constraints for larger water system sizes [50, 183, 233]. An alternative approach with promise is the use of implicit solvent models, which reduce the system size by approximately 10-fold by directly estimating solvation free energy [199, 200]. Implicit solvent models have shown success in simulating well-structured proteins and capturing structural features of certain IDPs [42, 234]. Among the popular implicit solvent models, the improved Generalized Born with molecular volume and solvent accessible surface area (GBMV2/SA) model can accurately reproduce the structures and stabilities of helical peptides and β-hairpins, while mitigating the over compaction bias seen in other implicit solvent force fields. This makes it particularly suitable for investigating the mechanisms of IDP interactions [124].

One main limitation of the GBMV2/SA implicit solvent model is its demanding computational cost and poor multi-core scaling due to the complex calculation of solvation free energy. However, there have been significant advancements in GPU-accelerated algorithms for protein force fields, enabling hundreds of speedups compared to conventional CPU-based algorithms and standard atomistic MD simulations at the microsecond level [217, 222, 235]. Similarly, a GPU-accelerated version of the GBMV2/SA implicit solvent model has been developed as an OpenMM plugin. Remarkably, this GPU-accelerated implementation achieves a ~60x speedup while

maintaining numerical equivalence to the original CPU-GBMV2/SA calculations. This enhancement in speed and efficiency greatly expands the applicability of the implicit solvent simulations using GBMV2/SA to larger systems and longer time scales.

It remains challenging to generate representative conformational ensembles through standard GPU-GBMV2/SA implicit solvent simulations due to diverse ensembles and high energy barriers [25, 236]. To address this, enhanced sampling methods have been developed, including the topology-based coarse-grained model for accelerating atomistic conformational sampling [34]. The multiscale enhanced sampling (MSES) technique utilizes this coarse-grained model to drive conformational transitions in atomistic simulations, aided by temperature/Hamiltonian replica exchange to remove bias potential effects, achieving faster transitions and maintaining accuracy of atomistic force fields. While other techniques like temperature replica exchange, umbrella sampling, and metadynamics have been proposed for accelerated conformational sampling [237], the MSES model stands out by enabling faster sampling of atomistic models through the use of a coarse-grained model for driving transitions [34, 122]. Previous studies demonstrated significant improvements in convergence for small but non-trivial IDPs using implicit solvent models (*e.g.*, GBSW and GBMV2) [122, 124]. Unfortunately, the CPU-only implementation of MSES limits its application in IDP conformational sampling and prevents harnessing the GPU capabilities of CHARMM/OpenMM. Therefore, the development of a GPU-accelerated MSES technique is crucial, as it would enable faster conformational sampling of IDPs.

In this chapter, the implementation of the MSES algorithms for sampling conformational ensembles and transitions of both folded and unfolded proteins will be

introduced. This approach will then be applied to assess the reliability of the GBMV2/SA implicit solvent model for conformational sampling of IDPs. It is anticipated that GPU-MSES will prove to be more efficient in sampling IDP conformations. The successful development of GPU-MSES will also greatly enhance the applicability of other implicit solvent models in IDP conformational sampling. The following sections will focus on introducing the methodology of the MSES model and its GPU implementation algorithms. The correctness of the GPU-GBMV2/SA model in simulating IDPs will also be tested and its reliability assessed.

4.2 Method

4.2.1 Multiscale enhanced sampling

The MSES method uses the CG model to accelerate the conformational sampling of AT model by introducing a coupling AT-CG energy term.

$$U_{\text{mix}}\left(\mathbf{r}_{\text{AT}},\mathbf{r}_{\text{CG}},\lambda\right)=U_{\text{AT}}\left(\mathbf{r}_{\text{AT}}\right)+U_{\text{CG}}\left(\mathbf{r}_{\text{CG}}\right)+\lambda U_{\text{AT-CG}}\left(\mathbf{r}_{\text{AT}},\mathbf{r}_{\text{CG}}\right), \quad \textbf{(4.1)}$$

where the U_{AT}, U_{CG} and $U_{\text{AT-CG}}$ are the atomistic, coarse-grained and coupled AT-CG potential energy functions, respectively. Given a proper coupling potential and a coupling scaling factor (such as $\lambda = 1$), it has been shown that the AT system can be effectively driven by the CG system, so a coupled potential plays an important role in accelerating the conformational transition of IDPs [34]. Motivated by the notion that native contacts dictate protein folding transitions, the CG and atomistic copies are coupled by using a harmonic type of penalty function, which depends on the differences of Cα–Cα distances between the AT and CG native contacts (Δd),

$$U_{\text{AT-CG}}\left(\mathbf{r}_{\text{AT}},\mathbf{r}_{\text{CG}}\right)=\sum_{m=1}^{M}k_m\phi_m\left(\Delta d\right),$$
$$\phi_m\left(\Delta d\right)=\begin{cases}\dfrac{1}{2}\left(\Delta d\right)^2,\ \Delta d\le d_s^m\\ A_m+\dfrac{B_m}{\left(\Delta d\right)^{s_m}}+f_{\max}^m\Delta d,\ \Delta d>d_s^m\end{cases}, \tag{4.2}$$

where the parameters A_{m} and B_{m} are identified by requiring both energy and its first derivative to be continuous at a switching distance (d_{s}^{m}) point,

$$A_m=\left(d_s^m\right)^2\left(1/2+1/s_m\right)-f_{\max}^m d_s^m\left(1+2/s_m\right),$$
$$B_m=\left(d_s^m\right)^{s_m+1}\left(f_{\max}^m-d_s^m\right)/s_m. \tag{4.3}$$

4.2.2 CPU/CUDA implementation of MSES method as an OpenMM plugin

The atomistic and coarse-grained potentials use the same energy function, which is determined by the protein force fields used in the simulations. Different from a regular harmonic potential, the MSES coupling potential uses the Cα–Cα distance of native contact as a basic variable to quantify the similarity of AT and CG models, which assumes that native contacts can be considered as a reaction coordinate to describe a protein folding process. It can be therefore considered as a four-body bonded interaction, so it could be better if we can implement them as an OpenMM plugin, rather than using a custom force. At present, both CPU and CUDA platforms have been implemented in the MSES plugin. It can be expected that the calculation of MSES coupling potential should be much faster than the calculations of AT and CG potentials, because the number of native contacts is much less than the number of atoms in the protein system. The architecture of MSES plugin has three important layers. The first layer is a python layer that is public to users. This layer is usually created by a SWIG interface automatically,

once we provide an interface input file which includes all function declarations (such as "msesplugin.i"). The second layer is the openmmapi. In this layer, it will define both force and forceimp classes, for example, "MSESForce" and "MSESImpl" in this MSES plugin. These classes will provide the function/data not only for the Python API, but also for the OpenMM platforms (such as CPU and CUDA). The input data will therefore be saved in this layer, including the input data from the users after using a Python API. The last layer is the platforms, which will implement the underlying algorithms of different platforms and compute all necessary operations.

4.2.3 Model systems and benchmark simulations

To evaluate the computational accuracy of GBMV2/SA implicit solvent model in the conformational sampling of proteins, we select the small proteins, including Ace-$(AAQAA)_3$-NH_2 and three β-hairpins, GB1p (residues: GEWTYD DATK TFTVTE), GB1m1 (residues: GEWTYD DATK TATVTE) and GB1m3 (residues: KKWTYN PATG KFTVQE), and also relatively large proteins, including the 28-residue segment of kinase-inducible domain (KID) of transcription factor CREB (residues 119−146: TD SQKRR EILSR RPSYR KILND LSSDA P), and p53-TAD domain (residues 1−61: MEEPQ SDPSV EPPLS QETFS DLWKL LPENN VLSPLPSQAM DDLML SPDDI EQWFT EDPGP D). The standard MD simulations at the temperature (270 K for GB1p and 300 K for others) are employed to run the folding simulations, where the initial conformation is an extend conformation melted from a MD simulation at a higher temperature (400 K).

4.2.4 Computational details and trajectory analysis

We perform the Langevin dynamics with a friction coefficient of 0.1 ps^{-1} with a time step of 2 fs, where all bonds involving hydrogen atoms are constrained using a SHAKE algorithm. The default optimized GBMV2/SA force field and parameters are used in all simulations, *e.g.*, the surface tension coefficient (γ = 5 cal/mol/Å^2), the angular numerical points (NPHI = 38). The multiscale enhanced sampling (MSES) simulations are carried out in CHARMM with a modified MMTSB toolset [238]. The default parameters are used, including force constant is 1.0 kcal/mol/Å^2, F_{max} is 0.5, R_{SWI} is 2 Å. The CHARMM36m protein force field is used to describe the protein-protein interactions. To monitor the convergence, the initial conformations include the mixed folded and unfolded structures, which are generated from a high-temperature simulation. All the results are analyzed using an in-house Python package. The error bar in each plot is shown in a standard error.

4.3 Results and discussion

4.3.1 CPU/CUDA implementations of MSES model

The MSES model involves three important energy terms, including the first two AT and CG energy terms that have been implemented previously, and the coupling energy term that is now implemented as an OpenMM plugin merged in the CHARMM program. Besides, two different platforms (such as CPU and CUDA) are implemented for a comparison, due to their simplicity. To ensure the correctness of MSES implementation, we compare their MSES energies and forces (Figure 4.1). It is clearly seen that their energies and forces are closer, and their difference is within a neglectable error, which suggests that the MSES model has been correctly implemented in the CHARMM program as an OpenMM plugin.

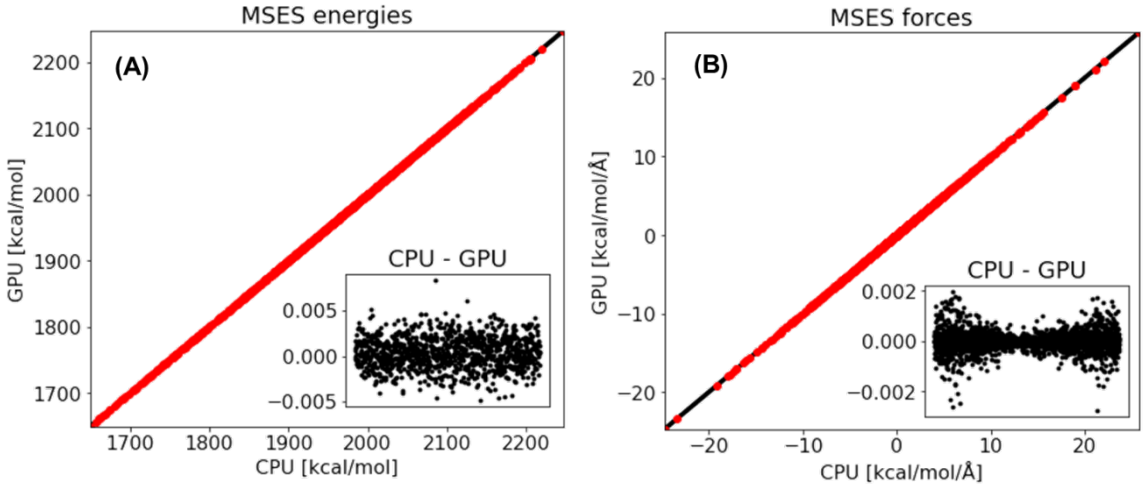

Figure 4.1 The comparison of CPU- and GPU-MSES model in calculating energies (A) and forces (B). The inserted images show the difference of CPU and GPU calculations in the same unit. The black line (y = x) is used as reference. The p53-TAD is used as a test system, and a variety of trajectories including both folded and unfolded structures are used for the energy calculations, but one folded structure is selected to calculate the molecular forces.

In addition to the conformational energies and forces, we test the energy conservation of MSES model (Figure 4.2). Conserved energy is often used to monitor the correctness of MSES model as well, because it can indicate that the conformational forces are derived from the energies. Unsurprisingly, total conformational energies are conserved well, and the observed energy fluctuation is much less. No obvious energy drift is observed within 300 ps, which also suggests a reliable implementation of MSES model as an OpenMM plugin.

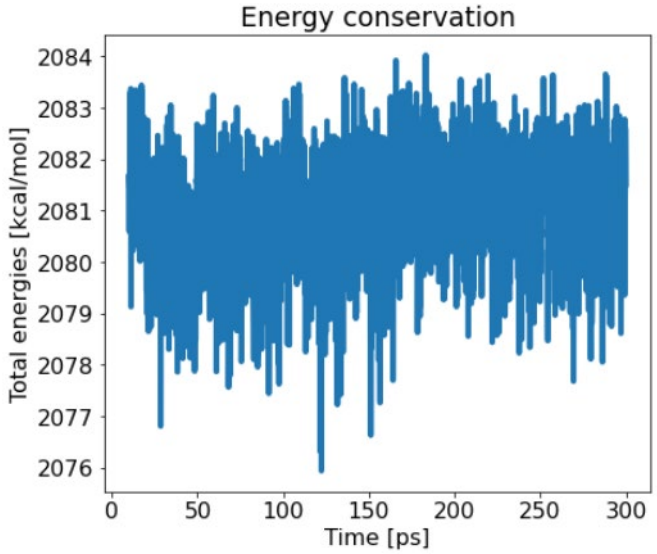

Figure 4.2 The energy conservation of MSES model. The p53-TAD protein is used as a test system. Mixed precision is used in the CUDA calculation.

4.3.2 Conformational equilibrium of protein simulations

Based on the correct implementation of MSES model, we then use it to accelerate the conformational sampling and to test whether this GBMV2/SA model can provide a reliable description in the conformational equilibrium of protein simulations. To remove the bias of MSES coupling potential, a series of Hamiltonian potentials are designed to achieve a replica exchange MD simulation, which can be found in previous simulations

as well [124]. To quantify the reliability of GBMV2/SA model, we select the secondary structure of several systems (such as Ace-$(AAQAA)_3$-NH_2, β-hairpins, KID and p53-TAD) as an important indicator, because of the experimental data provided.

4.3.2.1 Small peptide simulations

We first calculate the secondary structure of both helical and β-sheet systems (such as Ace-$(AAQAA)_3$-NH_2 and β-hairpins), because these two systems have been considered as a reference to quantify the capability of computational model (such as GBMV2/SA). We clearly see that the GBMV2/SA model reasonably describes the helicity of Ace-$(AAQAA)_3$-NH_2 system, and the folding stability of three β-hairpins, which are all consistent with the experimental observations (Figure 4.3). It should be noted that our results also agree with previous CPU-GBMV2/SA calculations. However, these peptide simulations are insufficient to quantify the reliability of GBMV2/SA model, because the GBMV2/SA model was highly tuned by reproducing the secondary structure of these two types of peptides.

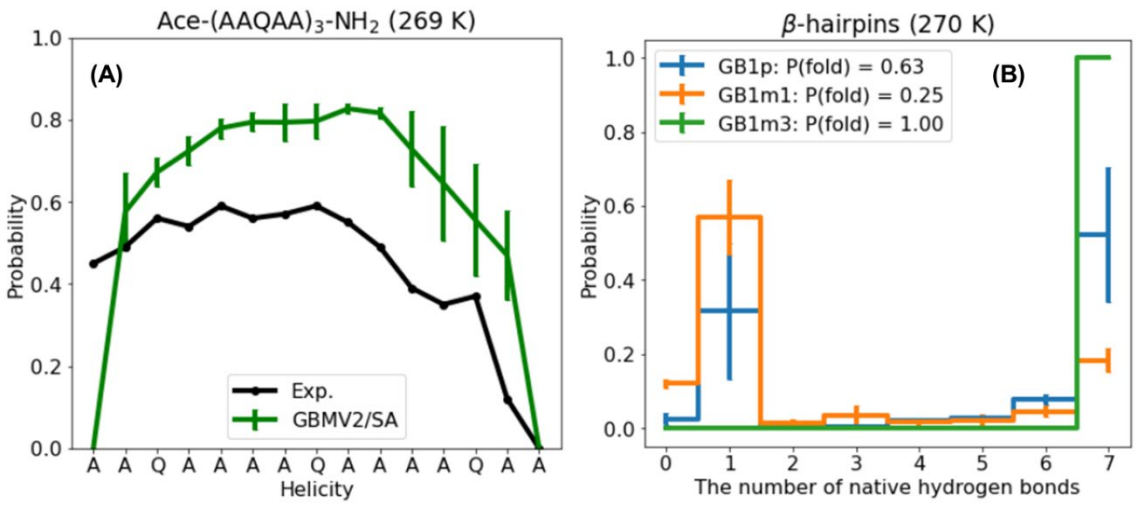

Figure 4.3 The population of (A) Ace-(AAQAA)$_3$-NH$_2$ helicity and (B) the number of hydrogen bonds of three β-hairpins, including GB1p, GB1m1, and GB1m3. Their stability is ordered as GB1m1 < GB1p < GB1m3.

4.3.2.2 KID and p53-TAD simulations

Larger proteins systems are therefore used to assess the GBMV2/SA model. Figure 4.4 shows the helicity profiles of KID and p53-TAD proteins. The KID protein includes two helical structures, and the experimental data shows that the first αA structure (residue ID: 120-129, 50-60%) is more helical than the second αB structure (residue ID: 134-144, 10%). Unfortunately, the results of GBMV2/SA model provide an opposite description, where the average helicity of αA structure is ~10%, while ~50% population is observed in the αB structure. This also agrees with the previous observation that αA structure is less stable than αB structure in the CPU-GBMV2/SA model [124]. For a larger p53-TAD protein system, we also observe that the GBMV2/SA model provides a higher population of its helicity of p53-TAD (residue ID: 17-29, ~ 40%), compared to the experimental data (~10%). Combined together, it suggests that the GBMV2/SA model provides a limited accuracy of secondary structures for larger helical

protein systems. This inconsistency could be attributed to the inaccurate description of nonpolar term, where a simple SASA model hardly captures some long-range dispersion solute-solvent interactions, which will be discussed in the following chapter.

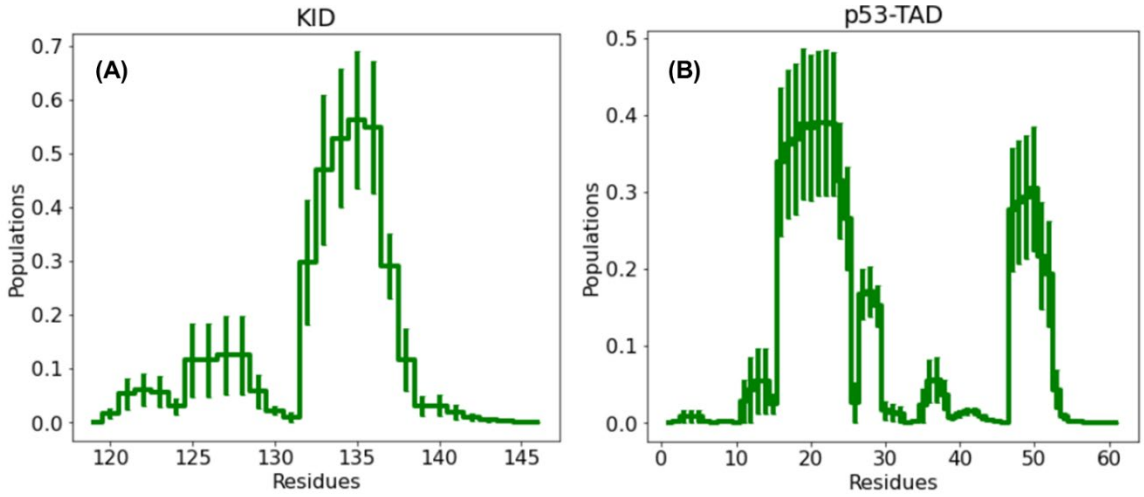

Figure 4.4 The helicity of KID (288 K) and p53-TAD (300 K) protein. Both systems are used to monitor the reliability of GBMV2/SA model in describing the conformational sampling of IDPs.

4.4 Conclusions

We have correctly implemented the MSES model as an OpenMM plugin, which is also validated by comparing molecular energies or forces of both CPU and CUDA calculations and confirming the energy conservation of p53-TAD protein simulation. We then use it to assess the capacity of GBMV2/SA model in sampling of IDP conformations, and potential limitations for a set of folded and unfolded proteins, focusing on the comparisons with experimental observations in protein secondary structures. Results show the GBMV2/SA model can provide consistent results with the experimental data for the small peptides but are limited to provide an accurate description of secondary structure of large helical proteins, such as KID and p53-TAD helical

proteins. This could be due to a missing nonpolar dispersion, which will be developed and discussed in the following chapter.

CHAPTER 5

IMPROVED IMPLICIT TREATMENT OF NONPOLAR SOLVATION FREE ENERGIES: THE GBMV2/NP MODEL

5.1 Introduction

Explicit solvent models are widely used to investigate the impact of solvent on solute molecules, as they provide a detailed and accurate description of solute-solvent interactions [197]. However, their computational cost significantly increases with system size, and sampling solute conformations becomes more challenging due to solvent friction. Implicit solvent models offer an alternative by reducing system size and enabling faster sampling of solute conformations. These models accurately capture solvation free energies that govern solvent effects on solute conformations [200, 239]. The use of implicit solvent models provides faster energy calculations and conformational sampling, with up to a 60-fold speedup compared to explicit solvent models in folding simulations of small systems [240].

Several implicit solvent force fields have been developed to accurately capture solute-solvent interactions, with the solvation free energy being a crucial physical quantity for describing the solute's free energy landscape [241]. However, accurately describing the solvation free energy is challenging. A thermodynamic cycle can be employed to account for different contributions, including the cavity free energy, nonpolar free energy, and electrostatic free energy, which capture the effects of solvent entropy, solute-solvent nonpolar interactions, and solute-solvent electrostatic interactions, respectively [242]. The Poisson-Boltzmann (PB) and generalized Born (GB) models are

commonly used to estimate the electrostatic solvation free energy through continuum electrostatics treatments of the solvent environment [243]. The GB model, in particular, allows for analytical evaluation of molecular forces and is widely employed in molecular dynamics (MD) simulations [239]. However, the GB model requires an accurate description of the effective Born radius of each atom, which depends on the solute's coordinates and molecular volume [225]. Several GB models have been developed to improve the accuracy of these descriptions, such as the GB with simple smoothing function (GBSW) and GB with molecular volume (GBMV) [145-147]. Notably, the GBMV2 model incorporates a vector-based scaling term to approximate the molecular volume, resulting in a reduction of unphysical regions within the solute's high dielectric interior [146, 147]. The GBMV2 model has shown better agreement with explicit solvent and experimental data, including the free energy profiles of amino acid side chain pairs, hydration free energies of small molecules, and accurate descriptions of folded and unfolded small protein conformations [124, 207, 212, 213]..

In addition to the electrostatic contribution, the nonpolar contribution is also crucial for describing solute conformations [149]. The conventional GBMV2/SA model utilizes solvent accessible surface area (SASA) to account for the nonpolar solvation free energy, employing a water probe radius of 1.4 Å [147]. While successful in describing folded and small disordered proteins, this model faces limitations in balancing the conformations of large proteins like KID and p53-TAD proteins (refer to section 4.3.2.2). One possible reason is the insufficient description of nonpolar solvation free energy, as the SASA model primarily captures short-range interactions and overlooks long-range solute-solvent nonpolar interactions [149]. Previous studies have pointed out

inconsistencies in the SASA-based model when describing nonpolar solvation free energies of different systems, such as cyclic, linear, and branched alkanes, and its limitations when calculating binding free energies of protein-protein or protein-ligand complexes [242, 244]. Furthermore, solute-solvent dispersion interactions are highly dependent on the conformation, with folded and unfolded conformations exhibiting different surface coefficients when fitted with SASA, indicating the inadequacy of the simple SASA model in capturing this conformational dependence. Therefore, improved implicit solvent models have been developed to directly include solute-solvent dispersion interactions and address this conformation-dependent effect [149]. For instance, solute-solvent van der Waals (vdW) and hydrogen bonding interaction terms, absent in the GB model, have been incorporated to describe the nonpolar component, leading to improved agreement with reference explicit solvent models in capturing conformational ensembles [143, 144]. Hence, it is necessary to develop an enhanced GBMV2/NP model capable of capturing this conformational dependence.

Graphic Processing Units (GPUs) have emerged as a powerful tool for accelerating the computational efficiency of MD programs [245]. Many MD engines, including CHARMM, AMBER, GROMACS, NAMD, and OpenMM, have incorporated GPU-based algorithms for protein force field calculations [191, 192, 246]. Consequently, significant efforts have been directed towards GPU acceleration of implicit solvent models, particularly in reducing the computational cost of calculating GB terms [220, 222]. For instance, the GB/OBC model has enabled routine microsecond MD simulations [222], and the GBSW model exhibited a remarkable 100x improvement over conventional CPU calculations for protein simulations [223]. Additionally, the nonpolar

solvation energy term has also been accelerated, as seen with the implementation of the SASA model into the GPU-accelerated GBSA/OBC model [224]. Recently, GPU-accelerated algorithms for GBMV2/SA were integrated into an OpenMM plugin, offering a substantial speedup of approximately 60x compared to the conventional CPU-based algorithms used in the CHARMM program [156].

In this chapter, we present a novel GBMV2/NP model and its GPU implementations as an OpenMM plugin. Building upon our previous GPU-accelerated GBMV2/SA model, we have devised a nonpolar model that combines SASA calculations with solute-solvent dispersion interactions. For consistency and compatibility, all implementations maintain the same architecture as the earlier OpenMM plugin, now referred to as the GBMV2/NP plugin, and have been integrated into both CHARMM and OpenMM programs. To comprehensively describe the GBMV2/NP model, this chapter offers an in-depth account of its methodology, implementation algorithms, and comprehensive testing and benchmarking analyses.

5.2 Method

In GBMV2/NP model, the total solvation free energy is generally divided into electrostatic and nonpolar contributions,

$$\Delta G_{\text{solv}} = \Delta G_{\text{elec}} + \Delta G_{\text{np}}, \tag{5.1}$$

where the nonpolar component involves the free energy cost of creating the solute cavity in the solvent and turning on the nonpolar solute-solvent vdW interaction, and the electrostatic component corresponds to the free energy cost of the subsequent step of charging up the solute [242]. The nonpolar contribution can be estimated directly from

$\Delta G_{\mathrm{np}}(\mathbf{X}) = \Delta G_{\mathrm{rep}}(\mathbf{X}) + \Delta G_{\mathrm{dis}}(\mathbf{X})$, these solvation free energy terms can be thus evaluated by the following expressions,

$$
\begin{aligned}
e^{-\beta\Delta G_{\mathrm{rep}}(\mathbf{X})} &= \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U^{\mathrm{rep}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}, \\
e^{-\beta\Delta G_{\mathrm{dis}}(\mathbf{X})} &= \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U^{\mathrm{rep}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U^{\mathrm{dis}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U^{\mathrm{rep}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}, \\
e^{-\beta\Delta G_{\mathrm{elec}}(\mathbf{X})} &= \frac{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}{\int d\mathbf{Y} e^{-\beta\left[U_{\mathrm{p\text{-}p}}(\mathbf{X})+U^{\mathrm{rep}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U^{\mathrm{dis}}_{\mathrm{p\text{-}w}}(\mathbf{X},\mathbf{Y})+U_{\mathrm{w\text{-}w}}(\mathbf{Y})\right]}}.
\end{aligned}
\tag{5.2}
$$

Similarly, the above three expressions can be described by the following thermodynamic cycle (Figure 5.1).

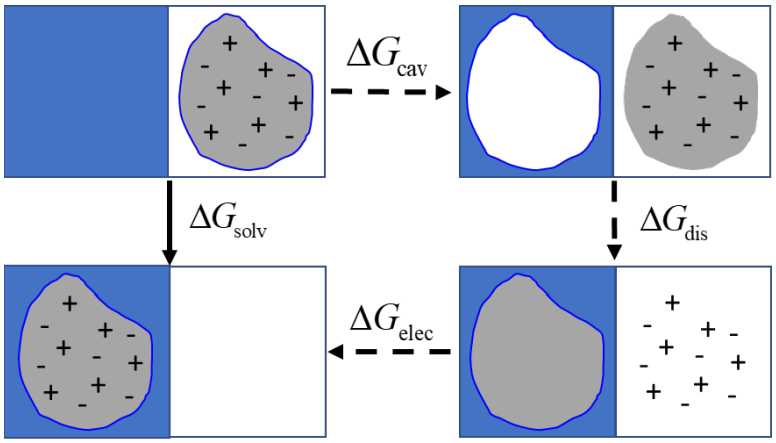

Figure 5.1 A thermodynamic cycle for calculating the solvation free energy into repulsive, dispersion, and electrostatic components.

5.2.1 Generalized Born electrostatic solvation free energy

A detailed description of GB electrostatic solvation free energy has been included in the section 3.5.1. In brief, the electrostatic term is approximated by the GB model

developed by Still and coworkers [225], which can be expressed as follows in a low concentration of salt,

$$\begin{aligned}
\Delta G^{\text{elec}} &= -\frac{1}{2}\sum_{i,j}\tau_{ij}\frac{q_i q_j}{f_{ij}^{\text{GB}}}, \\
\tau_{ij} &= \left(1/\varepsilon_{\text{solute}} - \exp\left(-\kappa f_{ij}^{\text{GB}}\right)/\varepsilon_{\text{solvent}}\right), \\
f_{ij}^{\text{GB}} &= \sqrt{\mathbf{R}_{ij}^2 + R_i^{\text{GB}} R_j^{\text{GB}} \exp\left(-\mathbf{R}_{ij}^2 / K_s R_i^{\text{GB}} R_j^{\text{GB}}\right)},
\end{aligned} \tag{5.3}$$

where q_i and R_i^{GB} the partial charge and Born radius of i^{th} atom, $\mathbf{R}_{ij}$ is a distance vector between two atoms, The K_s is usually set to 8 for GBMV2 electrostatic calculations, $\varepsilon_{\text{solute}}$ and $\varepsilon_{\text{solvent}}$ are the dielectric constant of solute and solvent, respectively, and κ is a Debye-Hückel screening parameter. It can be observed that the GB electrostatic model requires the calculations of Born radii that are related to the molecular volume by considering the approximated Coulomb and even high-order correction terms [147],

$$\begin{aligned}
R_i^{GB} &= \frac{P_1}{a_0 G_i^0 + a_1 G_i^1} + P_2, \\
G_i^0 &= \frac{1}{4\pi}\int_{\text{Solvent}} \frac{1}{\left|\mathbf{r}-\mathbf{R}_i\right|^4} d\mathbf{r},\ G_i^1 = \left(\frac{1}{4\pi}\int_{\text{Solvent}} \frac{1}{\left|\mathbf{r}-\mathbf{R}_i\right|^7} d\mathbf{r}\right)^{1/4},
\end{aligned} \tag{5.4}$$

where the parameters of Born radii are P_1 = 0.9085, P_2 = -0.102 Å, $a_0 = 1 - 1/\sqrt{2}$, and a_1 = 1. We usually used a standard numerical quadrature technique to calculate the integrals of the Coulomb field and high-order correction terms [147].

5.2.2 Nonpolar repulsive solvation free energy and forces

In GBMV2/NP model, the nonpolar model can be divided into both the cavity free energies to create a cavity to fit in the solute molecule and the nonpolar free energies

to turn on the long-range nonpolar dispersion interactions between the solute and solvent [143]. The nonpolar repulsive term is usually described by a SASA model,

$$\Delta G_{\text{rep}}(\mathbf{R}) \approx \sum_i \gamma_i A_i(\mathbf{R}, R_w), \tag{5.5}$$

where A_i and γ_i is the atomistic surface area and surface tension coefficient of each atom, respectively. The atomic SA can be expressed as,

$$A_i = \int_{|\mathbf{r}-\mathbf{R}_i|=R_i^{vdW}+R_w} f\left(\overline{V}_i(\mathbf{r})\right) d\mathbf{r}, \tag{5.6}$$

where the excluded volume, $\overline{V}_i(\mathbf{r}) = \sum_{j\neq i} V_j(\mathbf{r})$, involves the solvent accessible atomic volume except for i^{th} atom, and the smooth function f represents the exposed rate at **r** point, which should be one if the excluded volume is zero and be zero if the excluded volume is one. A straightforward numerical expression is then given after using a numerical integration,

$$A_i \approx 4\pi\left(R_i^{\text{vdW}} + R_{\text{w}}\right)^2 \sum_m w_m f\left(\overline{V}_i\left(\hat{\mathbf{r}}_m\left(R_i^{\text{vdW}} + R_{\text{w}}\right) + \mathbf{R}_i\right)\right), \tag{5.7}$$

nonpolar repulsive energy can be further expressed as follows,

$$\Delta G_{\text{rep}}(\mathbf{R}) = \sum_i 4\pi\gamma_i\left(R_i^{\text{vdW}} + R_{\text{w}}\right)^2 \sum_m w_m f\left(\overline{V}_i\left(\hat{\mathbf{r}}_m\left(R_i^{\text{vdW}} + R_{\text{w}}\right) + \mathbf{R}_i\right)\right). \tag{5.8}$$

The derivative of nonpolar repulsive energy in terms of each atomic position is,

$$\mathbf{F}_a^{\text{rep}}(\mathbf{R}) = -\frac{\partial \Delta G_{\text{rep}}}{\partial R_a} = -\sum_i 4\pi\gamma_i\left(R_i^{\text{vdW}} + R_{\text{w}}\right)^2 \sum_m w_m \frac{\partial f}{\partial \overline{V}_{m,j}} \frac{\partial \overline{V}_{m,j}}{\partial \mathbf{R}_a}. \tag{5.9}$$

5.2.3 Nonpolar attractive solvation free energy and forces

It has been proposed that the solute-solvent dispersion interactions can be described by a continuum vdW solvent model, which assumes that the average water

number density is constant outside of solute volume. Then, the solute-solvent dispersion interactions can be evaluated by a summation of all atomic solute-solvent dispersion interactions [143, 149],

$$\begin{aligned} \Delta G_{\text{dis}}\left(\mathbf{X}\right) &\approx \sum_i \alpha_i U_{\text{dis}}\left(\mathbf{x}_i\right) + \beta, \\ U_{\text{dis}}\left(\mathbf{x}_i\right) &= -\rho_w \int_{\text{solvent}} u_{\text{dis}}^{(i)}\left(\left|\mathbf{x}-\mathbf{x}_i\right|\right) d\mathbf{x}. \end{aligned} \tag{5.10}$$

where the $U_{\text{dis}}\left(\mathbf{x}_i\right)$ is the dispersion interaction energy of atom *i* with the solvent. It has been reported that the atomic dispersion energy can be accurately estimated by integrating over the solvent region [242]. However, an empirical expression was previously used to describe the solute-solvent dispersion free energy, $U_{\text{dis}}\left(\mathbf{x}_i\right) \approx 16\pi\alpha_i\rho_w\varepsilon_{iw}\sigma_{iw}^6 \Big/ 3\left(R_i^{GB}+R_w\right)^3$, so that we have,

$$\Delta G_{\text{dis}}\left(\mathbf{x}\right) \approx \sum\nolimits_i \frac{16\pi\alpha_i\rho_w\varepsilon_{iw}\sigma_{iw}^6}{3\left(R_i^{GB}+R_w\right)^3}, \tag{5.11}$$

where α_{i} is an adjustable parameter (default value is 1.0), ρ_{w} is the number of solvent molecules at standard conditions, which is 0.33428 Å^{-3} for water, σ_{iw} and ε_{iw} are the Berthelot rule between Lennard-Jones interaction parameters of solute atom I and oxygen atom of the water model, respectively [143, 144]. An advantage of using this Gallicchio-Levy (G-L) approximation is that this atomic dispersion term is associated with the atomic Born radius, R_i^{GB}. As a consequence, it is faster than calculate the forces of G-L dispersion energy as below, because this dispersion energy is directly related to the Born radius of each atom, so that they can share some data structure with the GB electrostatic calculations. Similarly, the nonpolar attractive forces can be written as follows,

$$F_a^{\text{dis}} = -\frac{\delta \Delta G_{\text{dis}}(\mathbf{x})}{\delta \mathbf{x}_a} = -\sum_i \frac{16\pi\alpha_i \rho_w \varepsilon_{iw} \sigma_{iw}^6}{\left(R_i^{GB} + R_w\right)^4} \frac{\delta R_i^{GB}}{\delta \mathbf{x}_a}. \quad \textbf{(5.12)}$$

Thus, the nonpolar attractive dispersion energy and forces are only dependent of the Born radii and their derivatives with respect to the coordinates. The Lennard-Jones parameters of atom *i* are $\sigma_{iw} = (\sigma_i + \sigma_w)/2$, and $\varepsilon_{iw} = \sqrt{\varepsilon_i \varepsilon_w}$, respectively, where the σ_w = 3.15365 Å and ε_w = 0.155 kcal/mol. R_w was set to 1.4 Å.

5.2.4 CUDA implementation as an OpenMM plugin

The existing GBMV2/SA kernels can be adapted for the CUDA implementation of GBMV2/NP model. First, it requires the calculations of Born radii and their derivatives that have been implemented in the GBMV2/SA kernels (such as "calcBornR" and "reduceGBMVForce"). Second, the energy and forces of SASA term can be computed in the "calcSASA" kernel. We therefore implement the calculations of nonpolar dispersion terms in the GB electrostatic kernels, instead of implementing a new kernel for the calculation of dispersion term, which avoids some unnecessary calculations. To provide an interface to the CHARMM and OpenMM programs, we also implement the C++/C, Fortran and Python API for the users according to the instruction of OpenMM development. At present, a standalone version of GBMV2/NP plugin has been implemented for OpenMM users. Besides, we provide another interface for the CHARMM users by merging it into the current CHARMM code. However, the CPU-GBMV2/NP calculations are not available in the GBMV2/NP plugin, but we implement the CPU-GBMV2/NP energy calculations in the CHARMM program, which is used to validate the correction of CUDA-GBMV2/NP implementation.

5.2.5 Computational details

Similar to the GPU-GBMV2/SA model, we validated the correctness and accuracy of nonpolar terms in the GPU-GBMV2/NP model by using the nonpolar energies of amino acid side chain pairs from the CPU-GBMV2/NP calculations. Besides, the increased computational cost and the percentage of time spent in different kernels are benchmarked in GPU-GBMV2/NP calculations. The energy conservation calculation is also used to evaluate the numerical stability of GBMV2/NP model, and we also test the effect of the number of numerical grids on its energy stability. To verify the reliability of G-L approximation, we follow the similar protocol published by previous study, to calculate the solute-solvent interaction energy by an explicit solvent MD simulations [242]. In brief, the CHARMM36m force field was used, and the solute molecules are kept rigid, and their atomic charges are set to zero to exclude the effect of electrostatic interactions. The periodic boundary condition is used, and a large water box is employed to ensure the solute cannot see its images along any direction. The system is first equilibrated for 100 ps using 1-fs as a time step, and 10 ns MD production simulation with a 2-fs time step is carried out to collect the trajectories. The attractive component of Lennard-Jones solute-solvent interactions is calculated by the Weeks-Chandler-Andersen (WCA) decomposition scheme [247]. For all calculations, the cutoff distance for the nonbonded calculations is set to 20 Å. All GPU calculations are carried out on an NVIDA TITAN X (Pascal) graphics card, and the CPU calculations use an Intel Xeon E5-2620 v4 2.10GHz CPU. The nvprof tool is used to report the performance of all GPU kernels.

It is critical but challenging to parameterize the GBMV2/NP model, because it includes many unknown parameters, including the global parameters (such as γ, R_w, and

α) and local parameters (such as the input radius of each atom). By following the philosophy of GBMV2/SA parameterization, we first obtain the solvation free energies of all side chains and the free energy profiles of side chain pairs by running the CHARMM36m explicit solvent simulations as reference, then use them iteratively parameterize the GBMV2/NP model. In addition, the vdW input radius can be regarded as a starting choice, so we use them to fit these global parameters first, and then tune the input radius to find out a good agreement with the reference data iteratively. We also use multiscale enhanced sampling simulations to run several small protein simulations (such as Ace-$(AAQAA)_3$-NH_2 and β hairpins) to ensure whether the GBMV2/NP can capture the equilibrium of both folded and unfolded conformational ensembles. Several control simulations of folded proteins (such as 1BDC protein) are also employed to test whether it can be used to simulate the folded proteins as well.

5.3 Results and discussion

5.3.1 Energy and forces of nonpolar solvation free energy

We first verify the correctness of GBMV2/NP GPU implementation by comparing the conformational energies and forces of 1BDC proteins (Figure 5.2). Expectedly, the energies and forces calculated from both CPU and GPU implementation are closer and their difference can be neglectable in terms of the total energies or forces. However, the CPU energies are always a little larger than the GPU energies, which could be attributed to the mixed precision we used in the GPU calculations. Differently, the average difference of both CPU and GPU forces are closer to zero, although we still can

observe several large deviations (~0.1 kcal/mol Å). Besides, the results can be reproduced in both CHARMM and OpenMM programs, so it suggests that the GPU-GBMV2/NP has been implemented correctly and is also available for users upon request.

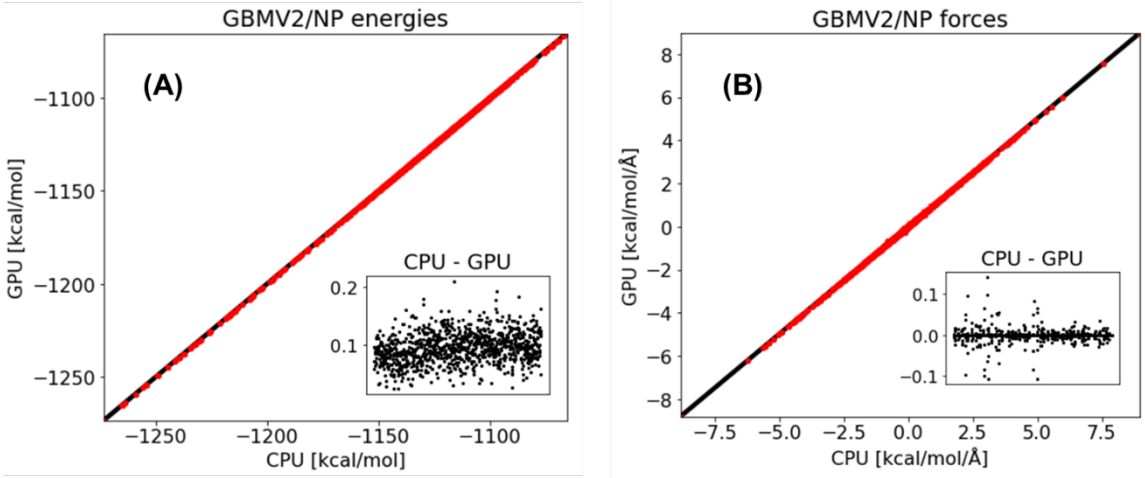

Figure 5.2 Comparison of GPU- and CPU-GBMV2/NP in calculating the energies (A) and forces (B) of 1BDC protein, where a variety of conformations are used to calculate the energies, while the forces are calculated from one structure. The diagonal black line (y = x) is shown as reference. The inserted panels are the difference between CPU and GPU calculations in the same unit. The CPU forces are calculated by the "test first" command from the CHARMM program.

5.3.2 Effect of NPHI on the numerical stability

It is noted that the GBMV2/NP model uses a numerical method to calculate the molecular volume and surface of solute. The number of numerical grid points will therefore affect the stability of calculating molecular energy and forces. We investigate the effect of angular grid points (NPHI) on the energy conservation of GBMV2/NP calculations (Figure 5.3). It is clearly seen that the energy conservation of GBMV2/NP model is highly associated with the NPHI values, where a large NPHI can give a small energy drift. However, a larger NPHI value usually results in a slower performance, due to an increase of total numerical grid points. It should be mentioned that the GBMV2/NP

model uses a larger surface tension coefficient (γ = 62 cal/mol Å^2), and thus is much larger than that of GBMV2/SA model (γ = 5 cal/mol Å^2), which requires a larger NPHI value to keep the energies less drifted. To keep a balance of computational cost and accuracy, it is found that a good choice of NPHI value for GB electrostatic and nonpolar calculations is 50 and 110, respectively.

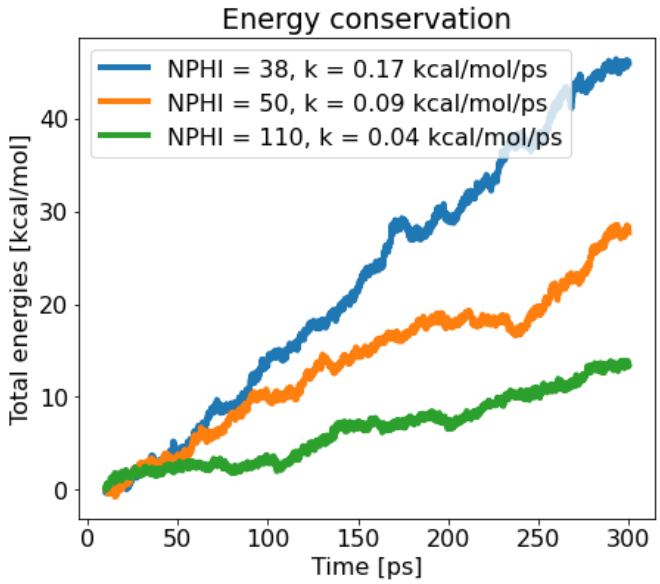

Figure 5.3 The effect of NPHI value on the energy conservation for GBMV2/NP MD simulations of GB1p peptide. The NPHI value is the number of angular numerical grid points used for the GBMV2/NP model. Both GB electrostatic and nonpolar terms use the same NPHI value. The default number of radial grid points is also used in the calculations of GB electrostatic energy.

5.3.3 Computational efficiency

We further explore the computational efficiency of GPU-GBMV2/NP model. Previous observations of GBMV2/SA model showed a ~60x faster speedup [156]. It is expected that the addition of this dispersion term will not significantly increase the computational cost of GBMV2/NP model, because it directly shares the data from the GB electrostatic calculations, including the calculation of Born radius of each atom and its

derivative with respect to atomic coordinates. This is confirmed by the benchmarking test of 3GB1 protein (Figure 5.4), which shows a little increase in the GBMV2/NP calculation. Unsurprisingly, the timings of the important GPU kernels are similar to our previous observations (Figure 3.8). This indicates that the inclusion of dispersion term has little effect on the computational efficiency of GBMV2/NP model.

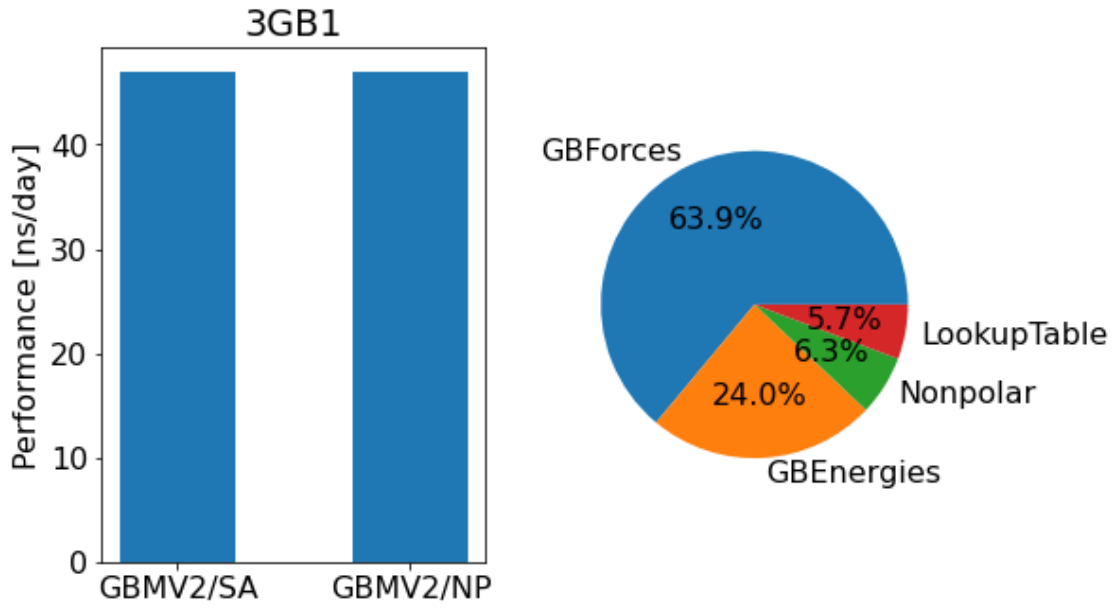

Figure 5.4 Comparison of GPU-GBMV2/SA and GBMV2/NP models in simulating a moderate size of 3GB1 protein, and the percentage of time spent in several important GPU kernels is shown in a pie graph.

5.3.4 Parameterization and benchmarking of GBMV2/NP model

5.3.4.1 Verify G-L approximation

The G-L approximation of vdW dispersion solute-solvent interaction energy has been used in previous implicit solvent models [143, 144]. It is still necessary to test whether it can provide a reliable approximation in the GBMV2/NP model, given that they use different approximations of molecular volume and Born radius. It can be seen from

Figure 5.5 that this G-L approximation can provide a good correlation between the implicit and explicit atomic solute-solvent dispersion interactions, which has an agreement with the previous observations [143]. This suggests that we can set the default value of its coefficient as one, which will be used in all following calculations.

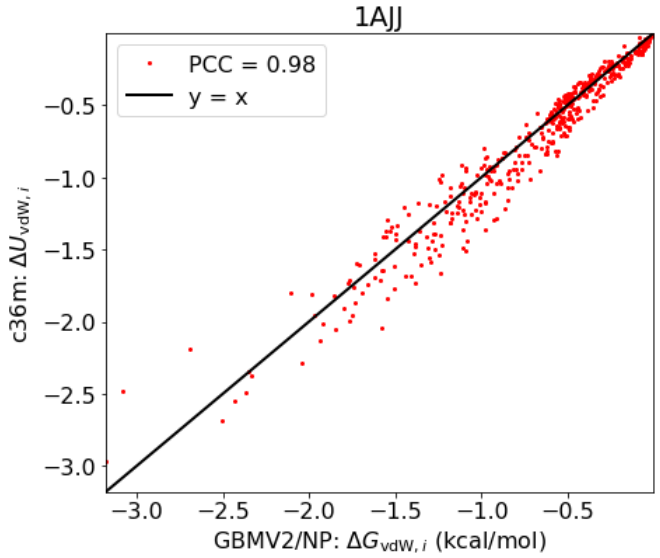

Figure 5.5 Comparisons of the atomistic solute-solvent vdW dispersion interactions between the GBMV2/NP and CHARMM36m (c36m) explicit solvent simulations. The PCC is the Pearson correlation coefficient, and it is better when it is closer to 1. The protein (PDBID: 1AJJ) was used in this calculation.

5.3.4.2 Parameterization of GBMV2/NP model

To parameterize the GBMV2/NP model, we use both solvation free energies of side chains and stabilities of many representative pairs to fit the key parameters, including the radius of water to probe the solvent accessible surface area (R_w), surface tension coefficient (γ), and the atomic input radius (R_i^{vdW}), the coefficient of vdW dispersion term (α_i), and possible backbone torsion profile (CMAP) to reproduce the secondary structures of both helical and β-sheet peptides. The reference data of solvation free energies of nonpolar side chains can be found from previous study [248]. We also

run both c36m and a99SBdisp explicit solvent simulations for all amino acids side chain pairs. In addition, we iteratively tune the hydrogen bond strength and CMAP profile to ensure a reasonable description of protein secondary structure for several peptides.

First, the R_w value is optimized to 0.8 Å, which reproduces the peak of potential mean forces (PMF) profile of nonpolar side chain pairs (such as wy_pd), which represents the first shell waters on the surface of solute (Figure 5.6). As we see, the GBMV2/NP model gives a very close match to the c36m explicit solvent simulation, compared to the GBMV2/SA model, in particular for the first peak of PMF profile. This can also be observed in other nonpolar side chain pairs.

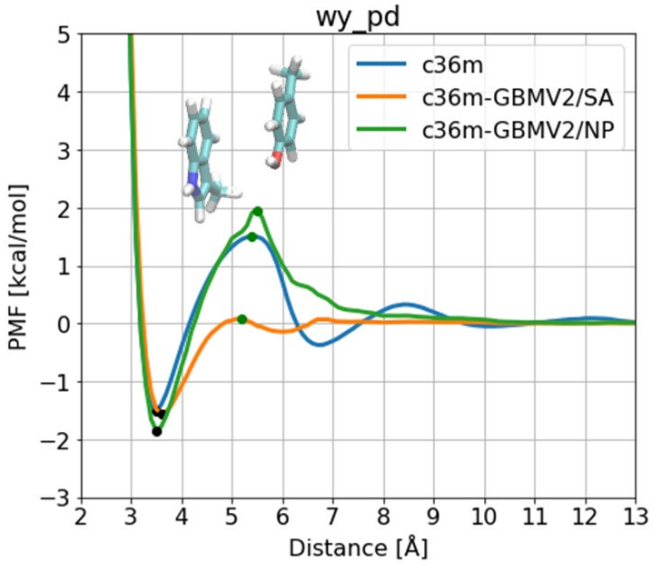

Figure 5.6 The free energy profile of Trp and Tyr side chain pair (wy_pd) for three protein force fields. The distance of CE2 and CE1 atom type is used as an order parameter to obtain the free energy profile. The inserted image shows the structure of wy_pd pair.

However, it is found that it is challenging to balance the solvation free energies of side chains and stabilities of side chain pairs when we fit the surface tension coefficient (γ). In the end, we find that the choice (γ = 0.062 kcal/mol Å^2) reaches a compromise and provides a good agreement with the explicit solvent data of both solvation free energies

(Figure 5.7) and stabilities (Figure 5.8), although it is relatively limited for the GBMV2/NP model to describe the stabilities of several nonpolar side chain pairs. This inconsistency could be attributed to the use of vdW volume in calculating the cavity free energy term, which could provide an inconsistent description of bound and unbound state of nonpolar side chain pairs.

Besides these nonpolar side chain pairs, we tune the polar or charged pairs to achieve a closer stability to the explicit ones by changing their input radii. We mainly focus on the backbone hydrogen bonding pairs (such as ala2 pair in Figure 5.8), which determines the stability of backbone hydrogen bond. It is noted that some charged pairs (such as rrsa, ek, and eks) cannot be described in the GBMV2 models (Figure 5.8**)**. However, both the state-of-the-art c36m and a99SBdisp protein force fields are also limited in describing these charged pairs, so this deviation could highly depend on the protein force fields.

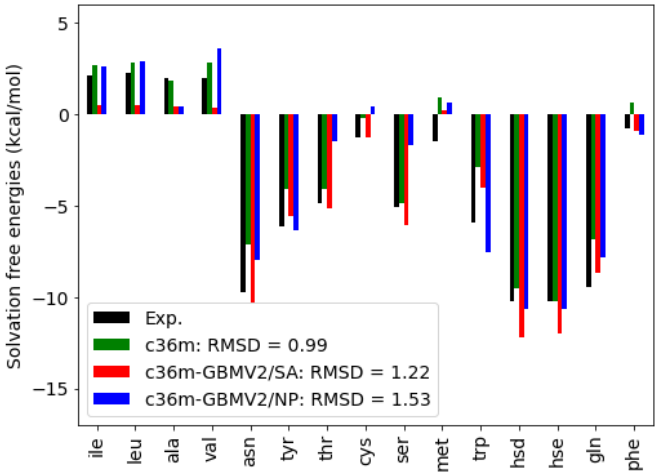

Figure 5.7 The solvation free energies of all nonpolar amino acids side chains. The data of experiment and c36m explicit simulations is obtained from the previous result [248]. The RMSD values of all calculations are calculated in terms of the experimental data, and a larger value means that it is less close to the experimental values.

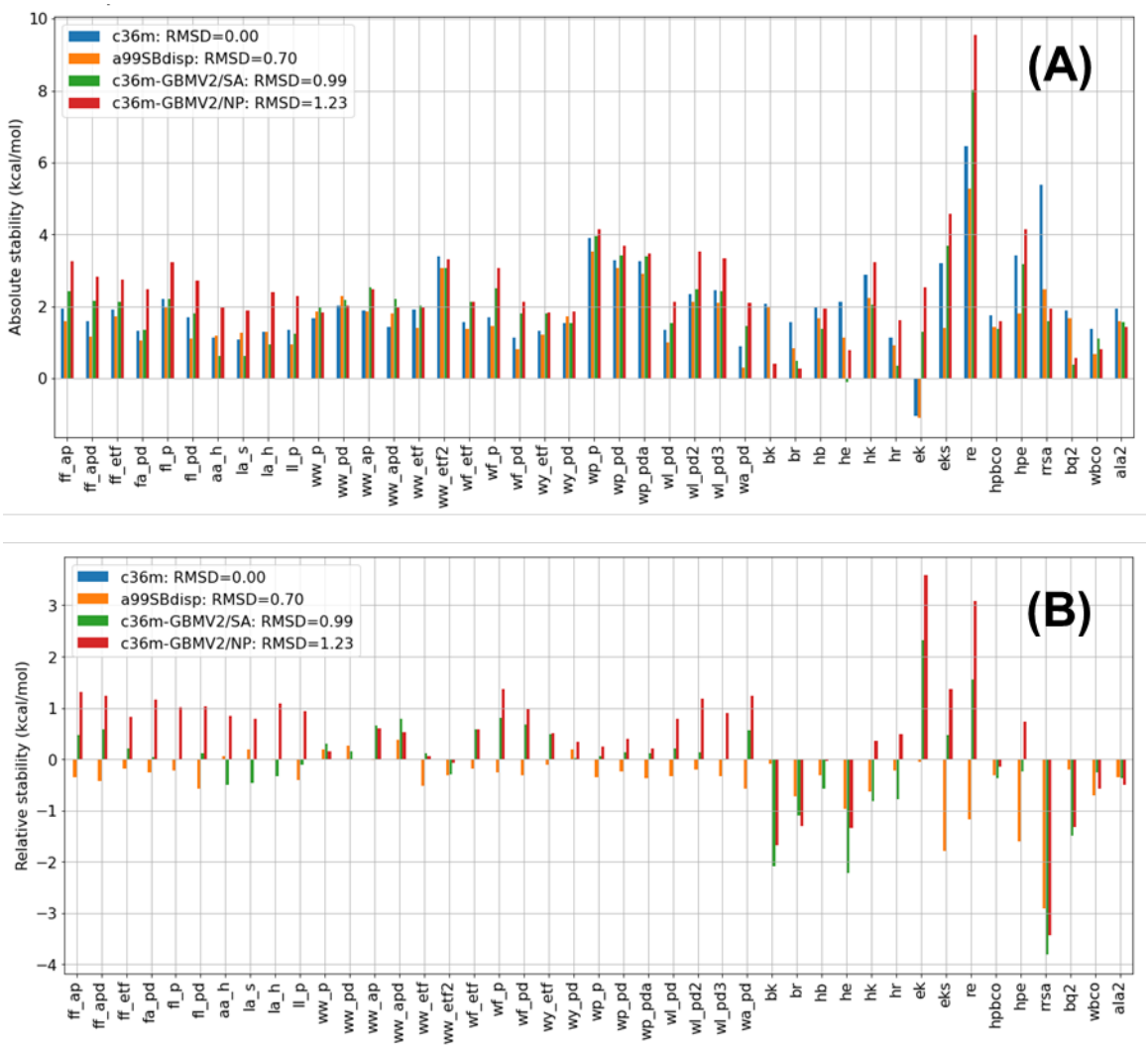

Figure 5.8 The stabilities of amino acids side chain pairs for the selected amino acid side chain pairs, and their descriptions can be found in previous paper [124]. The data from both c36m and a99SBdisp explicit solvent simulations are considered as reference. RMSD values of all calculations are calculated in terms of the c36m explicit solvent simulation.

5.3.4.3 Conformational equilibrium of peptide simulations

We further test the optimized GBMV2/NP to reproduce the conformational equilibrium of several peptide simulations (Figure 5.9); It shows that it can give a reasonable description of small peptide simulations. For example, the helicity profile of Ace-(AAQAA)$_3$-NH$_2$ peptide calculated from c36m-GBMV2/NP model is slightly higher

than the experimental observation. However, it describes the folding stability order of both GB1p and GB1m1 β-sheet peptide correctly, which are closer to the experimental data (~42% for GB1p and ~6% for GB1m1 at 278 K) [249]. This is partly due to an iterative optimization of input radius of each atom, in particular to reproduce the stability of backbone hydrogen bonding pairs.

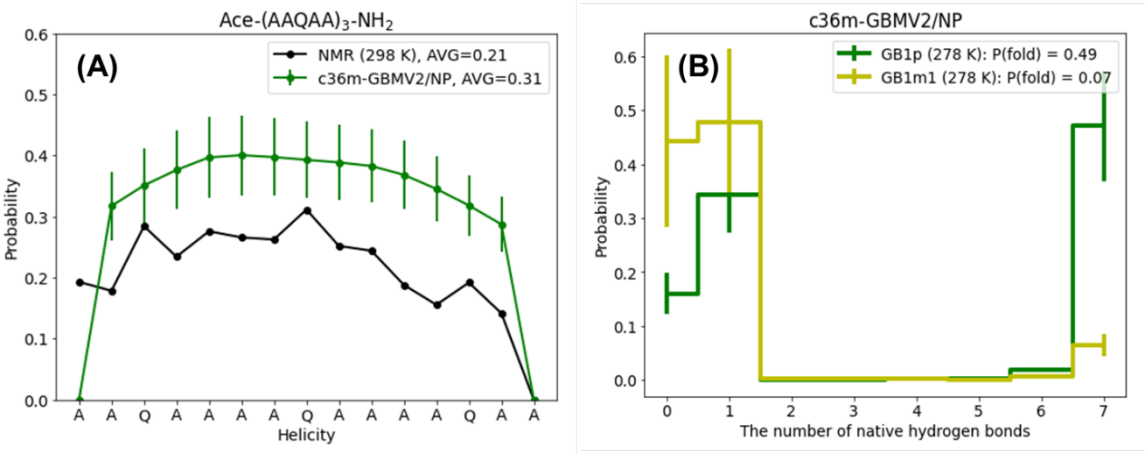

Figure 5.9 (A) The helicity profile of Ace-(AAQAA)$_3$-NH$_2$ peptide and (B) the population of the number of native hydrogen bonds. GB1p: GEWTYD DATK TFTVTE; GB1m1: GEWTYD DATK TATVTE; Experimental observation: GB1p is ~42% folded at 278 K and the stability of GB1p is higher than that of GB1m1 peptide.

5.3.4.4 Control simulations of folded proteins

Unfortunately, it seems that the GBMV2/NP model does not favor the stabilization of well-folded proteins (such as 1BDC protein in Figure 5.10). The native structure of 1BDC protein can be readily unfolded within a few nanoseconds, which is inconsistent with the experimental observation that this protein is pretty stable at 300 K. To understand the underlying reasoning, we plot its SASA value with the increase of R_w value. We can observe that both folded and unfolded structures give different behaviors. For example, they both have a minimum point, which is around a small R_w value. This

suggests that the SASA value includes the unphysical surface area inside the 1BDC protein when the radius of water probe is very small (Figure 5.10). This effect can be significantly observed in the folded structures, compared to the unfolded structures. It is because most protein side chains are exposed to the solvent for the unfolded proteins. This unphysical surface area resulted from the use of a lower R_w value (0.8 Å in the GBMV2/NP model) destabilize the folded structures. Although the GBMV2/SA uses the same SASA model, it has a very small surface tension coefficient (γ = 0.005 kcal/mol Å^2), compared to that of GBMV2/NP model (γ = 0.062 kcal/mol Å^2) and also uses a larger radius of water probe (R_w = 1.4 Å). As a result, this effect is not highlighted in previous GBMV2/SA model. This observation shows that the GBMV2/NP model needs a better model to calculate accurately the SASA of proteins, which needs to be addressed in the future.

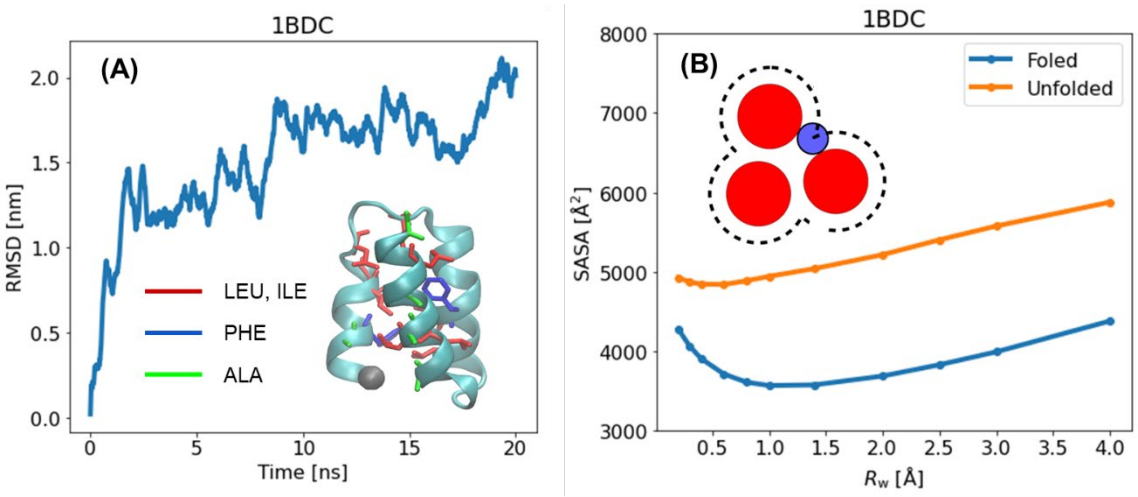

Figure 5.10 (A) The RMSD value of 1BDC protein during a control simulation at 300 K. The starting native structure is inserted, and the hydrophobic residues are shown in colors. (B) The SASA values are calculated from the SASA model with an increase of R_w value. Both folded and unfolded structures are used for a comparison. The inserted image shows a process to calculate the SASA value.

5.4 Conclusions

A dispersion-corrected term by a G-L approximation has been correctly implemented in a GPU-accelerated GBMV2/NP implicit solvent model, to improve the description of the nonpolar solvation free energies. It shows that its inclusion will not significantly increase the computational cost, compared to a more expensive calculation of electrostatic solvation free energy. Preliminary results also verify the reliability of G-L approximation to reproduce the explicit solute-solvent dispersion energies. It also shows that the optimized GBMV2/NP model gives a good agreement with the explicit solvent calculations, including not only the solvation free energies of amino acids side chains and stabilities for most of side chain pairs, but also the equilibrium of both folded and unfolded conformational ensembles for small peptides. However, this model favors the unfolded states for well-folded proteins, in particular large proteins (such as 3GB1 and 1BDC), which is inconsistent with the experimental observations. The underlying reason is attributed to the inaccurate description of repulsive SASA model that has a large surface tension coefficient, compared to that of GBMV2/SA model. As a result, the folded state of protein has a larger surface area that results from the description of unphysical vdW volume, which increases the instability of folded states of protein. This suggests that a sufficient description of the SASA term still needs to be improved.

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

Accurate and reliable molecular simulations are crucial for studying the detailed conformational ensembles of intrinsically disordered proteins (IDPs) in isolation, dynamic complexes, or biological condensates. By integrating computational capabilities with experimental studies, we can gain insights into how dynamic protein states respond to cellular stimuli, such as signaling and regulation, and establish a more rigorous understanding of the structure-function relationship of IDPs and intrinsically disordered regions (IDRs). In this dissertation, we investigated explicit and implicit atomistic simulations of IDP folding and dynamics. Our assessment of state-of-the-art protein force fields, including CHARMM36m and a99SB-disp, revealed that the CHARMM36m force field overestimated the stability of polar/charged pairs, particularly in electrostatic interactions with water molecules involving Arg and Glu residues. This indicates the need for a more accurate force field to describe protein-water electrostatic interactions and better describe the conformational ensembles and dynamics of IDPs.

Implicit solvent models have shown great potential in accelerating the conformational sampling of IDPs, but they may provide less accurate descriptions of IDP conformations. The GBMV2 models, including GBMV2/SA and GBMV2/NP, utilize an improved molecular volume to describe the electrostatic solvation free energy, striking a balance between accuracy and efficiency in IDP simulations. To further enhance the speed of GBMV2 models, we have implemented GPU acceleration for the calculation of

electrostatic and nonpolar terms, resulting in a significant ~60x speedup in GBMV2 simulations. However, the current GBMV2 models still struggle to accurately capture both folded and unfolded conformations. For instance, GBMV2/SA fails to capture the secondary structure of the p53-TAD protein compared to experimental observations. Additionally, the GBMV2/NP model, which includes an improved nonpolar description, faces challenges in effectively balancing the cavity and solute-solvent dispersion solvation free energies. Further optimization is necessary to fully realize the potential application of the GBMV2/NP model in sampling IDP conformations.

6.2 Future directions

The high dimensionality and complex nature of disordered protein conformation continues to push the limits of the force field and sampling capability. In particular, none of these methods alone appears to be generally applicable to simulate IDPs that are large (*e.g.*, more than a few dozens of residues) and/or contain nontrivial residual structural features. We still need more studies to provide a reliable and feasible computational method to simulate both folded and disordered proteins, including the development of protein force fields and enhanced sampling methods.

The optimization of current protein force fields to accurately describe protein-water interactions, especially for polar/charged residue pairs, is an ongoing endeavor. However, achieving a balanced representation of protein-protein and protein-water interactions for these residues remains challenging. A promising direction for improvement lies in the optimization of polarizable protein force fields specifically tailored for IDP systems, as classical protein force fields may not adequately capture their

unique conformations. Polarizable force fields have the potential to describe polar/charged residues in buried or water-exposed environments by incorporating polarizable effects. In addition to polarizability, a more accurate force field is needed to describe protein-water electrostatic interactions. Machine learning (ML) based protein force fields offer a promising avenue in this regard. While numerous studies are focused on developing ML-based force fields for protein systems, these potential ML-based force fields can strike a balance between ordered and disordered proteins. They rely on more accurate, albeit computationally expensive, models such as quantum chemistry or density functional theory calculations, which serve as a valuable foundation for the development of transferable protein force fields.

The development of more effective methods for sampling IDP conformations and dynamic interactions is an urgent need and presents exciting opportunities. One promising approach is the integration of various existing strategies, both CV-dependent and CV-free, to enhance sampling. A particularly exciting direction is the application of machine learning to design adaptive sampling strategies that can dynamically generate bias potentials to explore the free energy landscape more efficiently. Additionally, several protein models with different levels of resolution are being developed and refined for IDP simulations, especially for studying biological condensates. These models range from simplified Cα-only single-bead protein models to more complex implicit solvent models with atomistic representations. Many of the current models are designed to capture systems with minimal residual structures. A key challenge in multi-scale modeling and simulation of IDPs is finding the optimal trade-off between resolution, accuracy, and efficiency for the specific problem at hand. Nevertheless, multi-scale

simulations are expected to continue playing a central role in the study of IDPs and their dynamic interactions.

Indeed, the accurate prediction of binding free energies of protein-ligand systems remains a challenge for implicit solvent models, including GBMV2 models such as GBMV2/SA and GBMV2/NP. The reliability of these models in estimating the binding free energies of diverse drug-like molecules is still unknown. While many implicit solvent models struggle to provide accurate descriptions within a few kcal/mol, it is unclear whether GBMV2 models, with their improved molecular volume and accurate representation of electrostatic solvation free energy, can overcome this limitation. Therefore, further research is necessary to assess the potential of GBMV2 models in predicting the binding free energies of protein-ligand complexes. These studies will help elucidate the strengths and limitations of GBMV2 models and contribute to the development of more accurate computational methods for binding free energy predictions.

BIBLIOGRAPHY

1. Csizmok, V., et al., *Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling.* Chemical Reviews, 2016. **116**(11): p. 6424-6462.
2. Oldfield, C.J. and A.K. Dunker, *Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions.* Annual Review of Biochemistry, 2014. **83**(1): p. 553-584.
3. Wright, P.E. and H.J. Dyson, *Intrinsically disordered proteins in cellular signalling and regulation.* Nature Reviews Molecular Cell Biology, 2015. **16**(1): p. 18-29.
4. Uversky, V.N., *Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders.* Frontiers in Aging Neuroscience, 2015. **7**: p. 18.
5. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nature Reviews Molecular Cell Biology, 2005. **6**(3): p. 197-208.
6. Owen, I. and F. Shewmaker, *The Role of Post-Translational Modifications in the Phase Transitions of Intrinsically Disordered Proteins.* International Journal of Molecular Sciences, 2019. **20**(21).
7. Chen, J., *Towards the physical basis of how intrinsic disorder mediates protein function.* Archives of Biochemistry and Biophysics, 2012. **524**(2): p. 123-31.
8. Das, R.K., K.M. Ruff, and R.V. Pappu, *Relating sequence encoded information to form and function of intrinsically disordered proteins.* Curr Opin Struct Biol, 2015. **32**: p. 102-112.
9. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
10. Vacic, V. and L.M. Iakoucheva, *Disease mutations in disordered regions-exception to the rule?* Molecular Biosystems, 2012. **8**(1): p. 27-32.
11. Kulkarni, P. and V.N. Uversky, *Intrinsically Disordered Proteins in Chronic Diseases.* Biomolecules, 2019. **9**(4).
12. Oldfield, C.J., et al., *Comparing and Combining Predictors of Mostly Disordered Proteins.* Biochemistry, 2005. **44**(6): p. 1989-2000.
13. Chen, J., X. Liu, and J. Chen, *Targeting Intrinsically Disordered Proteins through Dynamic Interactions.* Biomolecules, 2020. **10**(5).
14. Mittag, T., et al., *Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase.* Structure, 2010. **18**(4): p. 494-506.
15. McDowell, C., J. Chen, and J. Chen, *Potential Conformational Heterogeneity of p53 Bound to S100B(betabeta).* Journal of Molecular Biology, 2013.
16. Wu, H. and M. Fuxreiter, *The Structure and Dynamics of Higher-Order Assemblies: Amyloids, Signalosomes, and Granules.* Cell, 2016. **165**(5): p. 1055-1066.
17. Krois, A.S., et al., *Recognition of the disordered p53 transactivation domain by the transcriptional adapter zinc finger domains of CREB-binding protein.* Proc Natl Acad Sci U S A, 2016.
18. Csizmok, V., et al., *An allosteric conduit facilitates dynamic multisite substrate recognition by the SCFCdc4 ubiquitin ligase.* Nature Communications, 2017. **8**(1): p. 13943.

19. Borgia, A., et al., *Extreme disorder in an ultrahigh-affinity protein complex.* Nature, 2018. **555**(7694): p. 61-66.
20. Clark, S., et al., *Multivalency regulates activity in an intrinsically disordered transcription factor.* Elife, 2018. **7**.
21. Fuxreiter, M., *Fuzziness in Protein Interactions-A Historical Perspective.* Journal of Molecular Biology, 2018. **430**(16): p. 2278-2287.
22. Weng, J. and W. Wang, *Dynamic multivalent interactions of intrinsically disordered proteins.* Curr Opin Struct Biol, 2019. **62**: p. 9-13.
23. Miskei, M., C. Antal, and M. Fuxreiter, *FuzDB: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies.* Nucleic Acids Research, 2017. **45**(D1): p. D228-D235.
24. Ganguly, D. and J. Chen, *Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states.* J Mol Biol, 2009. **390**(3): p. 467-77.
25. Fisher, C.K. and C.M. Stultz, *Constructing ensembles for intrinsically disordered proteins.* Current Opinion in Structural Biology, 2011. **21**(3): p. 426-31.
26. Ferreon, A.C., et al., *Modulation of allostery by protein intrinsic disorder.* Nature, 2013. **498**(7454): p. 390-4.
27. Garcia-Pino, A., et al., *Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity.* Cell, 2010. **142**(1): p. 101-11.
28. Berlow, R.B., H.J. Dyson, and P.E. Wright, *Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation.* Journal of Molecular Biology, 2018. **430**(16): p. 2309-2320.
29. Levine, Z.A. and J.-E. Shea, *Simulations of disordered proteins and systems with conformational heterogeneity.* Current Opinion in Structural Biology, 2017. **43**: p. 95-103.
30. Knott, M. and R.B. Best, *A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations.* Plos Computational Biology, 2012. **8**(7): p. e1002605.
31. Mao, A.H., et al., *Net charge per residue modulates conformational ensembles of intrinsically disordered proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(18): p. 8183-8188.
32. Ganguly, D. and J. Chen, *Atomistic details of the disordered states of KID and pKID. implications in coupled binding and folding.* Journal of the American Chemical Society, 2009. **131**(14): p. 5214-5223.
33. Zhang, W., D. Ganguly, and J. Chen, *Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins.* Plos Computational Biology, 2012. **8**(1): p. e1002353.
34. Zhang, W. and J. Chen, *Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models.* Journal of Chemical Theory and Computation, 2014. **10**(3): p. 918-923.
35. Moritsugu, K., T. Terada, and A. Kidera, *Scalable free energy calculation of proteins via multiscale essential sampling.* J Chem Phys, 2010. **133**(22): p. 224105.

36. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1): p. 141-151.
37. Liu, P., et al., *Replica exchange with solute tempering: A method for sampling biological systems in explicit water.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(39): p. 13749.
38. Mittal, A., et al., *Hamiltonian Switch Metropolis Monte Carlo Simulations for Improved Conformational Sampling of Intrinsically Disordered Regions Tethered to Ordered Domains of Proteins.* Journal of Chemical Theory and Computation, 2014. **10**(8): p. 3550-3562.
39. Peter, E.K. and J.E. Shea, *A hybrid MD-kMC algorithm for folding proteins in explicit solvent.* Physical Chemistry Chemical Physics, 2014. **16**(14): p. 6430-6440.
40. Zhang, C. and J. Ma, *Enhanced sampling and applications in protein folding in explicit solvent.* Journal of chemical physics, 2010. **132**(24): p. 244101.
41. Zheng, L.Q. and W. Yang, *Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering.* Journal of Chemical Theory and Computation, 2012. **8**(3): p. 810-823.
42. Best, R.B., *Computational and theoretical advances in studies of intrinsically disordered proteins.* Current Opinion in Structural Biology, 2017. **42**: p. 147-154.
43. Kmiecik, S., et al., *Coarse-Grained Protein Models and Their Applications.* Chem Rev, 2016. **116**(14): p. 7898-936.
44. Arai, M., *Unified understanding of folding and binding mechanisms of globular and intrinsically disordered proteins.* Biophysical Reviews, 2018. **10**(2): p. 163-181.
45. Bhattacharya, S. and X. Lin, *Recent Advances in Computational Protocols Addressing Intrinsically Disordered Proteins.* Biomolecules, 2019. **9**(4).
46. Rauscher, S., et al., *Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment.* Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5513-5524.
47. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles.* J Chem Theory Comput, 2012. **8**(9): p. 3257-3273.
48. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states.* Proceedings of the National Academy of Sciences, 2018. **115**(21): p. E4758.
49. Huang, J. and A.D. MacKerell, *Force field development and simulations of intrinsically disordered proteins.* Current Opinion in Structural Biology, 2018. **48**: p. 40-48.
50. Piana, S., et al., *Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States.* The Journal of Physical Chemistry B, 2015. **119**(16): p. 5113-5123.
51. Wu, H.-N., F. Jiang, and Y.-D. Wu, *Significantly Improved Protein Folding Thermodynamics Using a Dispersion-Corrected Water Model and a New Residue-Specific Force Field.* The Journal of Physical Chemistry Letters, 2017. **8**(14): p. 3199-3205.

52. Mu, J., et al., *Recent Force Field Strategies for Intrinsically Disordered Proteins.* Journal of Chemical Information and Modeling, 2021. **61**(3): p. 1037-1047.
53. Song, D., et al., *Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins.* Journal of Chemical Information and Modeling, 2020. **60**(4): p. 2257-2267.
54. Yang, S., et al., *Residue-Specific Force Field Improving the Sample of Intrinsically Disordered Proteins and Folded Proteins.* Journal of Chemical Information and Modeling, 2019. **59**(11): p. 4793-4805.
55. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins.* Nature Methods, 2017. **14**(1): p. 71-73.
56. Tian, C., et al., *ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution.* Journal of Chemical Theory and Computation, 2020. **16**(1): p. 528-552.
57. Rahman, M.U., et al., *Comparison and Evaluation of Force Fields for Intrinsically Disordered Proteins.* Journal of Chemical Information and Modeling, 2020. **60**(10): p. 4912-4923.
58. Abriata, L.A. and M. Dal Peraro, *Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization.* Computational and Structural Biotechnology Journal, 2021. **19**: p. 2626-2636.
59. Piana, S., et al., *Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes.* Journal of Chemical Theory and Computation, 2020. **16**(4): p. 2494-2507.
60. Song, D., R. Luo, and H.-F. Chen, *The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins.* Journal of Chemical Information and Modeling, 2017. **57**(5): p. 1166-1178.
61. Jing, Z., et al., *Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications.* Annual Review of Biophysics, 2019. **48**(1): p. 371-394.
62. Bedrov, D., et al., *Molecular Dynamics Simulations of Ionic Liquids and Electrolytes Using Polarizable Force Fields.* Chemical Reviews, 2019. **119**(13): p. 7940-7995.
63. Inakollu, V.S.S., et al., *Polarisable force fields: what do they add in biomolecular simulations?* Current Opinion in Structural Biology, 2020. **61**: p. 182-190.
64. Huang, J. and Alexander D. MacKerell, *Induction of Peptide Bond Dipoles Drives Cooperative Helix Formation in the (AAQAA)3 Peptide.* Biophysical Journal, 2014. **107**(4): p. 991-997.
65. Kamenik, A.S., et al., *Polarizable and non-polarizable force fields: Protein folding, unfolding, and misfolding.* The Journal of Chemical Physics, 2020. **153**(18): p. 185102.
66. Wang, A., Z. Zhang, and G. Li, *Higher Accuracy Achieved in the Simulations of Protein Structure Refinement, Protein Folding, and Intrinsically Disordered Proteins Using Polarizable Force Fields.* The Journal of Physical Chemistry Letters, 2018. **9**(24): p. 7110-7116.

67. Wang, A., et al., *Quality of force fields and sampling methods in simulating pepX peptides: a case study for intrinsically disordered proteins.* Physical Chemistry Chemical Physics, 2021. **23**(3): p. 2430-2437.
68. Yang, Y.I., et al., *Enhanced sampling in molecular dynamics.* The Journal of Chemical Physics, 2019. **151**(7): p. 070902.
69. Wang, A.H., Z.C. Zhang, and G.H. Li, *Advances in Enhanced Sampling Molecular Dynamics Simulations for Biomolecules.* Chinese Journal of Chemical Physics, 2019. **32**(3): p. 277-286.
70. Barducci, A., M. Bonomi, and M. Parrinello, *Metadynamics.* WIREs Computational Molecular Science, 2011. **1**(5): p. 826-843.
71. Barducci, A., G. Bussi, and M. Parrinello, *Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method.* Physical Review Letters, 2008. **100**(2): p. 020603.
72. Hamelberg, D., J. Mongan, and J.A. McCammon, *Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules.* The Journal of Chemical Physics, 2004. **120**(24): p. 11919-11929.
73. Torrie, G.M. and J.P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics, 1977. **23**(2): p. 187-199.
74. Kästner, J., *Umbrella sampling.* WIREs Computational Molecular Science, 2011. **1**(6): p. 932-942.
75. Gao, Y.Q., *An integrate-over-temperature approach for enhanced sampling.* The Journal of Chemical Physics, 2008. **128**(6): p. 064105.
76. MacCallum, J.L., M.I. Muniyat, and K. Gaalswyk, *Online Optimization of Total Acceptance in Hamiltonian Replica Exchange Simulations.* The Journal of Physical Chemistry B, 2018. **122**(21): p. 5448-5457.
77. Liu, N., et al., *Phosphorylation regulates the binding of intrinsically disordered proteins via a flexible conformation selection mechanism.* Communications Chemistry, 2020. **3**(1): p. 123.
78. Dickson, A., L.S. Ahlstrom, and C.L. Brooks III, *Coupled folding and binding with 2D Window-Exchange Umbrella Sampling.* Journal of Computational Chemistry, 2016. **37**(6): p. 587-594.
79. Sidky, H., W. Chen, and A.L. Ferguson, *Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation.* Molecular Physics, 2020. **118**(5): p. e1737742.
80. Chen, W., A.R. Tan, and A.L. Ferguson, *Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design.* The Journal of Chemical Physics, 2018. **149**(7): p. 072312.
81. Marinari, E. and G. Parisi, *Simulated Tempering: A New Monte Carlo Scheme.* Europhysics Letters (EPL), 1992. **19**(6): p. 451-458.
82. Wang, L., R.A. Friesner, and B.J. Berne, *Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2).* The Journal of Physical Chemistry B, 2011. **115**(30): p. 9431-9438.
83. Miao, Y., et al., *Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation.* Journal of Chemical Theory and Computation, 2014. **10**(7): p. 2677-2689.

84. Kokubo, H., T. Tanaka, and Y. Okamoto, *Two-dimensional replica-exchange method for predicting protein–ligand binding structures.* Journal of Computational Chemistry, 2013. **34**(30): p. 2601-2614.
85. Oshima, H., S. Re, and Y. Sugita, *Replica-Exchange Umbrella Sampling Combined with Gaussian Accelerated Molecular Dynamics for Free-Energy Calculation of Biomolecules.* Journal of Chemical Theory and Computation, 2019. **15**(10): p. 5199-5208.
86. Peng, X., et al., *Integrating Multiple Accelerated Molecular Dynamics To Improve Accuracy of Free Energy Calculations.* Journal of Chemical Theory and Computation, 2018. **14**(3): p. 1216-1227.
87. Bussi, G. and A. Laio, *Using metadynamics to explore complex free-energy landscapes.* Nature Reviews Physics, 2020. **2**(4): p. 200-212.
88. Galvelis, R. and Y. Sugita, *Replica state exchange metadynamics for improving the convergence of free energy estimates.* Journal of Computational Chemistry, 2015. **36**(19): p. 1446-1455.
89. Piana, S. and A. Laio, *A Bias-Exchange Approach to Protein Folding.* The Journal of Physical Chemistry B, 2007. **111**(17): p. 4553-4559.
90. Galvelis, R., S. Re, and Y. Sugita, *Enhanced Conformational Sampling of N-Glycans in Solution with Replica State Exchange Metadynamics.* Journal of Chemical Theory and Computation, 2017. **13**(5): p. 1934-1942.
91. Do, T.N., W.-Y. Choy, and M. Karttunen, *Binding of Disordered Peptides to Kelch: Insights from Enhanced Sampling Simulations.* Journal of Chemical Theory and Computation, 2016. **12**(1): p. 395-404.
92. Guo, J. and H.X. Zhou, *Protein Allostery and Conformational Dynamics.* Chem Rev, 2016. **116**(11): p. 6503-15.
93. Gianni, S., et al., *Fuzziness and Frustration in the Energy Landscape of Protein Folding, Function, and Assembly.* Acc Chem Res, 2021. **54**(5): p. 1251-1259.
94. Neupane, K., A.P. Manuel, and M.T. Woodside, *Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes.* Nature Physics, 2016. **12**(7): p. 700-703.
95. Pfaendtner, J. and M. Bonomi, *Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics.* J Chem Theory Comput, 2015. **11**(11): p. 5062-7.
96. Prakash, A., et al., *Biasing Smarter, Not Harder, by Partitioning Collective Variables into Families in Parallel Bias Metadynamics.* J Chem Theory Comput, 2018. **14**(10): p. 4985-4990.
97. Awasthi, S. and N.N. Nair, *Exploring high dimensional free energy landscapes: Temperature accelerated sliced sampling.* The Journal of Chemical Physics, 2017. **146**(9): p. 094108.
98. Chen, W. and A.L. Ferguson, *Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration.* Journal of Computational Chemistry, 2018. **39**(25): p. 2079-2102.
99. Galvelis, R. and Y. Sugita, *Neural Network and Nearest Neighbor Algorithms for Enhancing Sampling of Molecular Dynamics.* Journal of Chemical Theory and Computation, 2017. **13**(6): p. 2489-2500.

100. Salawu, E.O., *DESP: Deep Enhanced Sampling of Proteins' Conformation Spaces Using AI-Inspired Biasing Forces.* Frontiers in Molecular Biosciences, 2021. **8**(121).
101. Zhang, J. and M. Chen, *Unfolding Hidden Barriers by Active Enhanced Sampling.* Physical Review Letters, 2018. **121**(1): p. 010601.
102. Brown, S. and T. Head-Gordon, *Cool walking: A new Markov chain Monte Carlo sampling method.* Journal of Computational Chemistry, 2003. **24**(1): p. 68-76.
103. Neal, R.M., *Annealed importance sampling.* Statistics and Computing, 2001. **11**(2): p. 125-139.
104. Fukunishi, H., O. Watanabe, and S. Takada, *On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction.* The Journal of Chemical Physics, 2002. **116**(20): p. 9058-9067.
105. Huang, Y.-m.M., J.A. McCammon, and Y. Miao, *Replica Exchange Gaussian Accelerated Molecular Dynamics: Improved Enhanced Sampling and Free Energy Calculation.* Journal of Chemical Theory and Computation, 2018. **14**(4): p. 1853-1864.
106. Wang, J., et al., *Gaussian accelerated molecular dynamics: Principles and applications.* WIREs Computational Molecular Science, 2021. **11**(5): p. e1521.
107. Miao, Y., A. Bhattarai, and J. Wang, *Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics.* Journal of Chemical Theory and Computation, 2020. **16**(9): p. 5526-5547.
108. Wang, J. and Y. Miao, *Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): Enhanced sampling and free energy and kinetics calculations of peptide binding.* The Journal of Chemical Physics, 2020. **153**(15): p. 154109.
109. Kamiya, M. and Y. Sugita, *Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations.* The Journal of Chemical Physics, 2018. **149**(7): p. 072304.
110. Liu, X. and J. Chen, *Residual Structures and Transient Long-Range Interactions of p53 Transactivation Domain: Assessment of Explicit Solvent Protein Force Fields.* Journal of Chemical Theory and Computation, 2019. **15**(8): p. 4708-4720.
111. Shrestha, U.R., J.C. Smith, and L. Petridis, *Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations.* Communications Biology, 2021. **4**(1): p. 243.
112. Hicks, A. and H.-X. Zhou, *Temperature-induced collapse of a disordered peptide observed by three sampling methods in molecular dynamics simulations.* The Journal of Chemical Physics, 2018. **149**(7): p. 072313.
113. Brown, A.H., et al., *Equilibrium Conformational Ensemble of the Intrinsically Disordered Peptide n16N: Linking Subdomain Structures and Function in Nacre.* Biomacromolecules, 2014. **15**(12): p. 4467-4479.
114. Pang, X. and H.-X. Zhou, *Disorder-to-Order Transition of an Active-Site Loop Mediates the Allosteric Activation of Sortase A.* Biophysical Journal, 2015. **109**(8): p. 1706-1715.

115. Liu, X., Z. Jia, and J. Chen, *Enhanced Sampling of Intrinsic Structural Heterogeneity of the BH3-Only Protein Binding Interface of Bcl-xL.* The Journal of Physical Chemistry B, 2017. **121**(39): p. 9160-9168.
116. Liang, C., et al., *Modulation of Amyloid-beta42 Conformation by Small Molecules Through Nonspecific Binding.* J Chem Theory Comput, 2019. **15**(10): p. 5169-5174.
117. Liu, X. and J. Chen, *Modulation of p53 Transactivation Domain Conformations by Ligand Binding and Cancer-Associated Mutations.* Pac. Symp. Biocomput., 2020. **25**: p. 195-206.
118. Schrag, L.G., et al., *Cancer-Associated Mutations Perturb the Disordered Ensemble and Interactions of the Intrinsically Disordered p53 Transactivation Domain.* J Mol Biol, 2021. **433**(15): p. 167048.
119. Zhao, J., et al., *EGCG binds intrinsically disordered N-terminal domain of p53 and disrupts p53-MDM2 interaction.* Nat Commun, 2021. **12**(1): p. 986.
120. Wang, L., et al., *Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field.* Journal of the American Chemical Society, 2015. **137**(7): p. 2695-2703.
121. Zhou, H.X., *Theoretical frameworks for multiscale modeling and simulation.* Current Opinion in Structural Biology, 2014. **25C**: p. 67-76.
122. Lee, K.H. and J. Chen, *Multiscale enhanced sampling of intrinsically disordered protein conformations.* Journal of Computational Chemistry, 2016. **37**(6): p. 550-557.
123. Liu, X.R., X.P. Gong, and J.H. Chen, *Accelerating atomistic simulations of proteins using multiscale enhanced sampling with independent tempering.* Journal of Computational Chemistry, 2021. **42**(5): p. 358-364.
124. Lee, K.H. and J. Chen, *Optimization of the GBMV2 implicit solvent force field for accurate simulation of protein conformational equilibria.* Journal of Computational Chemistry, 2017. **38**(16): p. 1332-1341.
125. Liu, Y., et al., *Coupling Coarse-Grained to Fine-Grained Models via Hamiltonian Replica Exchange.* Journal of Chemical Theory and Computation, 2020. **16**(8): p. 5313-5322.
126. Yang, Y.I., H. Niu, and M. Parrinello, *Combining Metadynamics and Integrated Tempering Sampling.* The Journal of Physical Chemistry Letters, 2018. **9**(22): p. 6426-6430.
127. Shirts, M.R. and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states.* The Journal of Chemical Physics, 2008. **129**(12): p. 124105.
128. Kumar, S., et al., *THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.
129. Sinko, W., et al., *Population Based Reweighting of Scaled Molecular Dynamics.* The Journal of Physical Chemistry B, 2013. **117**(42): p. 12759-12768.
130. Ilie, I.M. and A. Caflisch, *Simulation Studies of Amyloidogenic Polypeptides and Their Aggregates.* Chemical Reviews, 2019. **119**(12): p. 6956-6993.

131. Zhou, H.X. and X. Pang, *Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation.* Chem Rev, 2018. **118**(4): p. 1691-1741.
132. Fassler, J.S., et al., *Protein Aggregation and Disaggregation in Cells and Development.* J Mol Biol, 2021: p. 167215.
133. Alberti, S., A. Gladfelter, and T. Mittag, *Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates.* Cell, 2019. **176**(3): p. 419-434.
134. Holehouse, A.S. and R.V. Pappu, *Functional Implications of Intracellular Phase Transitions.* Biochemistry, 2018. **57**(17): p. 2415-2423.
135. Brangwynne, C.P., P. Tompa, and R.V. Pappu, *Polymer physics of intracellular phase transitions.* Nature Physics, 2015. **11**(11): p. 899-904.
136. Mathieu, C., R.V. Pappu, and J.P. Taylor, *Beyond aggregation: Pathological phase transitions in neurodegenerative disease.* Science, 2020. **370**(6512): p. 56-60.
137. Chen, J., C.L. Brooks, and J. Khandogin, *Recent advances in implicit solvent based methods for biomolecular simulations.* Current Opinion in Structural Biology, 2008. **18**: p. 140-148.
138. Haberthür, U. and A. Caflisch, *FACTS: Fast analytical continuum treatment of solvation.* Journal of Computational Chemistry, 2008. **29**(5): p. 701-715.
139. Hawkins, G.D., C.J. Cramer, and D.G. Truhlar, *Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium.* The Journal of Physical Chemistry, 1996. **100**(51): p. 19824-19839.
140. Onufriev, A., D. Bashford, and D.A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(2): p. 383-394.
141. Mongan, J., et al., *Generalized Born Model with a Simple, Robust Molecular Volume Correction.* Journal of Chemical Theory and Computation, 2007. **3**(1): p. 156-169.
142. Nguyen, H., D.R. Roe, and C. Simmerling, *Improved Generalized Born Solvent Model Parameters for Protein Simulations.* Journal of Chemical Theory and Computation, 2013. **9**(4): p. 2020-2034.
143. Gallicchio, E. and R.M. Levy, *AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling.* Journal of Computational Chemistry, 2004. **25**(4): p. 479-499.
144. Gallicchio, E., K. Paris, and R.M. Levy, *The AGBNP2 Implicit Solvation Model.* Journal of Chemical Theory and Computation, 2009. **5**(9): p. 2544-2564.
145. Im, W., M.S. Lee, and C.L. Brooks Iii, *Generalized born model with a simple smoothing function.* Journal of Computational Chemistry, 2003. **24**(14): p. 1691-1702.
146. Lee, M.S., F.R. Salsbury, and C.L. Brooks, *Novel generalized Born methods.* The Journal of Chemical Physics, 2002. **116**(24): p. 10606-10614.
147. Lee, M.S., et al., *New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations.* Journal of Computational Chemistry, 2003. **24**(11): p. 1348-1356.

148. Chen, J., W. Im, and C.L. Brooks, *Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field.* Journal of the American Chemical Society, 2006. **128**(11): p. 3728-3736.
149. Chen, J. and C.L. Brooks III, *Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions.* Physical Chemistry Chemical Physics, 2008. **10**(4): p. 471-481.
150. Chen, J., *Intrinsically disordered p53 extreme C-terminus binds to S100B(betabeta) through "fly-casting".* Journal of the American Chemical Society, 2009. **131**(6): p. 2088-9.
151. Wang, Y., et al., *Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21.* Nature chemical biology, 2011. **7**(4): p. 214-21.
152. Ganguly, D. and J. Chen, *Modulation of the disordered conformational ensembles of the p53 transactivation domain by cancer-associated mutations.* PLoS Comput Biol, 2015. **11**(4): p. e1004247.
153. Nguyen, H., et al., *Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent.* Journal of the American Chemical Society, 2014. **136**(40): p. 13959-13962.
154. Maffucci, I. and A. Contini, *An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, β-Hairpin, and Intrinsically Disordered Peptides.* Journal of Chemical Theory and Computation, 2016. **12**(2): p. 714-727.
155. Tao, P. and Y. Xiao, *Using the generalized Born surface area model to fold proteins yields more effective sampling while qualitatively preserving the folding landscape.* Physical Review E, 2020. **101**(6): p. 062417.
156. Gong, X., et al., *Accelerating the Generalized Born with Molecular Volume and Solvent Accessible Surface Area Implicit Solvent Model Using Graphics Processing Units.* Journal of Computational Chemistry, 2020. **41**(8): p. 830-838.
157. Vitalis, A. and R.V. Pappu, *ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions.* Journal of Computational Chemistry, 2009. **30**(5): p. 673-699.
158. Vitalis, A. and A. Caflisch, *Micelle-Like Architecture of the Monomer Ensemble of Alzheimer's Amyloid-β Peptide in Aqueous Solution and Its Implications for Aβ Aggregation.* Journal of Molecular Biology, 2010. **403**(1): p. 148-165.
159. Mittal, A., et al., *Sequence-to-Conformation Relationships of Disordered Regions Tethered to Folded Domains of Proteins.* J Mol Biol, 2018. **430**(16): p. 2403-2421.
160. Choi, J.-M. and R.V. Pappu, *Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics.* Journal of Chemical Theory and Computation, 2019. **15**(2): p. 1367-1382.
161. Pak, A.J. and G.A. Voth, *Advances in coarse-grained modeling of macromolecular complexes.* Curr Opin Struct Biol, 2018. **52**: p. 119-126.
162. Wolynes, P.G., *Recent successes of the energy landscape theory of protein folding and function.* Quarterly Reviews of Biophysics, 2005. **38**(4): p. 405-410.

163. Hills, R.D. and C.L. Brooks, *Insights from Coarse-Grained Gō Models for Protein Folding and Dynamics.* International Journal of Molecular Sciences, 2009. **10**(3).
164. Law, S.M., et al., *Prepaying the entropic cost for allosteric regulation in KIX.* Proc Natl Acad Sci U S A, 2014. **111**(33): p. 12067-72.
165. Chu, X. and J. Wang, *Position-, disorder-, and salt-dependent diffusion in binding-coupled-folding of intrinsically disordered proteins.* Phys Chem Chem Phys, 2019. **21**(10): p. 5634-5645.
166. Ganguly, D., W. Zhang, and J. Chen, *Synergistic folding of two intrinsically disordered proteins: searching for conformational selection.* Molecular BioSystems, 2012. **8**(1): p. 198-209.
167. Ganguly, D., W. Zhang, and J. Chen, *Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins.* PLoS Computational Biology, 2013. **9**(11): p. e1003363.
168. Liu, Z.R. and Y.Q. Huang, *Advantages of proteins being disordered.* Protein Science, 2014. **23**(5): p. 539-550.
169. Ganguly, D. and J. Chen, *Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(4): p. 1251-1266.
170. Chu, W.T., S.L. Shammas, and J. Wang, *Charge Interactions Modulate the Encounter Complex Ensemble of Two Differently Charged Disordered Protein Partners of KIX.* J Chem Theory Comput, 2020. **16**(6): p. 3856-3868.
171. Baul, U., et al., *Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins.* The Journal of Physical Chemistry B, 2019. **123**(16): p. 3462-3474.
172. Liu, X. and J. Chen, *HyRes: a coarse-grained model for multi-scale enhanced sampling of disordered protein conformations.* Physical Chemistry Chemical Physics, 2017. **19**(48): p. 32421-32432.
173. Wu, H., P.G. Wolynes, and G.A. Papoian, *AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins.* The Journal of Physical Chemistry B, 2018. **122**(49): p. 11115-11125.
174. Wang, W., *Recent advances in atomic molecular dynamics simulation of intrinsically disordered proteins.* Physical Chemistry Chemical Physics, 2021. **23**(2): p. 777-784.
175. Shea, J.-E., R.B. Best, and J. Mittal, *Physics-based computational and theoretical approaches to intrinsically disordered proteins.* Current Opinion in Structural Biology, 2021. **67**: p. 219-225.
176. Ashbaugh, H.S. and H.W. Hatch, *Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space.* Journal of the American Chemical Society, 2008. **130**(29): p. 9536-9542.
177. Kim, Y.C. and G. Hummer, *Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding.* Journal of Molecular Biology, 2008. **375**(5): p. 1416-1433.
178. Dignon, G.L., et al., *Relation between single-molecule properties and phase behavior of intrinsically disordered proteins.* Proceedings of the National Academy of Sciences, 2018. **115**(40): p. 9929.

179. Latham, A.P. and B. Zhang, *Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins.* Journal of Chemical Theory and Computation, 2020. **16**(1): p. 773-781.
180. Nerenberg, P.S. and T. Head-Gordon, *New developments in force fields for biomolecular simulations.* Current Opinion in Structural Biology, 2018. **49**: p. 129-138.
181. Nerenberg, P.S., et al., *Optimizing Solute–Water van der Waals Interactions To Reproduce Solvation Free Energies.* The Journal of Physical Chemistry B, 2012. **116**(15): p. 4524-4534.
182. Henriques, J., C. Cragnell, and M. Skepö, *Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment.* Journal of Chemical Theory and Computation, 2015. **11**(7): p. 3420-3431.
183. Best, R.B., W.W. Zheng, and J. Mittal, *Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association.* Journal of Chemical Theory and Computation, 2014. **10**(11): p. 5113-5124.
184. Matsubara, D., et al. *Modified Protein-Water Interactions in CHARMM36m for Thermodynamics and Kinetics of Proteins in Dilute and Crowded Solutions*. Molecules, 2022. **27**, DOI: 10.3390/molecules27175726.
185. Qiu, Y., W. Shan, and H. Zhang, *Force Field Benchmark of Amino Acids. 3. Hydration with Scaled Lennard-Jones Interactions.* Journal of Chemical Information and Modeling, 2021. **61**(7): p. 3571-3582.
186. Gil Pineda, L.I., L.N. Milko, and Y. He, *Performance of CHARMM36m with modified water model in simulating intrinsically disordered proteins: a case study.* Biophysics Reports, 2020. **6**(2): p. 80-87.
187. Rieloff, E. and M. Skepö *Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison*. International Journal of Molecular Sciences, 2021. **22**, DOI: 10.3390/ijms221810174.
188. Rieloff, E. and M. Skepö, *Phosphorylation of a Disordered Peptide—Structural Effects and Force Field Inconsistencies.* Journal of Chemical Theory and Computation, 2020. **16**(3): p. 1924-1935.
189. Makowski, M., A. Liwo, and H.A. Scheraga, *Simple Physics-Based Analytical Formulas for the Potentials of Mean Force of the Interaction of Amino Acid Side Chains in Water. VII. Charged–Hydrophobic/Polar and Polar–Hydrophobic/Polar Side Chains.* The Journal of Physical Chemistry B, 2017. **121**(2): p. 379-390.
190. Masunov, A. and T. Lazaridis, *Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water.* Journal of the American Chemical Society, 2003. **125**(7): p. 1722-1730.
191. Abraham, M.J., et al., *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* SoftwareX, 2015. **1-2**: p. 19-25.
192. Eastman, P., et al., *OpenMM 7: Rapid development of high performance algorithms for molecular dynamics.* PLOS Computational Biology, 2017. **13**(7): p. e1005659.

193. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
194. Vaiana, S.M., et al., *The role of solvent in protein folding and in aggregation.* Journal of Biological Physics, 2001. **27**(2-3): p. 133-145.
195. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Protein Simulations, 2003. **66**: p. 27-85.
196. Mackerell, A.D., *Empirical force fields for biological macromolecules: Overview and issues.* Journal of Computational Chemistry, 2004. **25**(13): p. 1584-1604.
197. Wagoner, J. and N.A. Baker, *Solvation forces on biomolecular structures: A comparison of explicit solvent and Poisson-Boltzmann models.* Journal of Computational Chemistry, 2004. **25**(13): p. 1623-1629.
198. Onufriev, A., D.A. Case, and D. Bashford, *Effective Born radii in the generalized Born approximation: The importance of being perfect.* Journal of Computational Chemistry, 2002. **23**(14): p. 1297-1304.
199. Cramer, C.J. and D.G. Truhlar, *Implicit solvation models: Equilibria, structure, spectra, and dynamics.* Chemical Reviews, 1999. **99**(8): p. 2161-2200.
200. Roux, B.t. and T. Simonson, *Implicit solvent models.* Biophysical Chemistry, 1999. **78**(1): p. 1-20.
201. Feig, M. and C.L. Brooks, *Recent advances in the development and application of implicit solvent models in biomolecule simulations.* Current Opinion in Structural Biology, 2004. **14**(2): p. 217-224.
202. Baker, N.A., *Improving implicit solvent simulations: a Poisson-centric view.* Current Opinion in Structural Biology, 2005. **15**(2): p. 137-143.
203. Gilson, M.K., et al., *Computation of Electrostatic Forces on Solvated Molecules Using the Poisson-Boltzmann Equation.* Journal of Physical Chemistry, 1993. **97**(14): p. 3591-3600.
204. Nicholls, A. and B. Honig, *A Rapid Finite-Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation.* Journal of Computational Chemistry, 1991. **12**(4): p. 435-445.
205. Bashford, D. and D.A. Case, *Generalized born models of macromolecular solvation effects.* Annual Review of Physical Chemistry, 2000. **51**: p. 129-152.
206. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular Simulations.* Biopolymers, 2001. **56**(4): p. 275-291.
207. Chen, J.H., *Effective Approximation of Molecular Volume Using Atom-Centered Dielectric Functions in Generalized Born Models.* Journal of Chemical Theory and Computation, 2010. **6**(9): p. 2790-2803.
208. Ghosh, A., C.S. Rapp, and R.A. Friesner, *Generalized born model based on a surface integral formulation.* Journal of Physical Chemistry B, 1998. **102**(52): p. 10983-10990.
209. Chocholousova, J. and M. Feig, *Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations.* Journal of Computational Chemistry, 2006. **27**(6): p. 719-729.

210. Nguyen, H., D. Roe, and C. Simmerling, *Improved generalized Born solvent model parameters for protein and nucleic acid simulations.* Abstracts of Papers of the American Chemical Society, 2012. **244**.
211. Richards, F.M., *Areas, Volumes, Packing, and Protein-Structure.* Annual Review of Biophysics and Bioengineering, 1977. **6**: p. 151-176.
212. Juneja, A., M. Ito, and L. Nilsson, *Implicit Solvent Models and Stabilizing Effects of Mutations and Ligands on the Unfolding of the Amyloid beta-Peptide Central Helix.* Journal of Chemical Theory and Computation, 2013. **9**(1): p. 834-846.
213. Knight, J.L. and C.L. Brooks, *Surveying Implicit Solvent Models for Estimating Small Molecule Absolute Hydration Free Energies.* Journal of Computational Chemistry, 2011. **32**(13): p. 2909-2923.
214. Piana, S., J.L. Klepeis, and D.E. Shaw, *Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.* Current Opinion in Structural Biology, 2014. **24**: p. 98-105.
215. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program.* Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
216. Case, D.A., et al., *The Amber biomolecular simulation programs.* Journal of Computational Chemistry, 2005. **26**(16): p. 1668-1688.
217. Salomon-Ferrer, R., et al., *Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald.* Journal of Chemical Theory and Computation, 2013. **9**(9): p. 3878-3888.
218. Hess, B., et al., *GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.
219. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.
220. Tanner, D.E., J.C. Phillips, and K. Schulten, *GPU/CPU Algorithm for Generalized Born/Solvent-Accessible Surface Area Implicit Solvent Calculations.* Journal of Chemical Theory and Computation, 2012. **8**(7): p. 2521-2530.
221. Eastman, P., et al., *OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation.* Journal of Chemical Theory and Computation, 2013. **9**(1): p. 461-469.
222. Gotz, A.W., et al., *Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born.* Journal of Chemical Theory and Computation, 2012. **8**(5): p. 1542-1555.
223. Arthur, E.J. and C.L. Brooks Iii, *Parallelization and improvements of the generalized born model with a simple sWitching function for modern graphics processors.* Journal of Computational Chemistry, 2016. **37**(10): p. 927-939.
224. Huang, H. and C. Simmerling, *Fast Pairwise Approximation of Solvent Accessible Surface Area for Implicit Solvent Simulations of Proteins on CPUs and GPUs.* Journal of Chemical Theory and Computation, 2018. **14**(11): p. 5797-5814.
225. Still, W.C., et al., *Semianalytical treatment of solvation for molecular mechanics and dynamics.* Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.

226. Im, W.P., M.S. Lee, and C.L. Brooks, *Generalized born model with a simple smoothing function.* Journal of Computational Chemistry, 2003. **24**(14): p. 1691-1702.
227. Liu, X., J. Chen, and J. Chen, *Residual Structure Accelerates Binding of Intrinsically Disordered ACTR by Promoting Efficient Folding upon Encounter.* J Mol Biol, 2019. **431**(2): p. 422-432.
228. Dunker, A.K., et al., *Intrinsically disordered protein.* Journal of Molecular Graphics & Modelling, 2001. **19**(1): p. 26-59.
229. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.* Journal of Molecular Recognition, 2005. **18**(5): p. 343-384.
230. Schmidtgall, B., et al., *Dissecting mechanism of coupled folding and binding of an intrinsically disordered protein by chemical synthesis of conformationally constrained analogues.* Chemical Communications, 2017. **53**(53): p. 7369-7372.
231. Sugase, K., H.J. Dyson, and P.E. Wright, *Mechanism of coupled folding and binding of an intrinsically disordered protein.* Nature, 2007. **447**(7147): p. 1021-1025.
232. Ithuralde, R.E., A.E. Roitberg, and A.G. Turjanski, *Structured and Unstructured Binding of an Intrinsically Disordered Protein as Revealed by Atomistic Simulations.* Journal of the American Chemical Society, 2016. **138**(28): p. 8742-8751.
233. Palazzesi, F., et al., *Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States.* Journal of Chemical Theory and Computation, 2015. **11**(1): p. 2-7.
234. Click, T.H., D. Ganguly, and J. Chen, *Intrinsically Disordered Proteins in a Physics-Based World.* International Journal of Molecular Sciences, 2010. **11**(12).
235. Glaser, J., et al., *Strong scaling of general-purpose molecular dynamics simulations on GPUs.* Computer Physics Communications, 2015. **192**: p. 97-107.
236. Baker, C.M. and R.B. Best, *Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2014. **4**(3): p. 182-198.
237. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems.* Biochimica Et Biophysica Acta-General Subjects, 2015. **1850**(5): p. 872-877.
238. Feig, M., J. Karanicolas, and C.L. Brooks, *MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology.* Journal of Molecular Graphics and Modelling, 2004. **22**(5): p. 377-395.
239. Onufriev, A.V. and D.A. Case, *Generalized Born Implicit Solvent Models for Biomolecules.* Annual Review of Biophysics, 2019. **48**(1): p. 275-296.
240. Anandakrishnan, R., et al., *Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations.* Biophysical Journal, 2015. **108**(5): p. 1153-1164.
241. Onufriev, A., *Chapter 7 - Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview*, in *Annual Reports in Computational Chemistry*, R.A. Wheeler and D.C. Spellmeyer, Editors. 2008, Elsevier. p. 125-137.

242. Levy, R.M., et al., *On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy.* Journal of the American Chemical Society, 2003. **125**(31): p. 9523-9530.
243. Shivakumar, D., Y. Deng, and B. Roux, *Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model.* Journal of Chemical Theory and Computation, 2009. **5**(4): p. 919-930.
244. Zacharias, M., *Continuum Solvent Modeling of Nonpolar Solvation: Improvement by Separating Surface Area Dependent Cavity and Dispersion Contributions.* The Journal of Physical Chemistry A, 2003. **107**(16): p. 3000-3004.
245. Jones, D., et al., *Accelerators for Classical Molecular Dynamics Simulations of Biomolecules.* Journal of Chemical Theory and Computation, 2022. **18**(7): p. 4047-4069.
246. Phillips, J.C., et al., *Scalable molecular dynamics on CPU and GPU architectures with NAMD.* The Journal of Chemical Physics, 2020. **153**(4).
247. Weeks, J.D., D. Chandler, and H.C. Andersen, *Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids.* The Journal of Chemical Physics, 2003. **54**(12): p. 5237-5247.
248. Deng, Y. and B. Roux, *Hydration of Amino Acid Side Chains: Nonpolar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with Explicit Water Molecules.* The Journal of Physical Chemistry B, 2004. **108**(42): p. 16567-16576.
249. Fesinmeyer, R.M., F.M. Hudson, and N.H. Andersen, *Enhanced Hairpin Stability through Loop Design: The Case of the Protein G B1 Domain Hairpin.* Journal of the American Chemical Society, 2004. **126**(23): p. 7238-7243.