

GPU-accelerated Protein Modeling and Structure Prediction using Molecular Dynamics

A Dissertation presented

by

He Huang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2018

Stony Brook University

The Graduate School

He Huang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

**Carlos Simmerling - Dissertation Advisor
Professor of Chemistry**

**Daniel P. Raleigh - Chairperson of Defense
Professor of Chemistry**

**Ken A. Dill - Third Inside Member of Defense
Distinguished Professor of Physics and Chemistry**

**Dmytro Kozakov - Outside Member of Defense
Assistant Professor of Applied Mathematics and Statistics**

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

GPU-accelerated Protein Modeling and Structure Prediction using Molecular Dynamics

by

He Huang

Doctor of Philosophy

in

Chemistry

Stony Brook University

2018

Physical potential and solvation dictate the underlying laws in molecular modeling. Accurate descriptions of both of them are key to structural predictions of proteins. For the physical potential validation, we demonstrated the evaluation and comparison of two force fields on their capabilities in reproducing crystal structure dihedral angles and helical propensity measured from chemical shift, respectively. The findings cross validate our understanding in protein backbone parameter modifications and point out the necessity of improving amino acid specificity in the current model. For more complete solvation description in implicit solvent, we focused on the nonpolar term which hasn't been extensively addressed and sometimes even neglected in calculations. In biomolecules where hydrophobic interactions play a central role, excluding nonpolar contributions can negatively influence the stability of the system. In my thesis, a multidisciplinary method of physical chemistry and scientific computing is adopted for a fast and accurate way to estimate the Solvent Accessible Surface Area (SASA) and calculate nonpolar free energy. The corresponding forces derived analytically are included in MD simulations to stabilize folded conformations. Implemented in Amber software and tested on consumer GPUs, this novel algorithm reasonably reproduces the simulation results of current implementation (LCPO), but accelerates MD simulations over 30 times, which is greatly desirable for protein simulations facing sampling challenges. With the associated parameter empirically calibrated against explicit solvent simulation results, we validated the GPU-accelerated GB/SA simulating four small proteins (Trp-cage, CLN025, Homeodomain and HP36). The predictions of protein melting behavior and structural equilibria are more consistent with experimental measurements, compared to the predictions without the nonpolar term. Prior to the method developments, MD simulations were applied in CASP protein structure refinement and protein aggregation studies. All these findings will provide insight and experience as to further research needs.

Dedication Page

In loving memory of my Mum 罗蓉(1965-2015)

”A heavy weight of hours has chain'd
and bow'd
这被岁月的重轭所制服的生命
One too like thee: tameless, and swift,
and proud.
原是和你一样：骄傲、轻捷而不驯。”

Percy Bysshe Shelley
Ode to the West Wind
雪莱《西风颂》

Contents

Abstract	iii
List of Figures	viii
List of Tables	x
List of Abbreviations	xi
Acknowledgements	xii
1 Introduction	1
1.1 Four levels of protein structures	2
1.2 Protein modeling techniques	5
1.2.1 Molecular Mechanics and force field	5
1.2.2 Solvation and implicit solvent	6
1.2.3 Molecular Dynamics and Amber software	7
1.2.4 Secondary structure analysis and clustering algorithm	8
1.3 Improvements needed in current modeling tools	9
1.3.1 Nonpolar solvation and GPU-parallelization	9
1.3.2 Secondary structure balance and amino acid backbone specificity . .	11
1.4 Overview of the studied questions	12
2 GPU-accelerated Nonpolar Solvation of Proteins: Fast Calculation of Accessible Surface Area in Implicit Solvent Simulations	14
2.1 Abstract	14
2.2 Introduction	14
2.3 Methods	16
2.3.1 Current theory	16
2.3.2 Proposed fast pairwise analytical estimation algorithm	17
2.3.3 Simulated protein systems	22
2.3.4 MD simulation and analysis details	23
2.4 Results and Discussions	25
2.4.1 SASA estimation by the proposed algorithm	25
2.4.2 Estimation of molecular SASA in test set	27
2.4.3 Speed up in MD simulations	29
2.4.4 Stability of the hydrophobic core in the HC16 model system	30

2.4.5	Quantification of discrepancies between GB and TIP3P	31
2.4.6	The new pairwise algorithm closely matches LCPO	32
2.4.7	GB/SA solvation with reasonable surface tension can reproduce TIP3P profile	34
2.4.8	Application to unrestrained proteins	34
2.5	Conclusion	38
2.6	Supporting Information	39
3	Toolbox Development for Amino Acid Specificity Evaluations	54
3.1	Abstract	54
3.2	Introduction	54
3.3	Methods	56
3.3.1	System setup	56
3.3.2	Simulation details	57
3.3.3	Dihedral stability and inter-conversion ratio calculations	58
3.3.4	Helical propensity calculations	61
3.4	Results and Discussions	62
3.4.1	Diverse amino acid specific dihedral stabilities in different models	62
3.4.2	Helical propensities indicate the necessity of amino acid specific backbone parameters	68
3.4.3	Cross-validation of stability test and helical propensity test	69
3.5	Conclusions	70
3.6	Supporting Information	71
4	Strategic Refinement of Homology-Modeled and GB-Folded Protein Structures	74
4.1	Abstract	74
4.2	Introduction	74
4.3	Methods	76
4.3.1	Refinement targets	76
4.3.2	Simulation details	78
4.3.3	Evaluation Criterion	79
4.4	Results and Discussions	79
4.4.1	Refinements from GB predicted structures	79
4.4.2	Stability of CASP11 experimental structures in MD	81
4.4.3	Best frame in simulations for CASP11	85
4.4.4	Larger cluster size indicates higher confidence in refinement	88
4.5	Conclusions	92
4.6	Supporting Information	93
5	Study on Mechanism of IAPP Amyloid Fibril Initial Formation	98
5.1	Abstract	98
5.2	Introduction	98
5.3	Methods	105
5.3.1	Hypothesis validation plan and system setup	105

5.3.2	Simulation details	110
5.4	Results and Discussions	111
5.4.1	Benchmark of four computation models in modeling hIAPP monomer	111
5.4.2	Transient secondary structures in monomeric hIAPP	113
5.4.3	Helical fractions and kinetic rates in IAPP variants	114
5.4.4	Dimer models proposed as the smallest oligomer	116
5.5	Conclusions	122
5.6	Supporting Information	122
6	Conclusion and Prospective	126
	Bibliography	129

List of Figures

1.1	Structure of amino acid	2
1.2	Definition of secondary structures	3
1.3	Protein data set of diverse topologies with designated name and size	4
1.4	Thermodynamic cycle of the solvation process.	10
2.1	2D illustration of pairwise SASA calculation	19
2.2	Structure and restraints of HC16	23
2.3	Atomic SASA estimation for training set	26
2.4	Molecular SASA estimation for test set	28
2.5	Performance benchmark for GPU-accelerated MD simulations	30
2.6	Comparison of structural equilibria of HC16 in GB, GB/SA and TIP3P models	32
2.7	Convergence comparison of two GB/SA methods: LCPO and our method	33
2.8	Melting curves of four proteins	36
2.9	HP36 simulated structural equilibria using four models	37
S2.1	Illustration of transforming vdw equation into our formula	41
S2.2	Similarity of training set peptides and test set proteins	45
S2.3	Comparison of sorted frame and all frame pairs	50
S2.4	Atomic SASA estimation for test set	51
S2.5	Accumulative error in summing up atomic SASA for test set	52
S2.6	Linear regression of fitting max_SASA to adjusted_max_SASA	52
S2.7	LCPO approximation of molecular SASA for test set	53
S2.8	The decrease of fraction of unfolded for HP36 observed using LCPO and our GB/SA method	53
3.1	Two runs of short MD simulations for HiQ27 proteins using ff14SB and TIP3P at 290K	62
3.2	Two runs of short MD simulations for HiQ27 proteins using ff14SBonlysc and GBNeck2 at 290K	63
3.3	Two runs of short MD simulations for HiQ27 proteins using ff14SB and GB-Neck2 at 290K	64
3.4	Comparison of the stability ratios across 4 secondary structure basin dihedrals for 19 amino acids	66
3.5	Comparison of ff14SBonlysc and ff14SB backbone dihedral energy profiles	67
3.6	Comparison of amino acid specific helical propensities predicted from three model combinations	68

4.1	Refinement of GB-folded BBA structures	81
4.2	Refinement and native runs for CASP11 targets in TIP3P solvent	82
4.3	Refinement and native runs for CASP11 targets in GB solvent	83
4.4	Comparison of secondary structure elements in crystal structure and simulations	84
4.5	Refinement of CASP11 targets referring to the top cluster	88
4.6	TR829 refined from GB simulations	89
4.7	Other refinement examples from GB and TIP3P MD simulations	90
4.8	Refinement examples from REMD simulations	92
S4.1	Refinement of GB-folded HP36 structures	96
S4.2	Refinement of GB-folded proteinB structures	96
5.1	The proposed fibril formation mechanism	100
5.2	Schematic diagrams of how α -helical intermediates might promote amyloid formation	101
5.3	Fibril formation rates of IAPP and its variant from experiments	103
5.4	Four proposed dimer models for IAPP	105
5.5	Sequences of IAPP and its variants	106
5.6	Schematic helix length visualization	107
5.7	Hypothesized scheme of precursor and control groups of dimer models	109
5.8	Secondary structure fractions for different force field and model combinations	112
5.9	Distribution of helix lengths of WT IAPP in simulations	113
5.10	Top populated structures in IAPP simulations	114
5.11	α -helical contents for IAPP and its variants predicted in simulations	115
5.12	The spatial positions of N-terminal helices on the α -helical dimer.	117
5.13	Secondary structure fractions for the three systems of hIAPP in α -helical dimer simulations.	118
5.14	Interaction energy in α -helical dimers decomposed to pairwise residues	119
5.15	Comparison of structural features between precursor and the control groups .	121
S5.1	The structures of 7 monomers selected as representative α -helical N-terminus with extended C-terminus monomers ready for docking.	123
S5.2	The precursor group: 16 dimer structures docked from representative α -helical N-terminus with extended C-terminus monomers. Dimer 11 and Dimer 12 were triplicated due to their large similarity with the hypothesized dimer in Figure 5.4D)	124
S5.3	The control group: 28 dimer structures docked from representative α -helical N-terminus with extended C-terminus monomers.	125

List of Tables

S2.1	Defined 30 SASA types and their occurrences in the training and test set . . .	42
S2.2	Optimized parameters for 30 SASA types	44
S2.3	Summary of training set: 10 scrambled peptides	46
S2.4	Temperature ladders for all REMD simulations.	47
S2.5	Comparison of structural equilibria of HC16 in GB, GB/SA and TIP3P models	48
S2.6	Cluster analysis for HP36 combined trajectory at 250 K and occurrences of the top 7 cluster representative structures in the four 300 K trajectories, respectively	49
3.1	Experimental structures selected into HiQ27 data set	56
S3.1	Structural details for HiQ27 proteins	71
4.1	GB folded targets and their structures	77
4.2	CASPR11 targets and the template model quality	77
4.3	Refinement of CASP11 targets referring to the best structure	86
S4.1	All provided CASP11 targets in refinement category	93
S4.2	Temperature ladders for the CASP11 REMD simulations	95
S4.3	Comparison of secondary structure percentages in native and simulated struc- tures	97

List of Abbreviations

Å	Ångström (10^{-10} meter)
CASP	Critical Assessment of Protein Structure Prediction
DSSP	Dictionary of Secondary Structure Prediction
ff14SB	Force Field 14 Stony Brook
ff14SBonlysc	Force Field 14 Stony Brook with only side chain modifications of ff14SB
fs	femtosecond (10^{-15} second)
GB	Generalized Born
GBNeck2	Generalized Born with Neck correction version 2
GB/SA	Generalized Born/Surface Area
GPU	Graphics Processing Unit
HiQ27	H igh Q uality protein data set of 27
ICOSA	IC OSAhedron/numerical way of SASA calculation
IDP	Intrinsically D isordered P rotein
IAPP	I slet A myloid P olypeptide
K	Kelvin
LCPO	Linear Combinations of P airwise O verlaps
MD	Molecular D ynamics
MM	Molecular M echanics
μ s	microsecond (10^{-6} second)
ns	nanosecond (10^{-9} second)
NMR	Nuclear M agnetic R esonance
PDB	P rotein D ata B ank
PMF	P otential of M ean F orce
ppII	poly-proline II
QM	Q uantum M echanics
REMD	R eplica E xchange M olecular D ynamics
RMSD	R oot M ean S quare D eviations
SASA	S olvent A ccessible S urface A rea
SSE	S econdary S tructure E lement
T2D	T ype 2 D iabetes
TIP3P	T ransferable I ntermolecular P otential 3 P oints
vdw	van der W aals

Acknowledgements

It is an exceptional and unforgettable journey pursuing a PhD in **Laufer Center** at Stony Brook. Throughout the years, I appreciate it as a place of sunshine and warmth, conversations and freedom. To me it is very much like *the Flowers and Fruit Mountain*¹ existing outside of the book *Journey to the West*². My thanks and salutes first belong to all the significant people who keep this interdisciplinary center coming along: Mr. and Mrs. Laufer for supporting financially, Dr. Ken Dill and Dr. Carlos Simmerling for directing all the postdocs, graduate and undergraduate students, all the faculties for bringing us the cutting-edge seminars and speakers, and the staff Nancy Rohring, Eileen Dowd, Laura Lamonica, Dr. Feng Zhang for maintaining the research environment from different aspects.

I must thank my advisor Professor **Carlos Simmerling** for his support and guidance over the years. I really appreciate and learn a great deal from his expertise in molecular modeling and effective science communication, his persistence in producing valued work, and his ability to identify the key points and new directions when my projects were in exploration state or did not work out as expected. I am very grateful that he acknowledges my hard work by quietly upgrading my equipment, and more importantly, reassures me in struggle by offering help and allowing me time.

I want to thank my committee chair, Professor **Daniel Raleigh** for providing me opportunities to collaborate with and present to his lab. His great interests in my modeling results drove my research progress in the first project. I thank my third inside member of committee, Professor **Ken Dill** for his valued input and open questions in the committee meetings, and his encouragement through his life story telling, in which he tells that research is among the easiest thing after working on a ranch. I also thank Professor **Dmytro Kozakov** for agreeing to be the external member of my committee.

And the amazing **Simmerlingons**, they facilitate my computations and inform me about important things in doing research. I thank the Hardware/Software Crew: Dr. James Maier, Dr. Kevin Hauser, Koushik Kasavajhala, Kenneth Lam and Kellon Belfon for all their time and efforts paid out of their pocket in building and/or maintaining the Bell and Naga Cluster (previously Tabasco and Habanero Cluster). I thank the seniors: Dr. Hai Nguyen for being my close friend and taking me on board of the GB folding project; Dr. Kevin Hauser for helping me explore protein refinement strategies and suggesting great articles/books/advice;

¹Also known as Mount Huaguo (花果山), a paradise for monkeys because of its scenic view and rich resources.

²Written by Chinese writer Wu Cheng'en (吴承恩) and published in the 16th century during the Ming dynasty, it (西游记) tells a mythological story of a team obtaining sacred book from the west of Ancient China and returning after many trials and some suffering. It is recognized as one of the Four Great Classical Novels of Chinese literature.

Dr. Cheng-Tsung (Eric) Lai, Dr. Haoquan Li, Dr. James Maier, Dr. Yi (Miranda) Shang, Dr. Carmenza Martinez and Dr. Amber Carr for the valued conversations; I also thank other alumni of Simmerling Lab for the conversations in conference or correspondence opportunities. I especially thank my peers Kenneth Lam and Koushik Kasavajahala for always standing side-by-side with me, even when they know much more, learn a lot faster and contribute a greater deal in the teamwork; it is from them that I see how to undertake daunting tasks and resolve issues with unstoppable momentum. I thank Chuan Tian for the great conversations in untangling the force field puzzles; thanks to Dr. Angela Miguez, Kellon Belfon, Junjie Zou and Zackary Fallon for the discussions and input in the lab work. I thank all of them, for sharing their insight and foresight with me in the down-to-detailed technical explanations and up to career development.

Next I must thank my **family** and my **enlighteners**. I thank my parents, my cousins Dr. Yun Huang, Xu Huang, my auntie Dr. Tiange Shao and other relatives for their encouragements and support for my academic studies. I thank my childhood friend Sai Luo for staying an enthusiast about nature and science herself, and also fueling my curiosity and longing for scientific discoveries. I always feel lucky and grateful for my undergraduate advisor Dr. Xiaojun Yao at Lanzhou University who introduced me to the field of molecular modeling and scientific computing, for Dr. Weiwei Xue who demonstrated a diligent and fruitful researcher, and last but not least, Dr. Qifeng Bai who mentioned to me that ff99SB was the state-of-the-art force field in protein modeling while compiling Amber in 2010, when I had no idea about them at all, which coincidentally led me all the way to where I am now.

I greatly appreciate the friendships and interactions with some of the best people I know outside of the Lab. I thank Diana Melick and Kevin Brady for hosting me in their family and being my American life coaches, thank Bella Brady, Delia Brady and all the Mandarin-/Spanish-speaking "au pairs", especially Yujue Wang, being my sister-like families. I thank Lu Bai, Dr. Adam de Graff, Jiaye Guo, Ruoqian Lin, Dr. Qian Li, Cong Liu, Dr. Alberto Pérez, Yue Shi, Dr. Tamás Székely, Bihua Yu, Dr. Yiman Zhang, Mengru Zhang, Xiaoxue Zhang for being my close friends at Stony Brook. I also would like to thank the people who enhanced my experience and directed more possibilities: Xindi Li for starting the coding club and Tuoling Qiu, Cong Liu, Jiaye Guo, Yue Wang, Mengru Zhang solving problems together; Dr. Sergy Ovchinnikov for valued conversations about proteins and evolution; Jiaye Guo for being my Japanese learning partner, Mengru Zhang and Dr. Courtney Singleton for encouraging and joining us; Diane Papuzynski for being my skating coach and the Laufer Gang for inspiring activities etc. Lastly, I thank the two CAPS counsellors for their professional help and the SciFri (and many other) podcast host(s) for their spiritual support.

And finally, I thank my husband Shuai Liu for everything, especially for supporting me doing things even if he hasn't seen the value in them.

Chapter 1

Introduction

Living and non-living things in the universe are made of elements, atoms and molecules. Proteins are biological molecules that play essential roles in ubiquitous processes of life. Plants use proteins to harvest energy from the sun, to be more specific, proteins embedded in cell membranes gather light for photosynthesis; both plants and animals digest food by breaking bigger nutrient molecules into smaller pieces using enzymes and transport them across membranes or in bloodstream, which power up other processes of the cell; all level life forms carry out autonomous regulations using hormones and other signaling receptors, attack as or defend against pathogens using antibodies; other important roles are played by proteins, so as to support or move cells, build other proteins or read genetic instructions etc.[1]. All in all, so many processes that proteins are engaging have been uncovered, although all amazing things proteins could do are far less than being clearly understood.

How do we know so much about proteins? Because researchers around the world are studying these molecules at various levels from numerous perspectives. Proteins are related to the origin, survival and well-being of human beings including human researchers. That is why there have been many studies carried out to understand the structure, function and evolution of proteins, from as small as the atomic level, to molecular, cellular[2], and individual[3] level (human proteome), all the way up to as big as ecological level[4]. Scientists and researchers observe phenomena, ponder on the causes and set up model systems trying to reproduce the observations and distill the principles.

Interrogations and implications around proteins probably is indeed endless. The question of "how proteins fold to a specific 3D structure only from sequence information" puzzles generations of scientists. Proteins encode the mysterious force of nature as molecular machines, given they are originally just "ATGC..." DNA sequences wrapped in genomes and later get released along with "AUGC..." RNA molecules. Proteins have been discovered to be so relevant to human health and disease that the pharmaceutical companies produce drugs to target some key interactions between proteins and their biological counterparts, such as RNA, DNA, or other proteins. And a major part of proteins don't adopt deterministic structures or only do in some particular circumstances, called intrinsically disordered and folding upon binding. There are tremendous question marks and potential applications going after this central molecule, PROTEIN, or better described as a class of molecules made of basic units called **amino acids**.

This thesis describes a tiny portion of these endeavors directed at the atomic level and

structure prediction perspective. Thanks to the inventions from ancestral and contemporary researchers, efforts were able to be made into theoretically predicting the structures and structural ensembles of proteins. These proteins have been studied *in vitro* under experimental conditions mimicking physiological environment. In this thesis, these proteins are studied *in silico* simulating the experimental conditions, where proteins were observed and measured in their natural and denatured states.

1.1 Four levels of protein structures

There are 20 naturally occurring amino acids, which concatenate to polypeptides and proteins through condensation reactions forming peptide bonds. There are also other post-translational modifications, such as disulfide bond or other rarer types of covalent bond formation. All 20 amino acids share the same generic structure given in **Figure 1.1A**, where the amino group, carboxylic acid group and hydrogen are bound to the central carbon atom (called alpha carbon, C_α). The charge state of amino group to be +1 charge and carboxylic acid group to be -1 charge shown in **Figure 1.1A** is often found at neutral pH, called free termini (terminus); depending on the physiological conditions, experimentalists synthesize and computationalists simulate proteins with terminal cap residues. The R group (also known as the side chain) differentiate 20 amino acids in terms of size, polarity, charge, hydrophobicity etc. The sequence of amino acids of each polypeptide chain of which the protein is composed determines the **primary structure** of a protein.

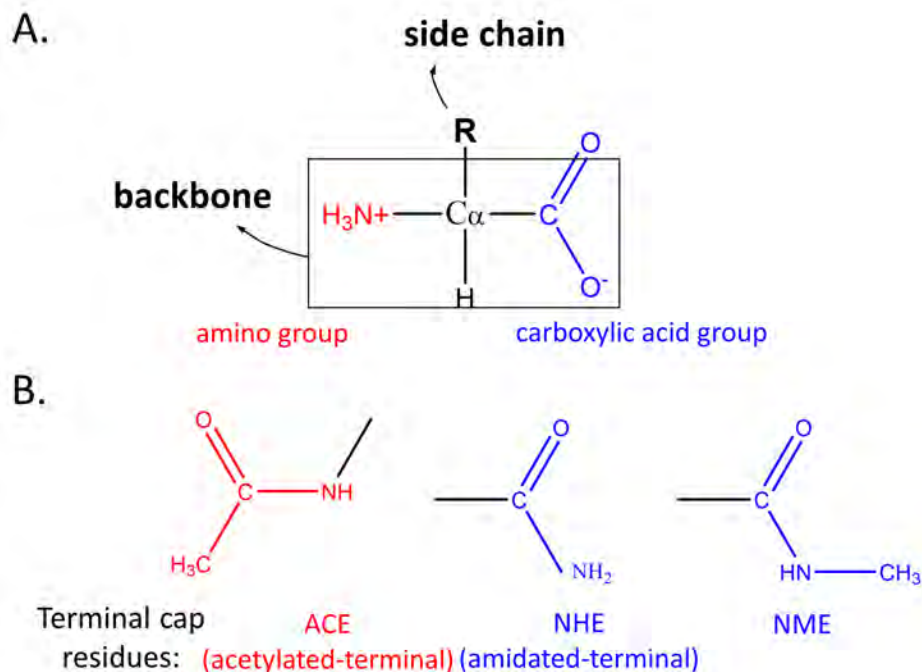


Figure 1.1: A. Generic formula of amino acid; B. Three types of cap residues in Amber, N-terminal cap in red and C-terminal caps in blue.

Delocalized electrons within the peptide bond result in a phenomenon called resonance, which gives peptide bond a partial double bond character and constrained to be planar without rotatable freedom (**Figure 1.2A**). Therefore, the only source of conformational freedom of polypeptide/protein backbone comes from the torsional rotation around the N–C α and C α –C single bonds. These two bonds are respectively designated as torsion angle ϕ and ψ . Thus these two parameters describe the backbone flexibility and conformational preferences of each amino acid. And they are the backbone dihedral parameters mentioned in the later contexts. The values of these two dihedral angles are represented on a ϕ vs. ψ 2D plot, called Ramachandran plot[5].

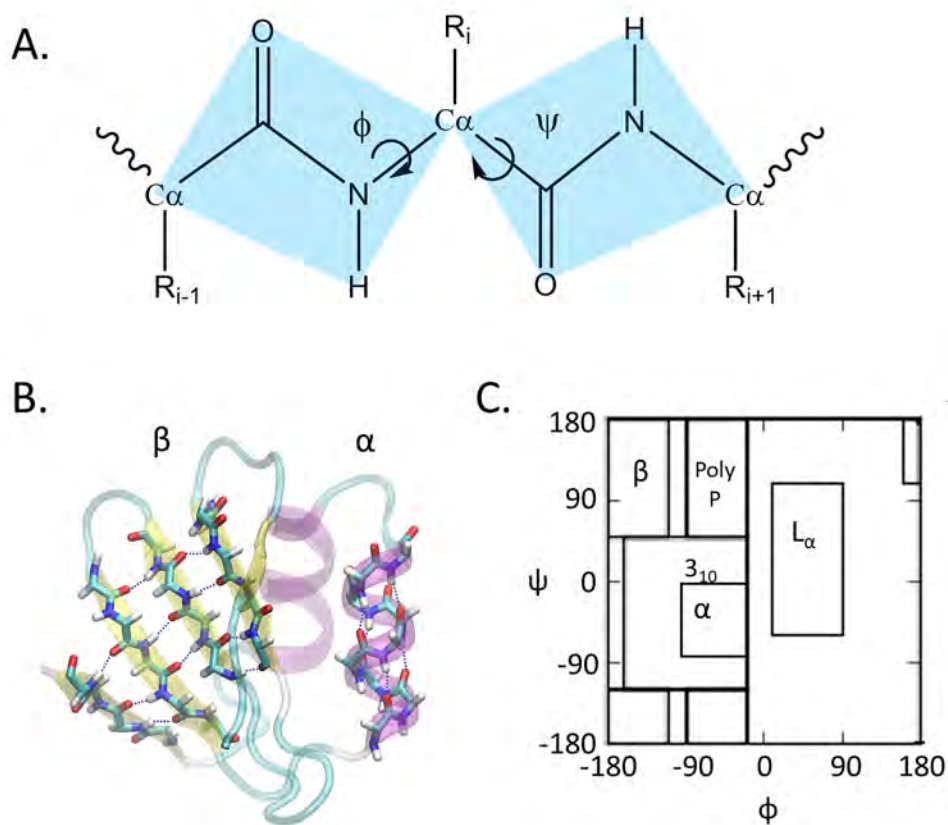


Figure 1.2: A. The torsion angles ϕ and ψ illustrated on a polypeptide fragment centering on the amino acid with side chain named R_i . The planes formed by peptide groups are in transparent light blue; B. α -helices (magenta) and β -sheets (yellow) highlighted in a protein structure; C. Definition of different secondary structural basins determined from backbone dihedral angles ϕ and ψ .

As a result of steric hindrance, not all combinations of ϕ and ψ are stereochemically feasible. But there are still patterns appearing in limited ϕ/ψ space, to be more specific, when certain allowed ϕ/ψ torsion combination is adopted and propagated consecutively along the polypeptide chain, some special 3D arrangements will appear as re-occurred patterns, called **secondary structures** of a protein. The two most common secondary structural elements

are α -helices and β -sheets (**Figure 1.2B**, which is the crystal structure of a protein TR829 studied in **Chapter 4**, PDB code: 4rgi[6]). These secondary structures are held together by hydrogen bonding between amide hydrogen and carbonyl oxygen. Other protein secondary structures including ppII, left-handed α helix and 3_{10} helix are also important. Some analysis approaches are introduced in details in Section 1.2.4.

In globular proteins, **tertiary structure** comes into place as results of hydrophobic residues packing and side chain interdigitation, often with salt bridges or hydrogen bonding formation. The different "fold" of tertiary structures can be further hierarchically classified as different levels, for example in the CATH protein structure classification database[7], "class" as the highest level, followed by "architecture" without connectivity info, and "topology" as the finer description. **Figure 1.3** displays proteins of diverse topologies used for SASA estimation test in **Chapter 2**.

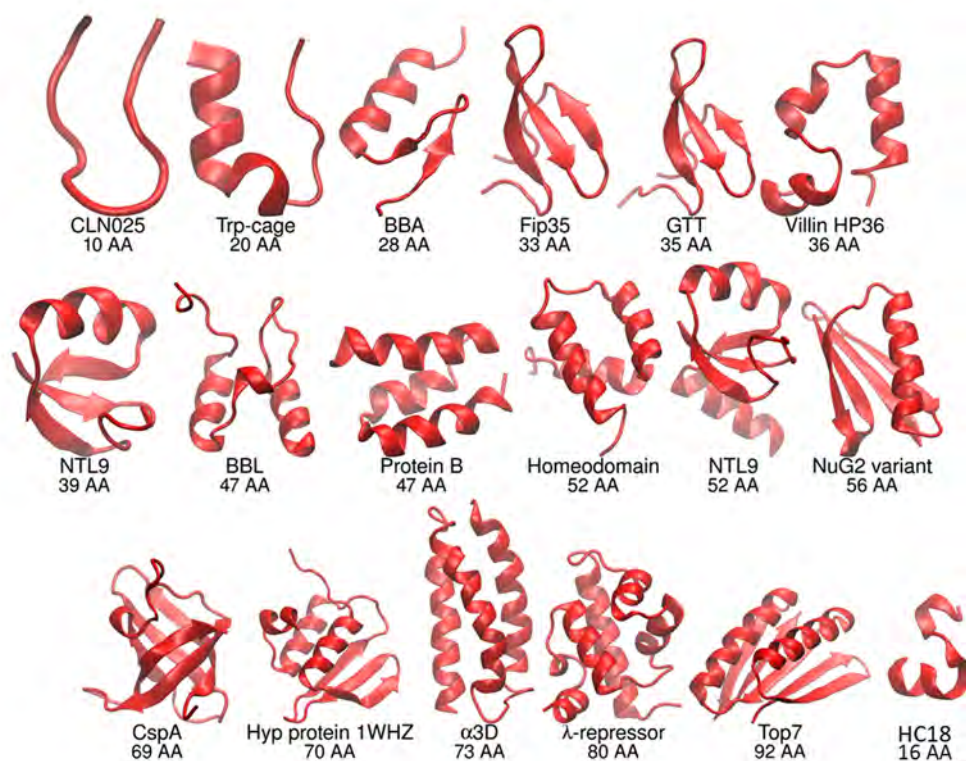


Figure 1.3: Protein data set of diverse topologies with designated name and size

Multiple chains of proteins form **quaternary structure** to carry out functions, for example HIV-protease performs cleavage of polypeptide as homodimer. However, oligomerization of amyloid-forming proteins are associated with pathogenesis in diseases such as Alzheimer's disease, Parkinson's disease, Huntington's disease and type II diabetes. If the mechanism of these protein oligomerization and pathogenesis could be understood atomically, modulations or interference could be carried out to prevent these quaternary structures formation. There is a long way to go at this point, but some efforts into the mechanism study of the initialization of amyloid fibrils have been paid and reported in **Chapter 5** of this thesis.

1.2 Protein modeling techniques

To seek answers for questions displayed at different levels of protein structures, we build models at atomic details in computer programs and strike to reproduce and predict the properties of protein molecules in their close-to-experimental conditions. These models are governed by Molecular Mechanics (MM). MM provides the underlying physics and defines the potential energy profiles differentiating various stability of protein conformations. For proteins to sample lower energy i.e. more stable conformations, Molecular Dynamics (MD) are used to calculate the interactive forces and simulate the motions. We carry out MD simulations and analyze trajectory data in Amber Software[8]. The analysis methods, such as RMSD calculations, Cluster Analysis, DSSP Analysis, are implemented in Amber Modules. For protein structure and trajectory visualizations, we mainly use VMD[9] software. In this following section, relevant protein modeling tools utilized in this thesis are briefly introduced.

1.2.1 Molecular Mechanics and force field

To understand the motions and energies of atoms and molecules, Quantum Mechanics (QM) ultimately provides the underlying physics on the subatomic scale, which accurately describes the probabilities of electron densities and energy level changes caused by orbital effects. However, QM is too complex thus it is not applicable for understanding proteins. In contrast, Molecular Mechanics (MM) is at the appropriate scale and efficacy, where only nuclear positions with masses and charges are used to calculate the motions and energies. The hybrid QM/MM approach is available and better for chemical reactions, catalytic sites, and transition state calculations, but still not applicable for protein structure studies yet.

In general, the MM potential energy functions are composed of energy terms representing bond stretching, angle bending, torsional angles correction, and non-bonded interactions. This equation below is a widely used expression of potential energy function[10]:

$$E(R) = \frac{1}{2} \sum_{bonds} k_b(b - b_0)^2 + \frac{1}{2} \sum_{angles} k_\theta(\theta - \theta_0)^2 + \frac{1}{2} \sum_{dihedral} k_\phi(1 + \cos(n\phi - \delta)) + \sum_{non-bonded} \left(\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr} \right) \quad (1.1)$$

As a function of the coordinates (R) of all atoms in a protein, potential energy (E) is calculated as a sum of four terms; each term is a function of bond length (b), angles (θ), dihedral angles (ϕ) and distances between two particles (r), respectively. The two term of bond/angle energy are modeled by Hooke's Law as a harmonic oscillator. k_b/k_θ is the bond/angle force constant determining the strength of the bond/angle, b_0/θ_0 is the ideal bond length/angle, which are derived from infrared stretching frequency and high resolution crystal structure geometry[11]. The third term in the equation is used for dihedral rotations, which is described as periodic and wave-like thus is modeled by a sum of cosine functions. k_ϕ , n and δ are barrier height of dihedral rotation, multiplicity and phase, respectively. This term has no physical inferences but is used to correct the discrepancy between QM and MM. Therefore, this dihedral term is usually fit against QM after all other energy terms are added up in place. The fourth term describes the non-bonded interactions, which are

made of van der Waals interactions and electrostatic contributions. Van der Waals energy consists of repulsive term and attractive term of dispersion energy, combined to be known as Lennard-Jones 6-12 potentials, has a minimum at the distance of the sum of the van der Waals radii of this certain atom pair; parameters A and B are determined by non-bonding distances in crystals, gas-phase scattering measurements[10] and organic liquids simulated by optimized potentials for liquid simulations (OPLS)[12]. The electrostatic interactions between two atoms use a Coulomb potential. D the dielectric function for the medium and r the distance between the two charges. There are no explicit hydrogen bond interactions, which is assumed to be taken care of by the electrostatic attraction when the donor and acceptor atoms possess opposite charges.

Different parameters can be determined for this equation by different systems and models, which result in various **force fields**. Force field in molecular modeling refers to the MM potential energy function and the parameter sets. Protein force fields used today date back to the late 1980s[13, 14, 12], and took generations of development and modifications. In one of the most widely used force fields, ff99SB[15], all the bond, angle and non-bonded terms (in **Equation 1.1**) are calculated in the same way as its ancestral force fields ff94[11] and ff99[16]. However, the torsion cosine parameters are refit against QM on the energy minimum of a larger set of Gly and Ala tetrapeptides conformations, instead of amino acid-like small molecules or limited set of Ala dipeptides. This refitting achieved a better secondary structure balance, thus it is widely used[17, 18, 19] and it is also used as a basis model for backbone variation updates such as ff99SB*[20], ff99SBnmr[21] and side chain updates ff99SB.ILDN[22]. More recently, ff14SB[23] was published with an empirical correction on the backbone dihedral parameters to better match scalar coupling data and a refit against QM gas-phase energies for the side chain dihedral parameters. These side chain modifications were then combined with ff99SB backbone dihedral parameters, termed ff14SBonlysc[23], along with a Generalized Born (GB) solvent, to successfully fold a set of proteins (all except HC16 as shown in **Figure 1.3**) from only the sequence information[24].

1.2.2 Solvation and implicit solvent

Proteins are typically studied in an aqueous environment. Biophysical study of their properties and functions requires an accurate description of their solvation and desolvation processes, i.e. the binding and removal of water[25] or solvent. To study how proteins fold or bind, the solvation free energy changes (ΔG_{sol}) associated with solute-solvent interactions and water reassembly are essential. In biomolecular modeling, these water molecules can be represented explicitly or implicitly. Explicit solvent models, which compute all the pairwise interactions over all solute and solvent atoms and are thus more detailed and complete in theory, however, are limited in usage, as water atoms dominate the calculations and friction slows the sampling of large conformational changes[26]. As an attractive alternative, implicit solvent models possess high efficiency in sampling, which has promoted their wide application in protein folding[24, 27], structure prediction[28], protein design[29] and refinement[30].

Implicit solvent model efficiently describes the electrostatics of molecules in water environment. It represents the solvent implicitly as continuum with the dielectric properties of water, and also includes the charge screening effects of salt. Some attractive features[8] of implicit solvent models are (1) the computation of implicit solvation is generally cheaper and

scale better for parallel computing than explicit solvent; (2) due to the absence of viscosity in explicit solvent, the molecule can explore through conformational space much faster; (3) Periodic Boundary Conditions and Particle Mesh Ewald summation are typically used to speed up explicit water calculations, but they might result in interacting artifacts. Implicit solvent with infinite volume does not need them thus the artifacts could be avoided.

1.2.3 Molecular Dynamics and Amber software

Given a potential energy function and a parameter set, there are various approaches to study the dynamics of proteins as well. One is molecular dynamics simulations in which Newton’s equation of motion are solved for atoms in the system[10].

Molecular dynamics is a computational method which simulates the motions of particles by calculating their interaction forces and potential[10]. This method has been applied to systems as small as an atom and as large as a galaxy. In microsystems, the complex forces between atoms and molecules are considered; between stars the simple gravitational interactions are calculated. The essential rule that these calculations obey is the classical Newton’s law of motion:

$$F = m \times a \tag{1.2}$$

At each time point, the force that acts on each particle of interest could be calculated. Integration of a series of time steps generates a trajectory that describes the positions and velocities of all the particles; the potential energy and kinetic energy are also computed accordingly.

Beginning 40 years ago[31], molecular dynamics simulations of proteins have been widely used as very powerful tools to study the structure and dynamics of peptides and protein molecules. The applications are ranged from, but not limited to, the energetic calculations of ligand binding and enzyme reaction mechanisms, protein folding and refolding, analysis of experimental data and refinement of structures. The widely used software “Amber” provides a framework for a suite of programs that allows users to carry out and analyze molecular dynamics simulations for proteins, nucleic acids, carbohydrates and lipids[8]. The term Amber also refers to the empirical force fields. In Amber, the simulation module and force fields are separated; other computer packages or platforms (Gromacs, Charmm, NAMD, OpenMM etc.) have implemented Amber force fields, and other force fields can be used within the Amber programs[32].

MD as a general sampling method is not efficient enough because the rough energy landscape of proteins makes the simulations get trapped easily in local minimum-energy states. Therefore, when conformational sampling is beyond straightforward MD simulations, we employ enhanced sampling method such as REMD. First introduced to work with Monte Carlo algorithm[33, 34, 35], Sugita and Okamoto[36] developed an implementation with MD, named replica exchange molecular dynamics method (REMD). In REMD, multiple copies (or replicas) of the same system are simulated simultaneously and independently at a ladder of temperatures. They exchange temperatures with neighboring replica based on potential energy overlap under Metropolis criterion. REMD has been shown to be more efficient than MD[37] and the convergence could be further improved by coupling to a reservoir of high-temperature generated structures[38].

In MD/REMD simulations carried out in this thesis, some more techniques employed are briefly introduced here. To reduce the degrees of freedom thus simplify force calculations, SHAKE algorithm, implicit solvent and periodic boundaries for explicit solvent are applied. SHAKE[39] imposes constraints on the bond lengths of the bonds involving hydrogen atoms, which average out the highest frequency vibrations. Implicit solvent and periodic boundaries have been introduced in previous section 1.2.2. To simulate proteins at constant temperatures, the solute and solvent system is coupled to an external heat bath that is fixed at the desired temperature, which acts as a source of thermal energy supply. The velocities are also scaled at each temperature in both MD and REMD. Langevin dynamics works as a stochastic heat bath, which approximates the frictional effect (frictional drag on the solute and random collisions associated with solvent thermal motions) of solvent molecules. A collision frequency associated with Langevin thermostat is used to control the magnitude of the frictional force and the variance of the random forces. The Hmass re-partition technique allows larger time step of integration, in which the mass of heavy atoms is re-partitioned onto the bonded hydrogen atoms. By doing this to slow the highest-frequency motions contributed from hydrogen atoms and to avoid numerical instability, the time step of simulations could be increased from the conventionally used 2 femtosecond (fs) to 4 fs[40].

1.2.4 Secondary structure analysis and clustering algorithm

In analysis carried out in this thesis, two different approaches are used to assign and characterize secondary structures in a protein: 1) ϕ/ψ dihedral angles represented on Ramachandran plot and 2) Dictionary of protein secondary structure prediction (DSSP). For the dihedral angles, we define secondary structure basins illustrated in **Figure 1.2C**, which is consistent with the previous studies[19, 23]. The definitions of the four secondary structure basins are as follows: right handed helix (α), (ϕ, ψ), (-160° to -20° , -120° to 50°); left handed helix (α_L), (3° to 90° , -60° , 110°); extended β -strand conformation, (-180° to -110° , 50° to 240° ; or 160° to 180° , 110° to 180°); and ppII, (-90° to -20° , 50° to 240°). Note that a more stringent (-100° to -30° , -67° to 7°) α -helix-forming basin is used for helical propensity calculations in **Chapter 3**. DSSP was developed by Kabsch *et al.*[41] as a program that recognizes hydrogen bonding and geometrical features for secondary structure assignment[41]. The types of secondary structure include α -helix, 3_{10} -helix, π -helix, parallel β -sheet, anti-parallel β -sheet, turn and bend. Bend is quantified to form if five $C\alpha$ ($i-2, i-1, i, i+1, i+2$) position a curvature angle larger than 70° . If a hydrogen bond is formed (with up to 4.1 \AA N-O distance and 60° misalignment angle), this bend is a turn. 4 consecutive turns of $i \leftarrow i+4$ hydrogen bonding makes a α -helix, 3 turns of $i \leftarrow i+3$ is 3_{10} -helix and 5 turns of $i \leftarrow i+5$ is π -helix. Helices are locally formed while β -sheets are more global. β -sheets can be formed in parallel or anti-parallel. This algorithm has been implemented in Amber cpptraj module[42].

Clustering is a technique used in many fields to dissect data into discrete sets based on similarities. In protein structure cluster analysis, pairwise RMSD values are evaluated and the geometric structures are separately by defining certain criterion, such as distances between clusters. In the cluster analysis carried out in this thesis, the top-down divisive (hierarchical) clustering algorithm has been used. Shao *et al.*'s work[43] has recommended that this algorithm performs the best if the cluster count is not defined prior. Benchmark analysis

has also been done and it is found that an epsilon value of 2.0 Å works the most efficiently at generating distinct clusters. The meaning of this epsilon value could be explained as follows, as larger clusters that originally have close distances further divide, the minimum distance between clusters is increasing; once the closest distance exceeds 2.0 Å, the clustering process finishes. If epsilon value is < 2.0 , there will be more cluster numbers of smaller sizes, which split the already very similar topologies even further; if epsilon value is > 2.0 , structures within each cluster are not similar enough. The distance between clusters is measured as the average distance between all members of two clusters. The representative structure (cluster centroid) is picked if this structure has the lowest cumulative distance to every other point in this cluster. This analysis tool had been implemented in Amber cpptraj module[42].

1.3 Improvements needed in current modeling tools

Essential compositions of protein modeling and simulations must include force field, solvent model, sampling method and implementation into computer programs. Aside from applying the current modeling tools to studying protein structures at different levels (**Chapter 4** and **Chapter 5**), investigations were done to explore and improve the implicit solvent models (**Chapter 2**) and Amber force field evaluations (**Chapter 3**).

1.3.1 Nonpolar solvation and GPU-parallelization

In implicit solvent models, the solvation process is put in the context of a thermodynamic cycle[44] (**Figure 1.4**), first solvating the uncharged solute by creating and accommodating a cavity (nonpolar term, ΔG_{np}) and then turning the charges back on by modeling water as a continuum high dielectric (polar term, ΔG_{pol}). The polar, or electrostatic part, is typically modeled with Poisson-Boltzmann (PB)[45] or Generalized Born (GB)[46] equations. The nonpolar part is often further decomposed into cavity (ΔG_{cav}) and van der Waals (ΔG_{vdw} contributions)[47]. The cavity term tends to be unfavorable, while the van der Waals interaction with solvent is typically favorable, thus some cancellation between these contributions gives rise to the overall ΔG_{np} . Both ΔG_{cav} and ΔG_{vdw} are thought to be proportional to the average number of waters making direct contact with solute (i.e. first solvation shell approximation)[48]. Thus the nonpolar term is often estimated by a SASA-based method[46], although it has been pointed out that SASA is not accurately proportional to solvation energies for small alkane solutes[49, 50], and the volume term may be more important[50, 51]. While SASA-based implicit solvent incorrectly predicted association stabilities of small molecule amino acid analogues when compared to explicit solvent results[48, 52], SASA-based nonpolar solvation has been shown to be useful for accurate prediction of native-like protein conformations[53] and protein-ligand binding affinities[54, 55] such as in MM/PBSA and MM/GBSA.

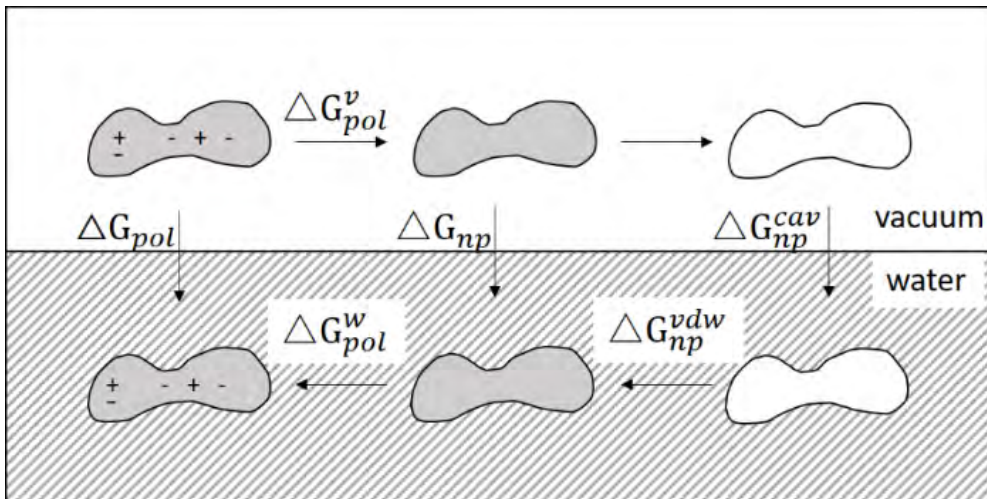


Figure 1.4: Thermodynamic cycle of the solvation process. Solvation free energy (ΔG_{sol}) is decomposed into polar (ΔG_{pol}) and nonpolar (ΔG_{np}) contributions. The steps involve uncharging the solute in vacuum (ΔG_{np}^v), removing the solute-solvent interaction in vacuum (no free energy change), creating a solute cavity (ΔG_{np}^{cav}), establishing uncharged solute-solvent interaction in solvent (ΔG_{np}^{vdw}), and charging the solute in solvent (ΔG_{pol}^w). The figure is adapted from Levy *et al.*[56]

Although much recent effort has been devoted to improving the polar solvation contribution, less attention has been paid to the nonpolar solvation term. This is likely because of its small magnitude relative to the polar part, questionable accuracy of simple nonpolar models, and significant computational cost. Its two sub-terms are of opposite signs in free energy change, thus this term is often treated as negligible; cavity creation loses entropy, while formation of attractive solute-solvent interaction gains enthalpy[49]. Compared to a solvation energy of -5.0 kcal/mol for a polar molecule, this number is only 1.8 kcal/mol for a nonpolar molecule of similar composition (e.g. ethanol vs. ethane)[57]. In some other reported literature, even if nonpolar contributions were considered, the implicit solvent accuracy was not improved with respect to experimental or explicit solvent results[58]. Even with demonstrated optimizations[50, 56, 59, 60], the cavity sub-term, particularly the SASA, remains a major resource-demanding calculation. Moreover, in contrast to the fact that all the other energy terms can be computed on GPUs in the most recent Amber implementation[8], the SASA-based nonpolar approaches can only be calculated on CPUs, producing a bottleneck that severely limits sampling in simulations.

In fact, to accelerate SASA calculation, we could refer to GB models and their efficient implementation on fast general purpose GPUs[61]. These GB models[62, 63, 64] are trained to reproduce the PB energies, along with the PB-based “perfect” effective radii[65], employing additive and pairwise analytical energy energies and their derivatives. This pairwise descreening algorithm[66] serves as an ideal platform for GPU parallelization[61]. When the same instruction is executed for every atom pair in the protein system, massively efficient GPU cores can compute the desired values simultaneously. Compared to parallel performance of CPU implementation with all double precision calculations, a single GPU using the mixed precision model[67] can achieve a factor of 2-5 speed up compared to massive CPU

cores, as CPU scaling plateaus long before it reaches the GPU performance[61].

Inspired by pairwise descreening algorithm used in GB models, a pairwise and GPU-friendly SASA estimation approach is developed to fill the gap in the current Amber implementation and accelerated nonpolar solvation. An efficient pairwise GPU-parallelizable algorithm requires the same instructions to be executed for every neighbor atom j of i indistinguishably. In the current implementation of Linear Combinations of Pairwise Overlaps (LCPO) algorithm, the SASA of a central atom i is dependent on not only the neighbors of i , but also the overlaps of the neighbors with each other. For example, in **Chapter 2 Figure 2.1A**, atom j_1 and j_2 are both neighbors of atom i . In determination of the SASA for atom i , not only atom pair (i, j_1) and (i, j_2) are involved, but atom pair (j_1, j_2) also contributes. This extra consideration makes LCPO a many-body algorithm and not as suitable for GPU parallelization. Therefore, even if GPU devices are employed for GB/(LCPO)SA simulations, the SA portion becomes a major bottleneck. Therefore, we proposed a new approach, where the same computation (see **Chapter 2 Equation 2.5**) is used for two atom pairs with different corresponding parameters, thus it is an ideal fit for GPU parallelization.

1.3.2 Secondary structure balance and amino acid backbone specificity

Understanding protein structures at each level are related to miscellaneous implications. Knowledge about secondary structure is important for the understanding of many biomolecular questions, such as protein folding, aggregation, protein-ligand interaction and conformational changes[20]. In the modeling and prediction of protein structural ensembles, accurate secondary structure balance is one of the key factors to achieve a consistent outcome with experimental measurements. Stability of different secondary structures could dictate the relative potential energy levels of certain tertiary structures; if certain secondary structure is destabilized or over-stabilized, the overall topology which is poor or rich in that secondary structure, respectively, would have larger statistical weight. As a result, certain tertiary structure would be of higher population unphysically, due to it is falsely favored by the computational model. In the modeling of disordered proteins or proteins in unfolded states, the secondary structure balance problem is exemplified and even magnified, as a more rugged energy landscape results in a much higher sensitivity.

In force field development, bond, angle and non-bonded terms listed in **Equation 1.1** are first fit against QM and experimental measurements, dihedral term is usually the last step to compensate the missing QM effects. In the case of ff14SB[23], backbone and side chain dihedral parameters are trained separately. Both dihedral parameters have been shown to shift secondary structure balances. ff99SB*[20] and ff99SBnmr[21] tweak backbone parameters for a better global balance between α -helical and β -strand basins. Backbone parameters were modified to improve the sampling of IDPs[68]. ff99SB-ILDN[22] and the more thorough ff14SB[23] improve side chain dihedral parameters and achieve better secondary structure balance.

So far, it is still challenging for us to reproduce the amino acid specific secondary structure balances. For one thing, the prediction of mutation effects is not always reliable. In a commonly studied variant of Trp-cage tc5b, when the first four residues NIYL are mutated

to DAYA in tc10b, simulated stability of tc10b (Koushik Kasavajhala and Carlos Simmerling *et al.*'s unpublished work) are not consistent with experimental measurements[69] nor theoretical predictions[70]. As observed in the work of Perez *et al.*[71], Alanine is predicted of a much lower helical propensity while experimentally it should be among the highest. Also the stability of secondary structures are observed to be sequence-dependent[72]: ff99SB underestimates α -helicity, whereas it accurately reproduces the melting temperature of Trp-cage. ff99SB*-ILDN cannot stabilize some beta-hairpin peptides. Charmm22* reproduces α -helicity in Ala-based peptides but cannot stabilize Trp-cage.

One hypothesis is that backbone parameters are not specific enough for different amino acids, given that in ff14SB and its ancestral force fields, except Gly and Pro, all the other amino acids share the same backbone parameters. Even though it has been shown that improvement in side chain parameters help with backbone specificity[23], whether current computational models are able to capture the amino acid backbone specificity awaits more investigations.

In principle, backbone specificity should come from side chain, but to capture them *in silico*, empirical corrections are needed because (1) fixed charge model has limitations; different side chains should induce charge redistribution on the backbone, but they are all the same except charged and Pro residues from ff94. (2) Variations in structure and energy arise from the complex interplay between torsional and non-bonded interactions[73], while our current terms are still clearly suffering from issues such as lack of ϕ/ψ coupling terms, short-range non-bonded problems etc. Empirical modifications have been shown as useful solutions, such as CMAP (correction map to compensate all the surface difference from MM to QM), used in Charmm force fields[74], and residue-specific force fields[72], considering the intrinsic conformational preferences of amino acid residues[75].

Before modifications can be applied to the current dihedral parameters, we need to build robust test sets and analysis tools to first evaluate the backbone specificity and accuracy in the current models. Therefore in **Chapter 3**, two test tool boxes are protocolized and demonstrated for backbone parameter evaluations.

1.4 Overview of the studied questions

In this thesis, all four chapters are displayed in reverse chronological order as the projects were developed and carried out:

I started with the project described in **Chapter 5** first as a rotation student in Dr. Simmerling's Laboratory, where monomeric and dimer models were built and analyzed for islet amyloid polypeptide. As (1) the structural equilibria are sensitive to force field nuances for intrinsically disordered protein, and (2) the experimental data used for comparison is at much longer time scale with respect to the simulations, it is challenging to validate the simulation predictions, thus we ascribed the inconclusive findings to a not-ready computational model and over-simplification in building dimer models.

I then participated in a project initiated by a former graduate student, where we folded a set of proteins using GB solvent and GPU. That project became a game-changing point, which leads to more wide testing of the same force field and solvent model combination on a even larger data set: CASP11 refinement data set (**Chapter 4**). The fact that none of

the helix bundle structures are stabilized in our simulations suggested that our computational model is specially challenged by proteins abundant in helical structures. There are two aspects of this instability issue: (1) it could be the local helices that are not stable and transformed to other secondary structures; (2) it is the tertiary structure of relative arrangement between the different helix bundles that go through a thermal unfolding. The local helical propensities point to the issues in backbone parameters, while the more global thermal instability is ascribed to the lack of nonpolar term in implicit solvent.

Therefore, I started to work on evaluating the secondary structure preferences for different force field/solvent model combination (**Chapter 3**), meanwhile started to tackle on incorporating the nonpolar term onto GPU calculations (**Chapter 2**).

To summarize, in this thesis, four questions and the endeavors directed to answering each question are reported in the four chapters, respectively.

1. How important is nonpolar term of solvation to protein structure stability, while it had been previously thought as a negligible correction term? (**Chapter 2**)
2. Whether the current all-atom force field and implicit solvent can reproduce the amino acid backbone specificity, given all (except Glycine and Proline) the amino acids share the backbone parameters trained on Alanine peptides. (**Chapter 3**)
3. How well current Amber force field and implicit solvent perform in the CASP refinement, if unrestrained MD simulations are applied? (**Chapter 4**)
4. Whether MD simulations of low-order oligomers of amyloid-forming protein IAPP could shed light on its fibril initialization mechanism. (**Chapter 5**)

Chapter 2

GPU-accelerated Nonpolar Solvation of Proteins: Fast Calculation of Accessible Surface Area in Implicit Solvent Simulations

2.1 Abstract

We propose a pairwise and readily parallelizable SASA-based nonpolar solvation approach for protein simulations, inspired by our previous pairwise GB polar solvation model development. In this approach, we develop a novel function to estimate the atomic and molecular SASAs of proteins, which results in comparable accuracy as the non-pairwise LCPO algorithm in reproducing numerical icosahedral-based SASA values. Implemented in Amber software and tested on consumer GPUs, our method reasonably reproduces LCPO simulation results, but accelerates MD simulations over 30 times compared to LCPO implementation, which is greatly desirable for protein simulations facing sampling challenges. The value of incorporating the nonpolar term in implicit solvent simulations is demonstrated on a peptide fragment containing the hydrophobic core of HP36, and evaluating thermal stability profiles of four small proteins. ¹

2.2 Introduction

Our motivation to revisit the nonpolar solvation aspect arose from our recent study of protein folding simulations using only polar solvation[24]. Although we could sample folding for proteins up to nearly 100 amino acids in standard MD on GPUs using only the polar solvation term (GBNeck2[62]), we observed that the proteins tested in our folding studies[24] and Perez *et al.*'s structure predictions[28] suffered from poor stability compared to experiment[24, 28]. In some of the small proteins (CLN025, Trp-cage, Villin HP36 etc.),

¹This chapter is adapted from the manuscript submitted, titled "A Fast Pairwise Approximation of Solvent Accessible Surface Area for Implicit Solvent Simulations of Proteins on CPUs and GPUs", He Huang and Carlos Simmerling*

even though native conformations are predicted from only sequences to as close as 1 Å without prior knowledge, and correct trends in the melting behavior could be reproduced[24, 28], simulated melting temperatures (T_m s) were usually off by tens of Kelvins (see Results). We hypothesized that this instability might be a result of neglecting nonpolar solvation in our model. It was also suggested by Chen and Brooks[48] that a fine tuning non-polar solvation model might be helpful or sufficient for proteins such as HP36. Shell and Dill *et al.* suggested [76] that more studies are needed to “examine the effect of turning off the surface area component of the implicit models”. Here, we investigate and quantify the effect of nonpolar term on protein stability and conformational equilibria in MD. Moreover, we study the extent to which a simple SASA-based approach could improve reproduction of experimentally determined properties such as folding free energy.

In our opinion, an analytical, GPU-compatible nonpolar solvation energy term is needed before we can carry out thorough investigations on the impact of nonpolar term in MD of larger proteins. Numerical approaches of Lee and Richards[77], and other geometric constructions[78], are computationally costly and not suitable for our purpose, since folding requires many microseconds of MD that remain intractable using these existing methods. Analytical approximations expressed as a function of interatomic distances are more attractive. Wodak and Janin[79] developed the first algorithm exploiting probabilistic method, Hasel and Still *et al.*[80] modified it for atomic surface areas. Dynerman *et al.*[81] implemented this algorithm on GPU and refit the parameters to calculate SASA changes in protein docking studies. However, their approach is not ideal for MD simulations because when atom pairs are considered, the derivatives are not mutually of the same value and are not pairwise additive. Weiser and Still *et al.*[82] derived an even faster formula approximating atomic surfaces from linear combinations of pairwise overlaps (LCPO), which is the current non-polar implementation (gbsa=1) for Amber simulations. Along with another pairwise algorithm developed by Vasilyev and Purisima[83], it has been implemented on CPU for MD simulations. These are not optimal for our purpose because we seek for a simple and fast approximation that can be embedded in the same code loops as the other nonbonded energy terms in the current Amber GPU-implementation[61, 8], without the need of additional, nested loops for nonpolar term evaluations. Richmond[84] and later Wesson and Eisenberg[85] provided area derivatives with respect to the atomic positions, but they are not pairwise additive and also not suitable for fast parallel GPU implementation. Different approaches taken by Schaefer and Karplus *et al.*[86] make use of the Born effective radii calculated in GB equations, which is not independent of polar term used in solvation. It may also be beneficial to have a method to estimate SASA without the need for the full GB polar solvation calculation, for use in SASA-based methods that also estimate the polar solvation by using atom type specific surface tensions, or atomic solvation parameters (ASP), such used in the work of Eisenberg *et al.* [87] and some preceding work[85, 88].

Here, we explore a simple pairwise approach that would be amenable to fast GPU calculations. Our algorithm therefore is designed to estimate SASA from short-range atom pair distances. For each atom, the SASA equals a maximum value, subtracting the sum of the areas that are buried or shielded due to other neighboring atoms preventing waters from accessing to the atom of interest. The ideal shielding function would re-use terms that are already being calculated for the non-solvation energies and forces. In principle, this could provide a SASA estimate with nearly no additional computational cost. Inspired by Vasilyev

and Purisima’s work[83], where they employed a recursive Lorentz function to compute the central atom’s SASA from distances to all other atoms, we adhere to a single function, but without recursive iteration complexity, to maintain its pairwise evaluation and minimal burden in speed. A monotonic and continuously differential function is chosen to best represent the pairwise burial term. Compared to the previous analytical approaches assuming protein atoms are randomly distributed in space[79, 80], we utilize the unique geometry environments for different protein atoms by defining 30 SASA specific atom types for parameterization. These atom types help us incorporate non-pairwise contributions in a mean-field manner. Each atom type possesses one parameter representing the base maximum SASA value and another two parameters describing how much this atom can shield other atoms’ SASA and how this shielding value changes over distance. Trained to reproduce numerical SASA values for all the atoms in a novel training set of multiple peptides spanning all 20 amino acids, we validate the 90 resulting parameters on a test set of proteins. In addition to comparing SASA profiles for LCPO and our new method, we also compared the ensembles sampled in MD simulations using both SASA calculation methods, as well as simulations without nonpolar solvation.

In the present work, we use the SASA to estimate only ΔG_{np} , thus a reasonable first approximation is that the same surface tension could be used for all atoms. Since a variety of surface tension values have been suggested from different training sets[46, 89, 90], we further calibrated the surface tension that best reproduces explicit solvent data in a model system with precisely controlled set-up. In this model system (HC16, a 16-residue hydrophobic core fragment of HP36), the surface tension was empirically adjusted to correct the discrepancy between GB and TIP3P simulation results. The optimized surface tension was then used for GB/SA simulations on additional systems.

In this work, we present a fast new algorithm for calculating SASA, implement the atomic SASA calculations in Amber software on consumer GPUs, and apply our GPU-encoded GB/SA method on four proteins, CLN025, Trp-cage tc5b, HP36 and Homeodomain, to explore our hypothesis that incorporation of a nonpolar term could improve the predicted protein stabilities. We compared well-converged ensembles obtained using a consistent protocol except for the inclusion or omission of the nonpolar solvation energy. Our findings suggest a potentially valuable role of this inexpensive nonpolar term in the accuracy of our computational model, particularly in improving the predictive ability of ensembles generated using the GB solvent model in microsecond-timescale implicit solvent simulations.

2.3 Methods

2.3.1 Current theory

SASA-based nonpolar energy

A SASA-based nonpolar solvation model[46] was used, where the free energy is approximated by taking the product of the surface tension scaling factor (γ) and the Solvent Accessible Surface Area (SASA), where i is an atom which iterates over all atoms of this

solute.

$$\Delta G_{np} = \gamma \sum_i SASA_i \tag{2.1}$$

SASA estimations by ICOSA and LCPO algorithm

ICOSA[77, 91] surface area (gbsa=2 in Amber) SASA is a numerical method that recursively rolls a 1.4 Å radius water probe on the van der Waals surface of the molecule, starting from an icosahedron. The current implementation does not include derivatives of the SASA, so it is not possible to use in MD where forces are required.

LCPO[82] (Linear Combinations of Pairwise Overlaps, gbsa=1 in Amber) is the algorithm used for GB/SA MD simulations in recent Amber versions. It considers the neighbor list of a central atom and subtracts the pairwise overlaps from its isolated sphere area. In practice, this is a three-body approach, as not only the overlaps between the central atom and its neighbor atoms are calculated, but also the overlaps of the neighbors with each other. This adds to the computational complexity compared to our desired (non-recursive pairwise) approach.

2.3.2 Proposed fast pairwise analytical estimation algorithm

Physical rationale

Our first step is to assume that the SASA of the molecule can be approximated by considering only the heavy (non-H) atoms, and that H atoms can also be excluded in the calculation of solvent shielding of the heavy atoms. Estimating SASA just for heavy atoms results in a substantial reduction of atom pairs and computational cost, which also has been used in LCPO[82] and other algorithms[83].

The SASA of each atom in a protein configuration is its maximum surface area (termed *max_SASA_i*) subtracting the patches shielded by close neighbor atoms (termed *shielded_SASA_i*):

$$SASA_i = max_SASA_i - shield_SASA_i \tag{2.2}$$

The simplest although impractical case is solvation of a single atom; Both the *SASA_i* and the *max_SASA_i* for this atom is the surface area of this isolated sphere. In the context of proteins, all atoms have at least one covalent bond, and thus atoms are never exposed entirely to solvent. Importantly, we decided to handle the shielding by covalent and non-covalent neighbors differently, since the covalent neighbors (bonds and angles) likely have larger overlaps and closer distances than those sampled by purely non-bonded neighbors. This simplifies our construction of a function to estimate the shielding of an atom based on the distance of each neighbor. Therefore, the *max_SASA_i* includes shielding from covalent neighbors, and implicitly accounts for multi-body effects such as those from overlaps between covalent neighbors, and differences in accessibility due to hybridization variants. We also assume that the shielding by covalent neighbors (1-2 and 1-3 neighbors) is approximately independent of conformation, and thus *max_SASA_i* also is independent of the specific conformation, and these pairs are not included in the shielding calculations of each SASA evaluation.

In this context, what is an atom’s *max_SASA_i* in proteins? The answer is that it depends on the local geometry of an atom, including atoms that are covalently linked (bonds, angles

etc.). To describe the protein local geometries, we define 30 atom types with each representing one specific local geometry of an atom found in proteins (see the detailed classifications in Supporting Info **Table S2.1** and parameters in **Table S2.2**). Each element (C, N, O, S) is divided first into different hybridization states, then further divided based on the number of bonded heavy atoms. Some types were subsequently divided further to improve quality of fitting. As the type of bonded elements also matters in its max_SASA_i value, 30 total atom types, termed “SASA type”, were used to describe all the protein local geometries. This procedure is essential for formulating our new algorithm.

Each atom type has an associated constant max_SASA_i that is calculated after the fitting of the second term, $shielded_SASA_i$ (i.e. the pairwise burial term, or pairwise shielding effect on each other’s accessible surface area). To adhere to the pairwise decomposability, we make two assumptions that (1) the atomic surface area shielded by all other atoms is a sum of pairwise effect, which only iterates once for all the i,j pairs, when atom j iterates over remaining atoms with respect to atom i; (2) this pairwise effect could be represented as a single function of distance separating this atom pair.

$$shielded_SASA_i = \sum_j shielded_SASA_{i,j} \quad (2.3)$$

$$shielded_SASA_{i,j} = f(R_{i,j}) \quad (2.4)$$

As a result, $shielded_SASA_i$ for a specific atom pair i, j contributes the same SASA reduction to both atoms i and j, with symmetric forces. But as every atom in a protein possesses its specific local geometry (as defined by diverse SASA types and involving different neighbor atoms), iteratively evaluating all the pairwise atoms results in a unique sum for each central atom in its specific conformation of the protein.

In the next section, we focus on the considerations of the functional form we selected, and the parameters used for pairwise burial term evaluations.

Formula and parameterization design

Given the basic idea elaborated in Section 2.3.2, to calculate atomic SASA, a defined constant max_SASA_i subtracts a term $shielded_SASA_i$, computed from summing pairwise burial terms considering all close neighbor atoms within certain cutoffs ($shielded_SASA_{i,j}$) (**Equation 2.2** and **2.3**). The pairwise $shielded_SASA_{i,j}$ is assumed as a function of pairwise distances (**Equation 2.4**) and it is conceptually physical. As depicted in **Figure 2.1A**, it varies as the two atoms are apart at difference distances: when the distance is beyond a certain cutoff, water can traverse the gap and the SASAs are not shielded by each other; when the distance gets smaller, the SASA shielded on each other increases, until the atom fully displaces solvent and thus the shielded SASA reaches its maximum. Therefore, a sigmoid-like function with pairwise combinatory parameters is desirable.

Many options for a sigmoidal form are possible, including adapting some of the values calculated for the GB polar term for the nonpolar calculation[86]. Our choice of formula is informed by the Lennard-Jones function (depicted in gray in **Figure 2.1B**) that is already being calculated during the simulation, minimizing the additional computational overhead. The curve is monotonic and continuously differential at all points, and more importantly, the

pairwise combinatory fashion of van der Waals parameters could be adapted to generate pair specific $shielded_SASA_{i,j}$ parameters. Some transformations (a reflection of the Lennard-Jones curve over the y-axis followed by an up/right shift) result in a curve that fits our conceptual prerequisites (black curve in **Figure 2.1B**). When the distance of an atom pair $R_{i,j}$ is beyond a certain point, the resulting $shielded_SASA_{i,j}$ is zero; as $R_{i,j}$ gets smaller, the burial term gets larger before it reaches a plateau and sensitivity to distance decreases as water is fully displaced. The $cutoff_{i,j}$ values are also taken in a pairwise combinatory way (**Equation 2.8**).

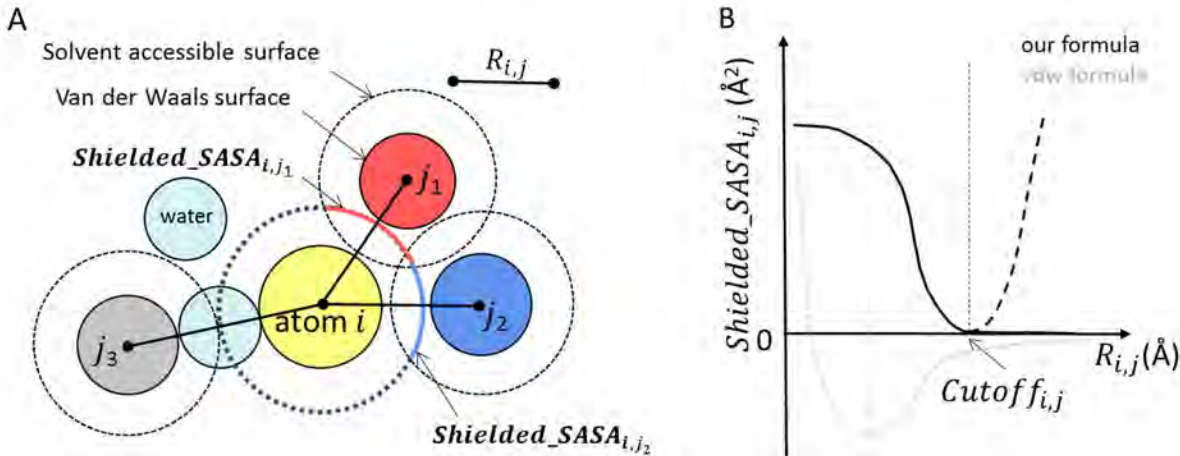


Figure 2.1: 2D illustration of the proposed algorithm and the key formula for calculating $shielded_SASA_{i,j}$. A. atom i in yellow is the central atom of interest; its SASA (central dotted circle) shielded by atom j_1 in red and atom j_2 in blue are calculated, respectively, using the pairwise distances $R_{i,j}$. Atom j_3 in gray is beyond the cutoff distance to atom i thus contributes zero to $shielded_SASA_{i,j}$. B. Our formula (black) is a transformation of the standard Lennard-Jones 6-12 formula (gray), by a reflection over y-axis followed by an up-right shift. Details of the derivation are provided in **Figure S2.1** and **Equation S2.1** to **S2.6**.

The stepwise derivation of the pairwise burial term $shielded_SASA_{i,j}$ is provided in **Equation S2.1** to **S2.6**, with the final equation as a function of $R_{i,j}$ shown below:

$$shield_SASA_{i,j} = \begin{cases} \varepsilon_{i,j} \left(\frac{\frac{n}{m-n}}{\left(1 + \frac{cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^m} - \frac{\frac{m}{m-n}}{\left(1 + \frac{cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^n} + 1 \right), & \text{if } R_{i,j} < cutoff_{i,j} \\ 0, & \text{otherwise} \end{cases} \quad (2.5)$$

where $\sigma_{i,j}$ and $\varepsilon_{i,j}$ are calculated from SASA-type specific parameters discussed below. The values of m and n are also discussed below. $Cutoff_{i,j}$ is a pairwise constant calculated from atomic radii.

For each atom in one SASA type, we need two parameters σ_i and ε_i to describe its ability to shield other atoms (hence 60 total). For each atom pair, we use the Lorentz-Berthelot combination rules to obtain the $\sigma_{i,j}$ and $\varepsilon_{i,j}$ values:

$$\sigma_{i,j} = \sigma_i + \sigma_j \quad (2.6)$$

$$\varepsilon_{i,j} = \sqrt{\varepsilon_i \varepsilon_j} \quad (2.7)$$

The cutoff distance is employed to ensure that when two atoms are far enough apart ($R_{i,j} \geq Cutoff_{i,j}$) they do not contribute to each other’s *shielded_SASA*. This eliminates the repulsive portion originally present in the Lennard-Jones-type function (dashed line in **Figure 2.1B**) and ensures force continuity through the cutoff distance. Cutoff distances are the sum of the atomic radius and the water probe radius (1.4 Å). The same atom radii for four elements (C 1.7 Å, O 1.5 Å, N 1.55 Å and S 1.8 Å) were used both here and in ICOSA. Different radii were used with LCPO (C 1.7 Å, O 1.6 Å, N 1.65 Å and S 1.9 Å) to be consistent with those used during the original training of the 54 LCPO parameters[92].

$$cutoff_{i,j} = cutoff_i + cutoff_j \quad (2.8)$$

$$cutoff_i = Atom_Radius_i + 1.4 \text{ \AA} \quad (2.9)$$

The exponents m and n determine the steepness of the *shielded_SASA_{i,j}* transition as the two atoms approach. Values for n were tested among 2, 4, 6 and 8; n = 4 gave the best resulting atomic SASA fitting correlation (data not shown). Correlation was less affected by the choice of m when 10 and 12 were used for comparison, so m = 10 was initially used in the optimization. As other values for m and n did not improve the accuracy of the algorithm (data not shown), and parameterizations of σ_i and ε_i values also affect the depth and steepness of the pairwise curves, we kept m = 10 and n = 4 for all atom pairs.

Training set and fitting strategy

The 60 SASA type specific shielding parameters were fit against ICOSA SASAs (also calculated using only heavy atoms) on a training set of 10 peptides. To cover a broad spectrum of atomic environments, possible atomic pairwise contacts and extents of burial, we designed a set of 10 sequences (**Table S2.3.**); each is a scrambled sequence made of all 20 natural amino acids (using all 3 protonation variants for the His side chain, thus each peptide is 22 amino acids in length). Together, conformational ensembles for these scrambled peptides provide significant statistics for atomic SASA ranges, and they encompass the distributions of pairwise distance distributions expected in real proteins (**Figure S2.2**).

For each sequence, 50 geometries of diverse structures were included in the training set ensembles. Ensembles were generated as follows: initial conformations were generated from fully extended structures constructed using *tleap*, with 1000 steps of minimization to ensure reasonable initial geometries. This was followed by 1 μ s of unrestrained MD simulation at 300 K (using a Langevin thermostat, the ff14SBonlysc[23] force field in GBNeck2[62] solvent without SA term) producing 5000 conformations equally spaced in time. The *cpptraj*

program[42] was used to separate each trajectory into 50 clusters using the hierarchical agglomerative algorithm, based on all 22 C α atoms. These 50 representative structures from each peptide sequence comprised the training set ensembles. **Table S2.3** shows the representative structure of the most populated cluster for each peptide. We calculated reference atomic SASA values for each heavy atom in each structure using a modified version of sander (where *Atom_Radius* for hydrogen was set to zero in icosahedron subroutine) in Amber 16[8].

Fitting of parameters was done as follows. Initial guesses for all 60 parameters were randomly generated, then were optimized using *l_bfgs_b* algorithm[93] in the Python Scipy package[94]. The objective function used for optimization was:

$$score = \sum_{scramble_peptide=1}^{10} \sum_i^{atoms} \sum_{frame_pair=1}^{250} (\Delta SASA_{atom_icosahedron} - \Delta SASA_{atom_fitted})^2 \quad (2.10)$$

where:

$$\Delta SASA_{atom_i} = \Delta SASA_{frame_a} - \Delta SASA_{frame_b} \quad (2.11)$$

where frame.a and frame.b represent two different conformations from the training set for that peptide. As Vasilyev and Purisima[83] have pointed out, the change in the accessible surface area is often of more interest than the absolute value. In addition, as *max_SASA_i* is a constant for one specific SASA type, fitting to $\Delta SASA_i$ results in isolation of the 30 *max_SASA_i* parameters since they cancel in the target $\Delta SASA_i$ values. For these reasons, we fit the 30 sets of σ_i and ε_i parameters to the SASA difference between pairs of conformations.

Instead of iterating over all combinations of conformation pairs, we sorted the atomic SASA of all 500 representatives, picked the 2 conformations with largest and smallest SASA as the first pair, then the second largest and the second smallest as the second pair, and so on. The reasons not to include all pairs of 500 conformations are (1) all 250,000 conformation pairs per atom for one evaluation of optimization is more time-consuming, (2) many data are redundant if each conformation to every other conformation is included, and (3) most importantly, most of the SASA differences are quite small if all conformation pairs are included, and the squared differences would weigh even less in the optimization function (**Equation 2.10**), resulting in inefficient data use. In the end, we adopted a sorted pair scheme that included 250 pairs of conformations for each atom in optimizations, and a flatter distribution of SASA difference values compared to the more costly all pairs scheme (**Figure S2.3**).

As discussed above, fitting the changes in SASA results in cancellation of the *max_SASA_i* in the scoring function. One extra step of calculating the 30 *max_SASA_i* values was done at the very end, when *shielded_SASA_i* for all atoms in the training set were calculated with the optimized 60 parameters. For each SASA type, the *max_SASA* was obtained by taking the arithmetic average of the difference between the icosahedral SASA and the calculated *shielded_SASA*, over all atoms of that SASA type:

$$max_SASA_i = \frac{1}{N} \sum_{peptide\ conformations}^{10} \sum_{type_i}^{50} (SASA_{atom_icosahedron} - shielded_SASA_i) \quad (2.12)$$

where N equals to the total numbers of atoms of that SASA type.

Test set

The proteins displayed in **Figure 1.3** were used as a test set to validate SASA estimation. This set of proteins of diverse topologies ranging from 10 to 92 amino acids corresponds to the set we previously used for ab initio protein folding[24]. The structural ensembles for the test set were extracted from the protein folding trajectories in that work to get a set of structures spanning diverse atomic and molecular SASA values to evaluate the new pairwise model. The model system HC16 was also included. Reference data were calculated for each structure using the ICOSA and LCPO algorithms.

2.3.3 Simulated protein systems

HC16 with helical restraints

HC16 (16-residues with ACE and NHE caps, with sequence DEDFKAVFGMTRSAFA) consists of the hydrophobic core of HP21 (a Villin headpiece HP36 fragment). HP21 was reported to transiently adopt native-like conformation[95] similar to that in full-length HP36. To facilitate obtaining converged data in explicit solvent, and also to maximally isolate the difference between simulations to the presence or absence of nonpolar solvation, we restrained 7 hydrogen bonds in the backbone of HC16 with 50 kcal/(mol·Å²) force constant: ACE.O-Phe47.H (1.94 Å), Asp44.O-Lys48.H (1.95 Å), Glu45.O-Ala49.H (2.41 Å), Phe47.O-Val50.H (1.87 Å), Thr54.O-Phe58.H (1.67 Å), Arg55.O-Ala59.H (2.24 Å), Ala57.O-NHE.H (2.07 Å). The distances for the restraints (listed respectively in the parentheses and depicted in **Figure 2.2**) were selected as those in the NMR structure[96] of HP36.

Unrestrained CLN025, Trp-cage, HP36 and Homeodomain

Chignolin variant CLN025 is a 10-residue mini-protein with sequence YYDPETGTWY. Reported in 2008, CLN025 adopts a stable hairpin conformation, determined by both crystallography (PDB code: 5AWL[97]) and aqueous state NMR (PDB code: 2RVD[97]).

Trp-cage variant tc5b is a 20-residue mini-protein with sequence NLYIQWLKDGPPSS-GRPPPS. Designed and solved in NMR (PDB code: 1L2Y[98]) in 2002, it is designated as the ‘Trp-cage’ motif because the burial of a hydrophobic Tryptophan side chain is thought to be a driving force of its folding. It has secondary structure of an α -helix, a short 3_{10} -helix and the Trp indole ring encapsulated in a cluster of Proline rings.

HP36 is the naturally found 36-residue Villin headpiece subdomain with a full sequence MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF. It is recognized to fold into a compact native state solved by NMR (PDB code:1VII[96]) with three α -helices.

Homeodomain is a 52-residue computationally re-designed variant of *Drosophila melanogaster* engrailed homeodomain, with sequence MKQWSENVEEKLKEFVKRHQRITQEELHQYAQRLGLNEEAIRQFFEEFEQRK. The NMR solved native structure (PDB code: 2P6J[99]) is thermally stable and also consists of three α -helices but adopts a different fold from HP36.

Experimental melting curves for CLN025[97], Trp-cage[98] and HP36[100] were obtained from CD experiments. The melting temperature of Homeodomain variant was measured from CD[99]. All 4 systems were previously studied in our ab initio folding experiments[24]

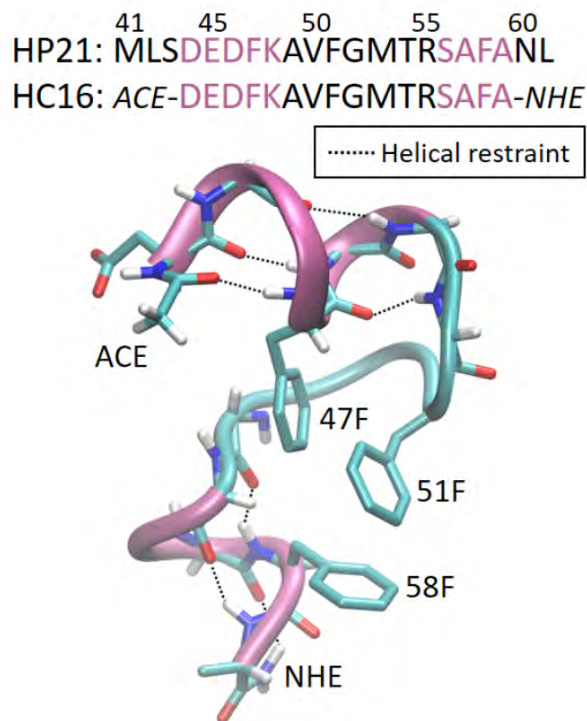


Figure 2.2: HC16 (Hydrophobic-Core 16-residue) sequence and conformation in NMR structure of HP36 (PDB code: 1VII[96]). The sequence of HP21 which has been characterized in experiment[95] is also listed for comparison. The two helices shown in pink are restrained with hydrogen bonds shown in black dotted lines. Side chains of three Phenylalanine (comprising the hydrophobic core of HP36) and capped termini are denoted.

using the same force field and solvent model used here, providing an excellent reference to quantify the possible improvement by addition of a nonpolar solvation term.

2.3.4 MD simulation and analysis details

Explicit solvent simulations of restrained HC16

Helical restraints described in **Figure 2.2** were applied to the HC16 system in explicit solvent simulations. Two sets of simulations were initiated from two conformations: one “restrained unfolded” and the other as observed in HP36 NMR structure. The “restrained unfolded” conformation was generated from a short high-temperature MD simulation starting from the NMR structure; after this 1 ns short MD run at 500K with helical hydrogen bonds and chirality restrained, the conformation of maximal end-to-end distance (25.9 Å vs. 16.0 Å as in NMR structure) was equilibrated with helical restraints at 300 K as “restrained unfolded” structure. Restrained HC16 parameterized in ff14SBonlysc[23] was solvated with 2187 TIP3P[101] water molecules in a truncated octahedral periodic box. The distance from solute to the edge of the box was 9 Å for the “restrained unfolded” structure, and increased to 11.061 Å for the NMR structure so that the total number of atoms was equivalent for the two simulations. For the equilibration, 10000 steps of energy minimization were first done with

100 kcal/(mol·Å²) restrained on all heavy atoms, followed by 100 ps of MD heated from 100 to 300 K at constant volume. Then 100 ps and 250 ps of constant pressure MD simulations were done with 100 and 10 kcal/(mol·Å²) force constant, respectively. Another 10000 steps of minimization with backbone restraints of 10 kcal/(mol·Å²) was followed by 100 ps of MD simulation at constant pressure and temperature. Then three 100 ps simulations (with 1, 0.1, 0 kcal/(mol·Å²) backbone restraints, respectively) were done with helical restraints. The helical restraints were applied throughout the production runs. Replica Exchange Molecular Dynamics (REMD) simulations were performed to help overcome viscosity barriers in explicit solvent, using 32 replicas in the NVT ensemble; 8.0 Å was used as the non-bonded interaction cutoff; Particle Mesh Ewald (PME) was used for long range electrostatics; Langevin dynamics with 1 ps⁻¹ collision frequency was used; thermostat temperatures ranged from 294.4 K to 394.4 K (the full temperature ladder is reported in **Table S2.4**). Each replica was simulated for > 2.6 μs, giving a cumulative 83 μs of simulation time and requiring about 15 days on Tesla K20X Amber 16 GPU (CUDA) version of PMEMD. The PMF profile at 300 K was calculated with temperature-biased weighted histogram analysis method (TWHAM)[102].

GB and GB/SA simulations of restrained HC16

SHAKE constraints[39] were applied on all hydrogen involved bonds. Langevin dynamics with 1 ps⁻¹ collision frequency (ntt=3) and hydrogen mass repartitioning[40] (allowing a 4 fs time step) were used in all implicit solvent simulations.

Restrained HC16 parameterized in ff14SBonlysc was simulated in GBNeck2 (igb=8) with mbondi3 radii[62]. GB simulations without nonpolar solvation used gbsa=0. Two runs of Langevin dynamics simulations starting from two conformations were run at 300 K, each for 16 μs. Clustering analysis comparing pairwise RMSD between structures were done on the last 8 μs of simulations (2 runs of 8000 frames, 16000 frames in total). The hierarchical agglomerative algorithm in cpptraj program was used for clustering, based on all 16 Cα atoms at a 2 Å cutoff.

REMD was used to enhance the sampling efficiency for all GB/SA simulations since compact conformations were stabilized relative to unfolded states, and simulations at 300 K sampled high RMSD conformations too rarely for precise quantification of stability. In Amber, gbsa=1 was used for LCPO algorithm and gbsa=3 was used for our new pairwise model. Surface tension values (surften flag) of 5, 7, 10 and 12 cal/(mol·Å²) were tested. For each surface tension, two production runs starting from “restrained unfolded” and NMR structure were simulated to 4 μs per replica of REMD with 6 replicas to get converged data; thermostat temperatures ranged from 279.5 K to 397.9 K (see **Table S2.4**). It took > 60 days for GB/(LCPO)SA to generate 4 μs of simulations on 4 cpu cores for each replica, while 4 days were sufficient to collect the same amount of data for GB/(pairwise)SA, on 1 GXT680 GPU for each replica.

GB and GB/SA simulations of unrestrained proteins

CLN025, Trp-cage, HP36 and Homeodomain variant were simulated without restraints in REMD, employing ff14SBonlysc and GBNeck2; both LCPO and our pairwise SA were used. Surface tension was 7 cal/(mol·Å²) unless otherwise specified. For each system, two

production runs starting from fully extended or NMR structure were simulated. For CLN025, 6 replicas (252.3 K – 389.1 K) REMD were done for 1.3 μ s in GB, 1.5 μ s in LCPO and 8 μ s in pairwise SASA. A backbone RMSD cutoff of 2.2 Å was used for calculating fraction of folded, consistent with our previous study. For Trp-cage, 8 replicas (247.7 K – 387.3 K) REMD were simulated for 1.7 μ s in GB, 1.4 μ s in LCPO and 4 μ s in pairwise SASA. A backbone RMSD cutoff of 2.0 Å was used for calculating fraction of folded. In both CLN025 and Trp-cage, the last half of the trajectories of the two runs were used for melting curve plotting. For Homeodomain variant, 12 replicas (288.7 K – 440.3 K) REMD were simulated for 4 μ s for GB and pairwise. A backbone RMSD cutoff of 5.0 Å was used for fraction of folded calculations. For HP36, 8 replicas (250.0 K – 349.0 K) REMD were simulated for 4.2 μ s in GB. As simulations in LCPO used surface tension of 10 cal/(mol·Å²) and were run for 650 ns, the pairwise SASA used the same surface tension to be consistent. REMD simulations were run for 24 μ s to converge. Cluster analysis was done on the lowest temperature trajectories (250 K) using the same protocol as for HC16 GB trajectories, based on all 36 C α atoms. Another set of HP36 REMD simulations were carried out in ff14SB and GBNeck2 for 20 μ s to show the observed misfolding of HP36 could be ascribed to a force field issue.

PMF calculations

Potential of Mean Force (PMF) structure equilibria profiles were calculated using a collective variable of RMSD of all C α atoms, against native structure as in HP36. This can be interpreted as the reconstructed free energy landscape projection onto the RMSD space. We first histogrammed the RMSD values for all sampled structures at 300 K (either directly from MD simulations running at 300 K or extracting 300 K trajectories from REMD simulations), using a bin size of 0.1, in the range 0-7 Å. We then defined the relative free energy for each bin, using **Equation 2.13**:

$$\Delta G_i = -RT \log \frac{N_i}{N} \quad (2.13)$$

where R is the gas constant ($1.9858775 \times 10^{-3} \text{kcal}/(\text{mol}\cdot\text{K})$), T is 300 K, and N is the largest bin population. The error bars on PMF plots reflect the absolute deviation of free energies for each bin calculated from two independent simulations starting from different conformations.

2.4 Results and Discussions

2.4.1 SASA estimation by the proposed algorithm

Parameterization on atomic SASA of training set

As stated in Methods, we defined 30 SASA types, each with two parameters σ and ε , to characterize variation of SASA with the possible pairwise atomic contacts found in proteins. All 60 parameters were optimized to achieve the least square errors with respect to the ICOSA-based SASA numerical changes for all the heavy atoms in the training set. The optimization took multiple rounds to best reproduce $\Delta \text{SASA}_{atom.icosa}$ in **Equation 2.10**. We verified that reducing the SASA types or the peptide species worsens the fit quality, but

using fewer frames for each peptide ensemble made less difference. The resulting σ and ε values are provided in **Table S2.2**, along with the calculated max_SASA parameters.

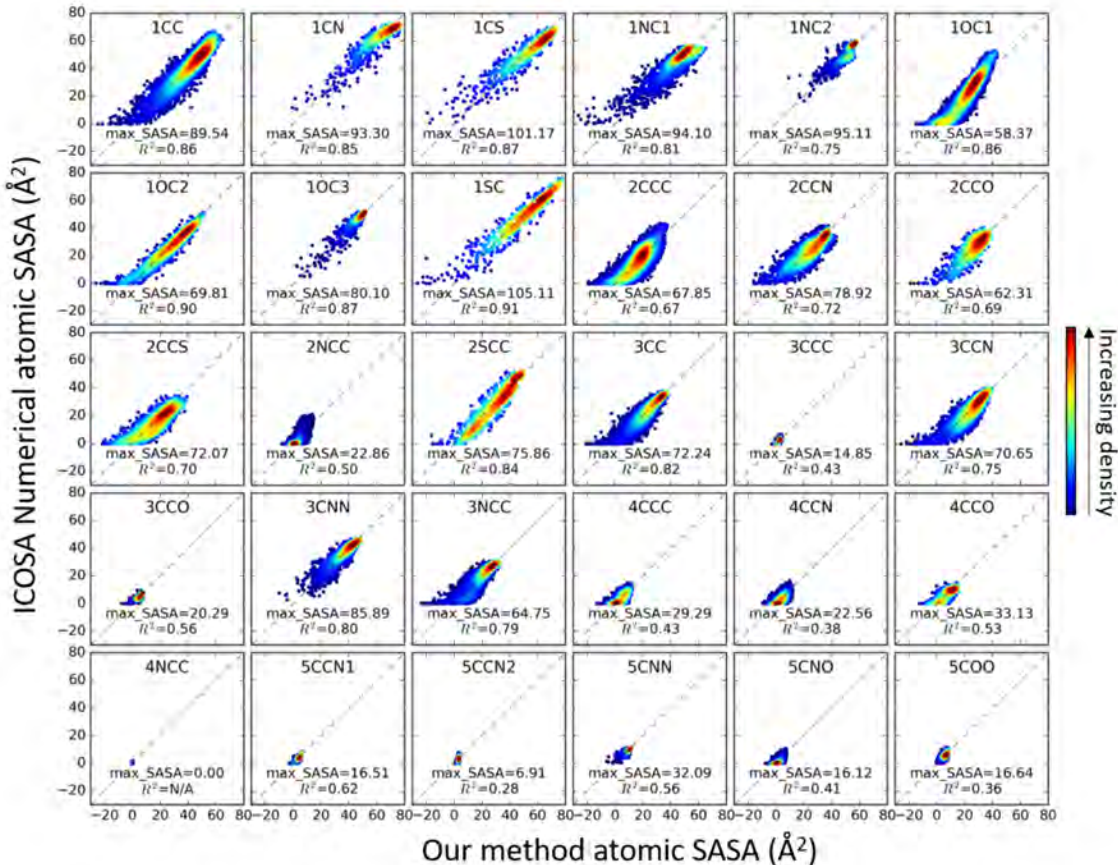


Figure 2.3: 2D histograms of pairwise atomic SASA of each SASA type, versus ICOSA-based numerical values in the training set. Perfect agreement would coincide with the diagonal dashed lines. The color indicates the kernel density estimated using `scipy gaussian_kde`[103].

The final set of parameters reasonably reproduces the atomic SASA for heavy atoms in the trained peptides, shown separately for each of the 30 SASA types in **Figure 2.3**. Among all the types, hydrogen atoms are defined as ‘1H’ type and neglected in both reference and estimation. Nitrogen atoms ‘4NCC’ that bond with 3 other heavy atoms in sp^3 hybridization were set to zero SASA, for they are highly buried in trained peptides and test proteins. The estimated atomic SASA values scatter around the diagonals that represent perfect fittings. In particular, the diagonals go through the densest data (dark red) regions for all atom types, which indicate excellent agreement for the most frequently sampled atomic SASA values. The coefficients of determination (R^2 for the linear regression between ICOSA values and our estimations) vary from 0.28 (‘5CCN2’) to 0.91 (‘1SC’). However, those with lower correlations tend to adopt a small range of SASA values (e.g. ‘2NCC’, ‘3CCC’); the R^2 for atom types sampling atomic SASA over 20 Å are all above 0.50. For the atom types that

seldom exposed to solvent (e.g. “4xxx”, “5xxx”), the pairwise estimate also indicates burial with close to 0 Å atomic SASA.

For the totally buried heavy atoms, our algorithm sometimes produces negative SASA values. The appearance of such unphysical results and the inaccuracies of estimation root in the simplification of a two-body algorithm. This pairwise burial algorithm assumes the mutually buried surface areas could be averaged to a pairwise fashion, which could be captured by one monotonic function analytically. But we could tell from the estimation that this assumption works better for the exposed atoms. When our fitting works well for the more exposed instances of a SASA type, the accuracy suffers for the most buried examples of that SASA type. For example, for the ‘1SC’ type atoms that have SASA > 30 Å², data points fall closely around the diagonal and visually correlate well, compared with lower accuracy for the instances with SASA < 30 Å². This observation applies to almost all other SASA types. When atoms become deeply buried, our current algorithm continues to assign (small) shielding contributions from atoms in the tail of the sigmoidal function. A better-designed switching function could eliminate these negative SASA values, but in the current implementation we did not explore this more since our goal was to develop a simple, fast approach, and the frequency of observing the slightly negative SASA values is quite low overall. Furthermore, the changes in the SASA are more important than the absolute values.

2.4.2 Estimation of molecular SASA in test set

Generally, we would expect that high correlations of the atomic SASA values (calculated to obtain forces) would also result in accurate molecular SASA values when the atomic values are summed. However, we observed that the sum of estimated atomic SASA values (**Figure S2.4**) systematically deviates from the numerical molecular values (**Figure S2.5**), which was also encountered in Dynerman *et al.*’s work where computed SASA values (desolvation energy changes calculated from SASA, to be specific) systematically deviate from numerical numbers in a proportional manner[81]. We ascribe it to be a negative consequence of tolerating inaccuracies in atomic SASA pairwise estimation. The occurrence of negative SASAs, along with correlation in errors for different atoms, attribute to cumulative errors in molecular SASA estimations, which was further adjusted by linear transformations.

Given the systematic error from summing our simple pairwise atomic SASA estimates, we decided to empirically adjust the sum of our atomic estimations to more closely match molecular values. By comparing total SASA values we found that a universal scaling factor 0.6 worked well; in terms of energy and forces, this is equivalent to scaling the designated surface tension γ by 0.6. In Amber, we encoded the scaling factor directly so users could obtain similar results for different SASA algorithms when setting a particular surface tension value. It is recommended for the users to bear in mind that the total SASA in **Figure 2.4** are obtained from the atomic SASA shown in **Figure S2.4** using the following transformation (see detailed description and **Figure S2.6**):

$$SASA_{molecule} = adj_max_SASA - 0.6 \times \sum_i^{natoms} shield_SASA_i \quad (2.14)$$

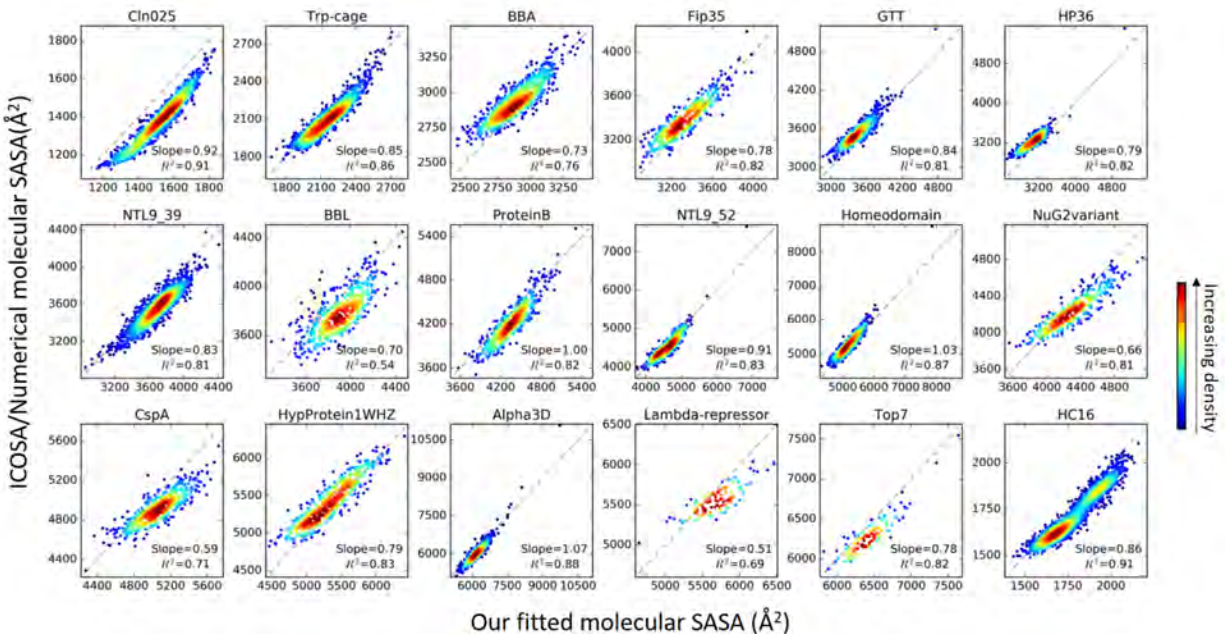


Figure 2.4: 2D histograms for each protein, showing fitted molecular SASA versus ICOSA numerical values for the test set. Perfect agreement is indicated by the diagonal dashed lines. The color indicates the kernel density estimated using `scipy gaussian_kde`[103].

$$adj_max_SASA = 0.6814 \sum_i^{natoms} max_SASA_i + 361.1 \quad (2.15)$$

After transformations using **Equation 2.14** and **2.15**, the estimated molecular SASA values become better estimates of the numerical values for the 18 test protein systems in **Figure 2.4**. The coefficients of determination range from 0.54 for BBL, 0.69 for λ -repressor, to above 0.8 for CLN025, Trp-cage, Fip35, GTT, HP36, HC16, NTL9 (39 and 52 residues), ProteinB, Homeodomain, NuG2variant, Hypothetical protein 1WHZ, α 3D, and Top7. Overall, in 15 out of 18 protein test systems, we can estimate the SASA to well correlate with numerical calculations (Pearson correlation efficient, $R^2 > 0.81$) across the range of sampled conformations. This is encouraging given that the parameters were trained on short peptides with scrambled sequences; even though pairwise atom contacts are similar between the training and test set, the transferability to larger proteins is still reassuring.

In most cases, our fast estimations tend to slightly overestimate the ICOSA molecular SASA differences (indicated by slope < 1), but the same effect is also observed in LCPO-based SASA predictions for the same test set (**Figure S2.7**); this can be attenuated by decreasing the chosen surface tension. Notably, the cases for which our estimation qualities are worse than average (BBA, BBL, NuG2variant and λ -repressor) are also challenging and among the worst predictions for LCPO. This suggests there may be some specificities in these proteins, where local geometric features are insufficient for predicting solvation properties. It is possible that the estimation qualities could be further improved by refining the functional forms for our pairwise estimates, or fitting shielding parameters for atom type pairs. However, a SASA-only nonpolar term is itself a crude estimation of non-electrostatic

solvation, perhaps suggests that adding further complexity and computation cost may not be worthwhile. However, before this work, except nonpolar term, all the other energy terms were accessible on GPUs. Having the nonpolar term left out hindered the possibility of extensive tests with a more complete description of solvation, and quantitative analysis of the impact of SASA-based nonpolar solvation on well-converged ensembles for non-trivial systems. Thus our focus here is not on an ideal SASA calculation, but what benefits, if any, can be obtained from simple SASA-based approaches used during protein MD simulations. Once these are implemented in a form fast enough to converge ensembles for non-trivial proteins, it will become possible to examine the extent to which further optimization can improve agreement with experiment. In the next section, the acceleration in MD simulations achieved by GPU implementation is illustrated and described in detail. The efficiency of our algorithm is compared to LCPO. Convergence is comparable within the same simulation time, but the overall wallclock speed (computational cost) of the simulations is sped up by more than an order of magnitude using our approach.

2.4.3 Speed up in MD simulations

After implementation in the Amber software, simulation benchmarks establishing the performance of simulating unrestrained HP36 are shown in **Figure 2.5** below. On CPUs, simulations using GB, GB/SA (LCPO) and GB/SA (pairwise) are similar in speed, with LCPO being slightly slower. However, our method was really targeted to GPU-style massive parallelism. Compared to less than 40 ns/day with 8-core CPU clusters, the slowest GPU calculation (GTX 970) provides 665 ns/day using our pairwise approach. Importantly, adding the pairwise SASA calculation incurs little additional overhead compared to simulations without it (676 ns/day). As the compute capability of GPU increases, the speed accelerations over LCPO reached 31x (single GTX 980). These accelerations are comparable to standard Amber GPU performance[61], and are also consistent with our design of the algorithm. The only information needed is how far each central heavy atom is from its close neighbor atoms within the solvent accessible distances, and with no recursive neighbor-neighbor calculations required. These distances have already been pre-calculated and cached for electrostatic, van der Waals, and polar part of solvation computation and the SASA calculation can be embedded in the same loops and parallel decomposition schemes. Our nonpolar calculation is also implemented fully on the GPU, without the need to transfer back and forth between GPUs and CPUs, as is necessary by the current LCPO code.

The parameter set for our proposed algorithm was coded in a modified version of Amber version 16[8]. Setting the gbsa flag to 3 in GB simulations activates GB/SA using the new nonpolar solvation term in the sander, pmemd or pmemd.cuda (all precisions) programs. Compared with the existing hybrid GPU/CPU algorithm[104] needing the CPUs for the LCPO algorithm and GPU for remaining terms in the force field, our method calculates all energy/force terms on GPUs, if designated, thus accelerates the MD production runs by tens of times.

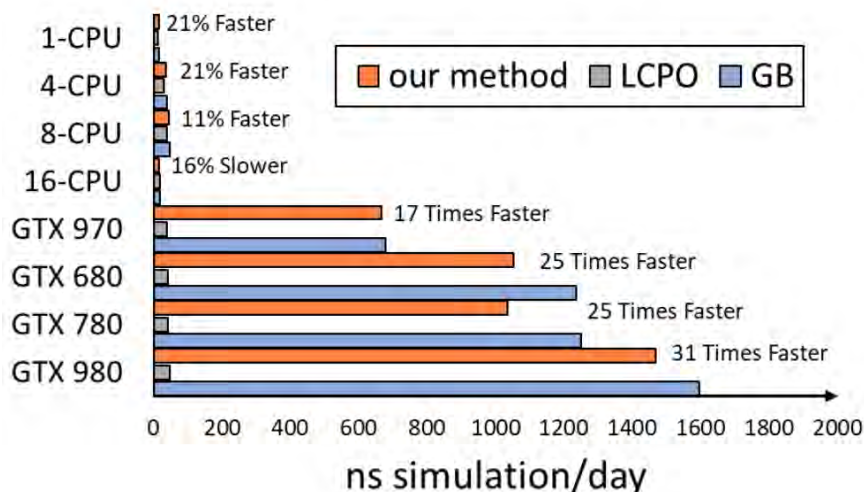


Figure 2.5: Performance benchmarks on CPUs and single gaming GPUs, simulating HP36 in GB and GB/SA (LCPO and our pairwise method) models. The speed up multiples (percentages) denoted are calculated from the respective ns simulation/day achieved in our method divided by that obtained using LCPO on the same architecture.

2.4.4 Stability of the hydrophobic core in the HC16 model system

Calibration model system and rationale

We next carried out a quantitative comparison of explicit and implicit solvent results on a controlled peptide fragment, in which the role of the solvent model could be isolated from other variables that confound direct comparison to experiment such as protein force field accuracy. We use the hydrophobic core of HP36, a peptide of 16 residues termed “HC16” (**Figure 2.2**), including a neutral acetyl capping group (ACE) at the N-terminus and amidation (NHE) at the C-terminus. HC16 retains the structured region of HP21, which was previously reported to adopt HP36 native-like structure as a fragment[95]. Used as a model system for hydrophobic residue clustering study, this packed hydrophobic core is made of side chains protruding from two α -helices, particularly the three phenylalanine residues, Phe47, Phe51, and Phe58 (we adopt the widely used numbering of residues derived from intact Villin headpiece). On such a small and fast folding peptide, it is more practical to obtain converged sampling across available configuration spaces, using both explicit and implicit solvent models. As stated earlier, explicit solvent, by default, contains all solvation effect including nonpolar interactions, whereas implicit solvent simulations with and without non-polar term, are likely to show variations in the configuration space of HC16. The HC16 model system is precisely controlled by setting nonpolar term as the single variable in benchmark simulations; we hypothesize that when the two helices in HC16 are rigorously restrained to the secondary structures adopted in folded conformations, the thermodynamic stability in hydrophobic core formation and breakdown is dominated by the effectiveness of nonpolar term. We restrained the backbone of the two helices and sampled their relative orientation and packing in MD. Restraining the helices has the double benefit of (1) simplifying sam-

pling in explicit solvent (still highly challenging to fully converge for 16 amino acids), and (2) reducing the potential influence of differences in secondary structure propensity from the polar portion of the implicit/explicit solvent[105] (although we note that the GBneck2 model used here has excellent agreement with TIP3P in this respect[62]). Thus, our expectation is that the discrepancies observed between explicit and implicit solvent simulation ensembles should largely arise from the differences in the nonpolar solvation treatment.

In the restrained HC16 model system, we used consistent computational methods and simulation protocols except for the nonpolar term: GB: GB as the polar term and no nonpolar term used; TIP3P:TIP3P as a full solvation description of both polar and nonpolar terms; and GB/SA: GB as the polar term and nonpolar term incorporated through SASA, modulated by scaling the surface tension. Comparing ensembles from LCPO and our pairwise method evaluates our SASA approximation, and comparing the TIP3P, GB and GB/SA simulations allows tuning of an appropriate surface tension value and evaluation of the extent to which this approximation can improve reproduction of explicit solvent results.

2.4.5 Quantification of discrepancies between GB and TIP3P

As stated earlier, proteins solvated in GB model alone exhibit low folding stability[24]. We hypothesized that this is due to lack of nonpolar solvation stabilizing the hydrophobic core, and that an explicit solvent model like TIP3P may produce a more accurate result. Therefore, we first investigate structurally and energetically the conformational equilibrium of HC16 in both GB and TIP3P to see if stability differences are recapitulated, by comparing well-converged simulations that are largely identical except for nonpolar solvation.

Although the PMF profiles all exhibit dominant global minima at low RMSD values as shown in **Figure 2.6A**, differences in the nonnegligible stability manifest discrepancies in the sampled structural ensembles. Without the nonpolar term, GB predicts a smaller energy gap and flatter energy surface for the unfolded conformations. The GB PMF falls below the TIP3P PMF as soon as the RMSD advances beyond the native-like minimum, with maximum energy difference close to 2 kcal/mol (at around 4 Å C α -RMSD). Furthermore, the cluster analysis (see **Table S2.5**) of the simulated GB trajectory manifests the compositions of three dominant conformations of various SASA values (shown in **Figure 2.6C**). Compared to the second dominant cluster (4.1 Å C α -RMSD, cluster 2), the native cluster (1.0 Å C α -RMSD, cluster 1) has smaller SASA values, suggesting that a nonpolar term could stabilize the native-like cluster by 1-2 kcal/mol. The third cluster (5.4 Å C α -RMSD), with SASA falling between cluster 1 and 2, would also be modestly stabilized with respect to cluster 2. The combination of the lower hydrophobic core stability in GB MD, along with the difference in SASA between the clusters with and without hydrophobic core suggests that a SASA-based algorithm might appropriately stabilize the native-like cluster and improve agreement between implicit and explicit solvent.

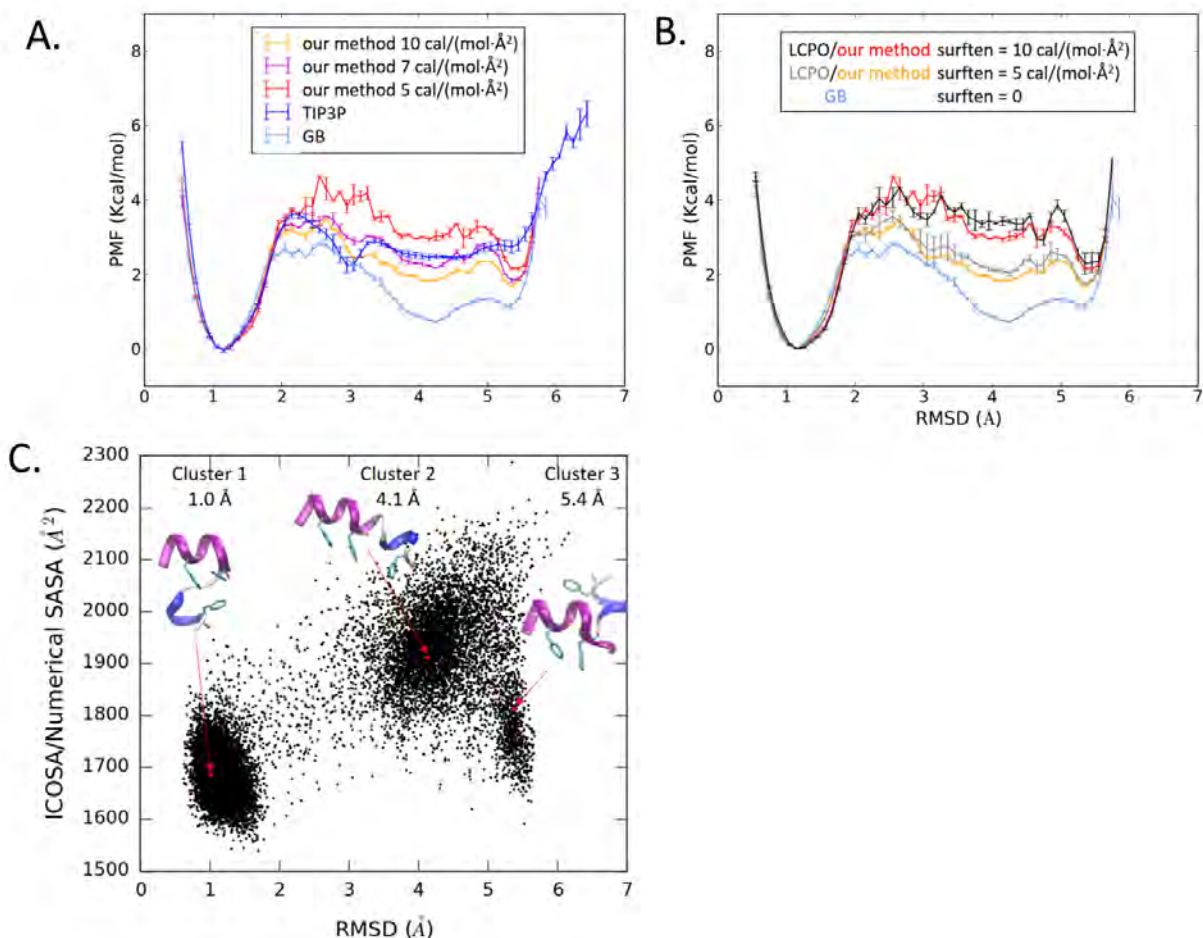


Figure 2.6: Structural equilibria of restrained HC16 simulated in GB, TIP3P and GB/SA water models at 300 K. A. PMFs for structural equilibria of HC16 measured by $C\alpha$ -RMSD, by varying the effectiveness of nonpolar solvation, from no nonpolar effect (pure GB), to increased nonpolar effect as surface tensions in GB/SA simulations increase, and to full solvation with TIP3P; B. PMFs for structural equilibria of HC16 measured by $C\alpha$ -RMSD comparing two GB/SA methods (LCPO and our method) and GB; C. 2D scatter plot of ICOSA/Numerical SASA versus $C\alpha$ -RMSD against NMR structure fragment of HC16. The top three cluster representative structures are indicated in the figure.

2.4.6 The new pairwise algorithm closely matches LCPO

Before we compare the effect of SASA-based nonpolar solvation (GB/SA) to explicit solvent result, we first compared the effects obtained using two different GB/SA methods in Amber: `gbsa=1` and `3` for LCPO and our pairwise method, respectively. This allows us to evaluate the ability of our pairwise approximation to recapitulate the ensemble shifts obtained with LCPO, as compared to the analysis in 2.4.2 and **Figure S2.7** that focused solely on the accuracy with which we could reproduce LCPO-based SASA values.

As shown in **Figure 2.6B**, the PMFs illustrating the free energy landscape profiles using

LCPO and our method agree with each other quite well (within ± 0.3 kcal/mol) when the same surface tension value is used for both methods. Using either model, increasing the surface tension results in less unfolded structures in the structural ensemble, which suggests that at least for this peptide, the nonpolar term plays a modulating role in hydrophobic core stability in implicit solvent.

There are still small local disagreements between our method and LCPO, at the scale of < 0.3 kcal/mol. These are reasonable for two reasons: (1) the SASA estimations for atoms and molecules are of somewhat different accuracies compared with the numerical references; (2) although the PMF uncertainties appear small when using the RMSD as collective variable, these may underestimate the true uncertainty in the data. In **Figure 2.7** we show an alternate convergence analysis in which the population of native-like structures (< 2.0 Å all $C\alpha$ -RMSD) is accumulated as a function of time for two independent REMD simulations for each of the two GBSA methods. Even after several microseconds of REMD, the fractions of native-like vary for LCPO by around 10% depending on the initial structure. Our pairwise model appears to converge more quickly than LCPO, but more extensive testing would be needed to determine the generality of this observation.

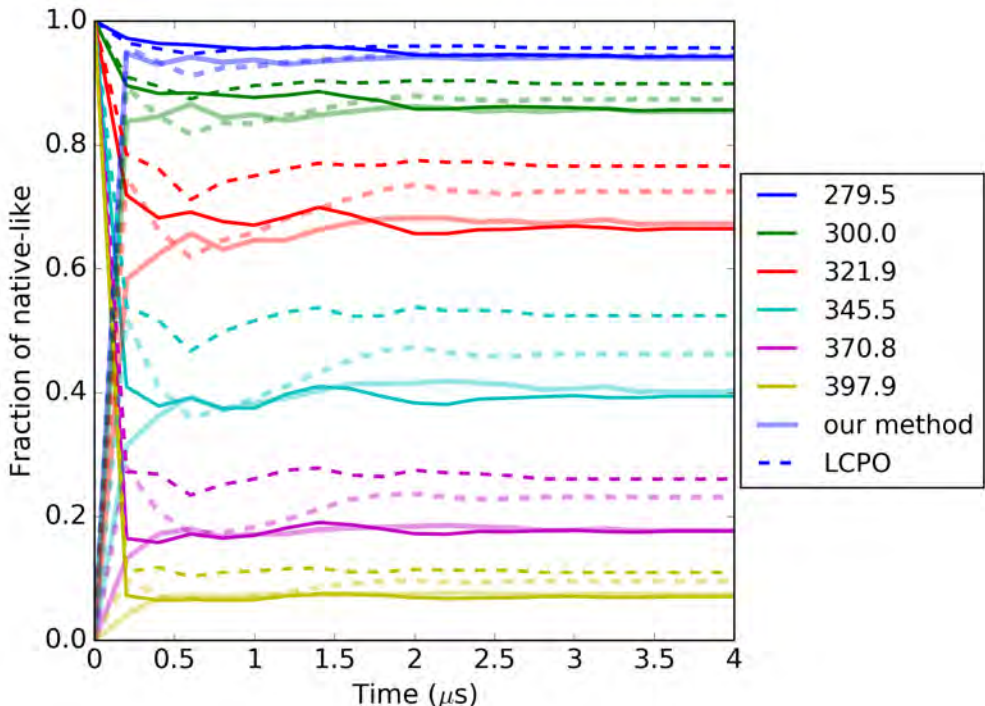


Figure 2.7: Fraction of folded calculated on HC16 for each temperature replica throughout the REMD simulations. Convergences from two different initial starting structures (NMR: opaque lines, unfolded: semi-transparent lines) are observed in our method (solid lines) and LCPO (dashed lines) using surface tension of $5 \text{ cal}/(\text{mol}\cdot\text{\AA}^2)$.

2.4.7 GB/SA solvation with reasonable surface tension can reproduce TIP3P profile

When nonpolar solvation energy is incorporated, GB/SA models could resurface the energy landscapes of HC16 structural ensembles towards the TIP3P result, by stabilizing the dominant native-like conformation while sampling less of the unfolded conformations (**Figure 2.6C**). A surface tension γ at 7 cal/(mol·Å²) is found to agree the best with TIP3P result. The choice of the calibrated surface tension is close to the value of 7.2 cal/(mol·Å²) used in MM/PBSA and MM/GBSA methods implemented in Amber as the Free Energy Workflow (FEW[106]); this is encouraging that the good agreement obtained with our method is not simply a result of empirical fitting.

Further consistency in GB/SA and TIP3P simulations is evident with closer examination of the PMF profiles shown in **Figure 2.6A**. When the nonpolar term is absent, the cluster 2 structure (4.1 Å C α -RMSD) with extended helix shown in **Figure 2.6C** has an occurrence of 15.0% (see occurrence data in **Table S2.5**) at 300 K, measured by within 2 Å from this 4.1 Å misfolded structure. This structural ensemble is not abundant in explicit solvent results, with occurrence < 0.2% in TIP3P ensemble. In GB/SA simulations, this misfolded structure is diminished to < 2% (in two GB/SA methods with $\gamma = 7$ cal/(mol·Å²)). But we also noticed that by increasing surface tension, in both LCPO algorithm and our method, another energy minima appears at around 5.4 Å in **Figure 2.6C** with close to 3% occurrence, with respect to < 0.2% in explicit solvent results. This 5.4 Å misfolded structure inversely orients the two helices of HC16 with misplacement of core Phenylalanine residues, and of relatively smaller SASA value. It is hard to attribute the cause as it could be a force field or solvent inaccuracy, or it may also be the convergence is still challenging in explicit solvent simulations.

2.4.8 Application to unrestrained proteins

Our algorithm provides a fast way to estimate the SASA of atoms and molecules in various conformations. Validated on a carefully controlled short peptide, we demonstrated that the nonpolar term is beneficial for core stability. With GPU compatibility, it is now possible to rapidly evaluate the extent to which a simple SASA-based nonpolar term can improve prediction of complex conformational ensembles. Such analyses on multiple systems were largely out of reach in the past due to the computational cost of SASA calculations on larger peptides and proteins during MD.

We included the GPU-compatible nonpolar solvation term while simulating the four proteins (CLN025, Trp-cage tc5b, HP36 and Homeodomain variant) without restraints. The simulated ensembles, with nonpolar term (our method and LCPO) or without (GB polar solvation only), were compared to experimental measures (CD or NMR). As always, one must use caution in such comparisons, since inaccuracies in the solute force field also impact agreement with experiment. Furthermore, the accuracy of the solvent models employed here is likely less reliable away from 300 K. Nevertheless, the trends in the data may provide useful insight within these limitations.

As shown in **Figure 2.8A**, compared with CLN025 GB-only simulations, conformational ensembles across the simulated temperature range show higher population of native-like

conformations using our SASA method, which also agrees reasonably with LCPO results. While still not as thermally stable as measured in CD[97], the improvement in stabilizing β -hairpin structures is encouraging; experimentally, the fraction of native folded hairpin is over 90% at 300 K, while it is less than 20% in our Amber ff14SBonlysc and GBNeck2 results without SASA. By incorporating the nonpolar solvation term, this value is elevated to around 70% in our method and around 80% in LCPO. This discrepancy between these two nonpolar methods corresponds to only around 0.30 kcal/mol, consistent with the differences observed for HC16. It is possible that better agreement could be obtained by increasing surface tension from 7 cal/(mol·Å²) to a larger value, but we decided to only test the value optimized using TIP3P as discussed above.

We next simulated Trp-cage tc5b, and again observed a significantly better agreement with experiment when the nonpolar term was added (**Figure 2.8B**). With GB/SA, we obtained near-quantitative agreement between our simulated Trp-cage tc5b and experimental thermal stability profiles. This further suggests that the ability to perform GB/SA with adequate sampling may significantly improve protein modeling efforts. At 300 K, our method and LCPO both accurately reproduce the experimental value of around 80%, compared to less than 30% fraction of folded as seen in the GB-only result. Our predicted T_m of 323K also is close to the experimental value of around 317 K[98]. This thermal stability of Trp-cage shows better accuracy than the GB-only model (predicted T_m at 283K) and other models, compared with predicted T_m down-shifted to 206K[107] using Charmm22* force field and modified TIP3P water model, or up-shifted to above 400K using ff94 force field and GB-HCT model[108], or OPLS-AA force field and TIP3P water[109].

When our pairwise SASA-based nonpolar term is incorporated in Homeodomain variant simulations, the increase of thermal stability with respect to GB-only is again observed (**Figure 2.8C**), although has not elevated to what experimental measurements are, compared to Trp-cage tc5b. Better agreement is possibly achievable with a larger surface tension, similar with CLN025. However, as the fold and topology of a protein gets more complex, ascribing the simulated thermal instability to the lack of nonpolar term alone needs a second thought; as more atoms and degrees of freedom are integrated in the simulations, the inaccuracies in models are likely to be magnified instead of being cancelled or concealed. And the errors in computational models could come from solvent models as well as force fields.

In the case of HP36 Villin headpiece, when only polar solvation with GB is included, at 300 K, less than 5% of conformations adopt folded structures (measured by fractions of conformations < 3.5 Å $C\alpha$ -RMSD excluding flexible termini), see Figure 2.8D. With our pairwise SASA-based nonpolar term, the stability of native-like conformations is predicted to be over 20% at 300 K, which is much higher. At 300 K, the two native-like conformations populated in GB trajectory that have occurrences of 1.36% and 1.70%, have been stabilized to be 18.1% and 14.1% shown as cluster 1 and cluster 3 in Figure 2.9A (see detailed measurements in Supporting Information **Table S2.6**). In GB-only results, the simulated melting curve has shifted to low temperature by around 100K (**Figure 2.8D**).

Interestingly, when the SASA term is added, the native-like structure is significantly increased in stability at 300 K and above, as we observed for CLN025 and Trp-cage, however, decreasing fractions of folded at 288.4 K and below result in a melting curve of a downward bell shape. Further analysis of the lower temperatures trajectories indicates that a misfolded

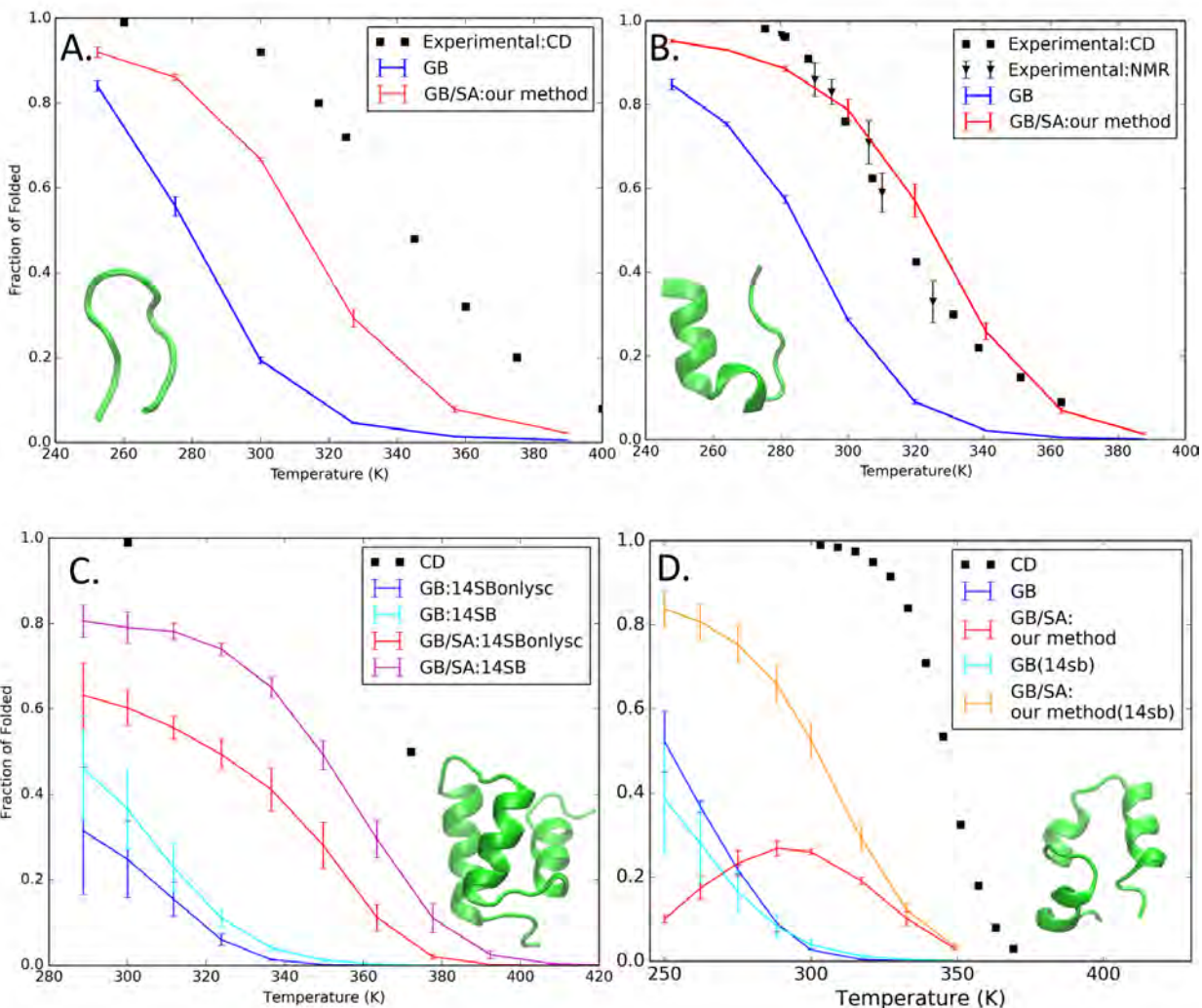


Figure 2.8: Thermal stability profiles for A. CLN025, B. Trp-cage tc5b and C. Homeodomain; D. HP36 respectively in GB and GB/SA REMD simulations, compared to experimental data.

structure become more dominant at lower temperatures, which reduces the fraction of folded. At 250 K, the previously populated native-like[24] structures (cluster 1, 30.4% and cluster 3, 18.3%) are diminished to 8.69% and 2.75% respectively in our GB/SA simulations (see **Figure 2.9B** and more details in **Table S2.6**). A misfolded structure ensemble instead occupies 75.9% of our GB/SA simulations shown as cluster 2 in **Figure 2.9B**. Compared to native-like structures, this 6.87 Å misfolded structure ensemble of smaller SASA has been stabilized by around 2 kcal/mol in potential energy contributed from nonpolar term (see data in **Table S2.6**). As a result, the native-like conformations are predicted as less favorable structural ensembles. Consequently, at all temperatures in REMD simulations, NMR structures unfold in the first hundreds of nanoseconds (**Figure S2.8**), which is not only observed with our method, but also with LCPO.

Two explanations for both GB/SA methods destabilizing the native structure of HP36 are explored and discussed. One explanation lies in force field as the SASA term may not

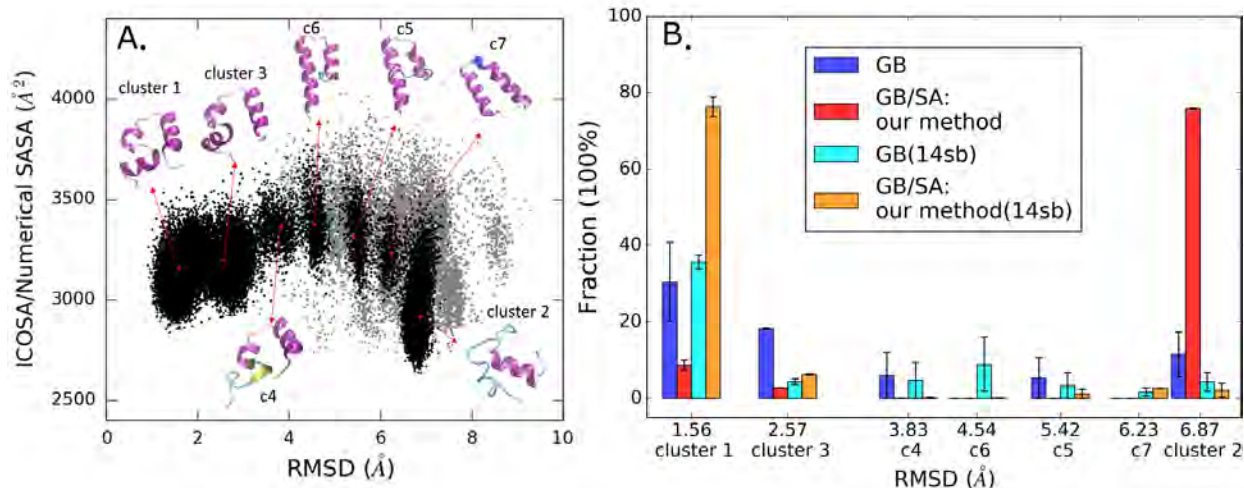


Figure 2.9: HP36 simulated structural equilibria using four models (GB with 14SBonlysc, GB/SA with 14SBonlysc, GB with 14SB, and GB/SA with 14SB). A. 2D scatter plot of our estimated SASA-based nonpolar energy SASA versus $C\alpha$ -RMSD excluding flexible termini against NMR structure for all structures in the combined 250 K trajectories simulated using four models. The structures clustered into top 7 populated clusters are black dots, and the rest structures are in gray. The top 7 cluster representatives are colored by secondary structures and illustrated with arrows pointing from their corresponding (RMSD, SASA) coordinates shown as red dots. B. Comparison of the top 7 cluster populations across four models. Each bar in the chart refers to the fraction (population) of a certain cluster in the simulated 250 K trajectory using a certain model. $C\alpha$ -RMSD excluding flexible termini values and the cluster order are denoted on the x-axis. The error bars are calculated from the first and second halves of trajectory.

be the cause of the error, despite the misfolded structure being lower in SASA. It is possible that the misfolded structure is an artifact arising from insufficient force field penalization. Another explanation is that the SASA-based nonpolar term fails to accurately recapitulate the missing nonpolar effect. The solute-solvent dispersive interactions might be indispensable for HP36 stability in simulations; as suggested by Gallicchio *et al.*[49], this dispersive term is almost independent of SASA but depend strongly on atomic composition.

Force field 14SBonlysc has been employed throughout all the training (HC16) and test cases (CLN025, Trp-cage and Homeodomain), as it was previously demonstrated to be capable of folding small proteins[24] with GBNeck2[62] implicit solvent. But its backbone parameters have not reached optimal secondary structure balances[23]. Empirical adjustment in the backbone ϕ parameter published as ff14SB[23] (i.e. ff14SBonlysc with mod1 ϕ backbone parameters), although trained to reproduce *Ala*₃ populations in TIP3P at 300 K[19], has been shown to stabilize the right-handed α and ppII dihedrals over β for all amino acids in TIP3P[23]. We thus applied ff14SB with GB and GB/SA implicit solvent to simulate HP36. As seen in **Figure 2.8D**, modified backbone force field does produce more consistent outcomes and backup our hypothesis that force field attributes to the unfolding of HP36 native structure in GB/SA solvation. Due to the lack of nonpolar term, GB-only simulations using force field 14SBonlysc 2 and 14SB predict similar melting behavior for HP36

in the simulated temperature range; the predicted T_m is around 100 K lower (346 K in experiment[100] versus < 250 K in GB) using both force fields. With SASA term, unlike the misfolded behavior using ff14SBonlysc, ff14SB is able to better maintain the balance in secondary structure propensities, elevate the stability of HP36 at all simulated temperatures and match better with experimental melting data[100]. **Figure 2.9B** illustrates and **Table S2.6** summarizes the predicted structural ensembles at 250 K and 300 K compared across four models. With SASA-based nonpolar term, folded structures (native-like or misfolded) are stabilized with close to 80% fractions, while more intermediate ensembles (cluster 4-7, denoted as c4, c5, c6 and c7 in **Figure 2.9A** and **2.9B**) populate at 250 K without this nonpolar term for both force fields. This in turn backups our overall hypothesis that non-polar solvation is not a negligible term in implicit solvation, instead, it could reshape the potential energy landscape thus it is crucial for accurate implicit solvation, as were observed in all the proteins demonstrated in this work.

Although clear weaknesses have been recognized, the SASA-based nonpolar model has been shown to work reasonably well with extensive parameterization against experimental solvation free energies of small nonpolar molecules[89, 110]. The application in biomolecules faces more challenges due to the uneconomic trade-off between computational cost and accuracy issues. Complete nonpolar solvation is a combination of solute-solvent dispersion energy (ΔG_{vdw})[50, 54], along with the hydrophobic effect and surface tension that depend on the size scale and shape[89, 111, 112, 113], the curvature[112], and temperature[113] of molecules. Methods to accurately calculate these contributions have not reached a consensus and are not readily calculated on GPUs to test impact on complex protein ensembles. But with the implementation of our new algorithm, despite its relative crudeness, the bottleneck in computational cost is overcome with order of magnitude accelerations for peptide and protein modeling. This can permit a greater exploration of success and failure cases for more complex biomolecules, possibly improving structure prediction and refinement, and also providing insight into future, more accurate nonpolar solvation models.

2.5 Conclusion

In this work, we proposed a fast, GPU-friendly pairwise SASA-based nonpolar solvation approach for protein simulations inspired by Amber’s pairwise GB solvation model[62, 66, 114, 115] development. In this approach, we developed a novel algorithm to estimate the atomic and molecular SASAs of proteins, which results in comparable accuracy as LCPO algorithm[82] in reproducing numerical ICOSA[77] SASA values. By calculating pairwise burial SASA from atom distances, our method accelerates MD simulations up to 30 times compared to LCPO implementation, with only around 20% overhead compared to CPU or GPU simulations that omit the SASA term. The main speed advance arises from employing GPU devices for SASA calculations and reducing constant communications with CPUs; the previous CPU/GPU implementation [104] using LCPO suffers from dramatic speed loss when the SASA calculation for every time integration step is still done on CPUs, even though all other energy terms are evaluated on the GPU[61]. Compared with other analytical approaches [79, 80, 82, 84, 85] including LCPO, our two-body algorithm is suitable for inexpensive gaming GPU devices that are built for highly parallelization calculations.

To ensure that our purely two-body algorithm is able to capture reasonably the SASA values in proteins, we pre-treat all protein atoms by grouping them into SASA types. This allows implicit incorporation of many-body contributions based on local geometry, and the remaining neighbor shielding is calculated using the non-recursive pair distances. Parameters were optimized on a peptide library covering all of the defined protein SASA types, sampling diverse conformations and SASA ranges. The objective function for training was designed to reproduce the SASA changes in atomic numerical ICOSA values, instead of the absolute atomic or molecular SASA numbers. The resulting 90 parameters are encoded in a new implementation as `gbsa=3` in Amber.

The evaluation of our nonpolar term and the calibration of surface tension are done in a helically restrained system which is derived from the hydrophobic core of HP36. This small peptide is also simulated in LCPO and explicit solvent TIP3P. Our method achieves similar outcomes as LCPO as well as TIP3P solvent when surface tension adopts $7 \text{ cal}/(\text{mol}\cdot\text{\AA}^2)$.

Four small proteins (CLN025, Trp-cage, Homeodomain and HP36) without restraints are simulated and compared to experimental results. The simulated melting curves for CLN025, Trp-cage and Homeodomain, with nonpolar term, are more consistent with experimental measures compared to without this term. Our method reasonably reproduces LCPO algorithm. In the case of HP36, it is more complicated. HP36 for both LCPO and our method, destabilize the NMR structure but switching to ff14SB rescues the situation, which points out the limitation in force field accuracy.

Compared with highly challenging convergence, much less sensitive thermal profiles[116] and misfolded structures[26] observed in explicit solvents, it is promising that with a non-polar term included, protein modeling in implicit solvent continues to be gaining in physical accuracy as well as increasing in speed. This is an important distinction since current protein simulations are typically limited by conformational sampling, rather than accuracy (especially for protein folding/misfolding, aggregation, IDPs and many more areas in which simulations could provide valuable insight).

2.6 Supporting Information

Definition of SASA types and parameters

We defined 30 atom types (just for SASA estimation, so we term them SASA types) based on one atom’s bonded heavy atoms and hybridization state. The nomenclature system of SASA type is 1 digit followed by several capital letters. The digit indicates the category the central atom falls into, depending on how many heavy atoms are bonded to the central atom or the just the group index. For example, for Carbon atoms, we categorize their bonding environments into 5 groups, group 1 means the central atom is single bonded to one heavy atom, group 2 means it is bonded to two heavy atoms. As for group 3, it also contains two heavy atoms bonded central carbons, but instead of a sp^3 hybridization as in group 2, the central carbon is double-bonded (or conjugated to) one or two heavy atoms, so that the bond lengths in group 3 are shorter than those in group 2, therefore less exposed to solvent are the central carbons in group 3 than those in group 2. Furthermore, in group 4, three

heavy atoms are bonded with the central carbon and in group 5, there are conjugated double bonds so that the central atoms in group 4 and 5 are categorized respectively. Depending on the element type of the heavy atoms, the 5 groups are further divided into sub-groups. All the detailed division and definitions are included in **Table S2.1** below.

Description of multiple rounds of optimization

The initial parameters for σ were randomized in the range of 2.7-3.6 Å, and the range for ε was 1.5-1.9 Å². When we attempted the optimizations by varying functional forms, m and n values, it was always the best performed parameter set saved and used as input for the next round of minimization; the final parameters were not bounded to the initial ranges. There were four rounds of optimizations, each searching for the best option for one thing. In the first two rounds, the functional form used was hyperbolic function, as was for the vdw dispersion energy. The objective function was molecular SASA and residual SASA. In the first round we used the hyperbolic functional form to optimize molecular SASA, starting at n=6 (same order as van der Waals dispersion term); with the converged parameter set outcome, the second round of optimization was made of 6 runs optimizing molecular SASA by varying the n value from 1 to 6, we found n=3 with a cutoff at 12 Å or n=4 with no cutoff performed the best; then we kept the two parameter sets, applied to MD simulations but found that we could not reproduce LCPO results. Then we changed to the current functional form, aiming to minimize the atomic SASA differences for another round. With the two sets from previous fitting, one of the resulting parameter set assigned 0 to C α atoms (4CCN SASA type), the simulation results are not as effective either, while in the other set, C α atoms contribute to pairwise *shield_SASA*, and LCPO could be reproduced effectively, so we selected this parameter set.

The derivation of shield_SASA formula

When m=12, n=6, the van der Waals Lennard-Jones potential is an expression as below:

$$formula(vdw) = \varepsilon_{i,j} \left(\left(\frac{\sigma_{i,j}}{R_{i,j}} \right)^{12} - \left(2 \frac{\sigma_{i,j}}{R_{i,j}} \right)^6 \right) \quad (S2.1)$$

But the more general form is as below:

$$formula(vdw - like) = \varepsilon_{i,j} \left(\frac{\frac{n}{m-n} \sigma_{i,j}^m}{(R_{i,j})^m} - \frac{\frac{m}{m-n} \sigma_{i,j}^n}{(R_{i,j})^n} \right) \quad (S2.2)$$

A, B, and C steps correspond to the transformations described in **Figure S2.1** below:

$$formula(after A) = \varepsilon_{i,j} \left(\frac{\frac{n}{m-n} \sigma_{i,j}^m}{(-R_{i,j})^m} - \frac{\frac{m}{m-n} \sigma_{i,j}^n}{(-R_{i,j})^n} \right) \quad (S2.3)$$

$$formula(after B) = \varepsilon_{i,j} \left(\frac{\frac{n}{m-n} \sigma_{i,j}^m}{(cutoff_{f_{i,j}} + \sigma_{i,j} - R_{i,j})^m} - \frac{\frac{m}{m-n} \sigma_{i,j}^n}{(cutoff_{f_{i,j}} + \sigma_{i,j} - R_{i,j})^n} \right) \quad (S2.4)$$

$$formula(afterC) = \varepsilon_{i,j} \left(\frac{\frac{n}{m-n} \sigma_{i,j}^m}{(cutoff_{i,j} + \sigma_{i,j} - R_{i,j})^m} - \frac{\frac{m}{m-n} \sigma_{i,j}^n}{(cutoff_{i,j} + \sigma_{i,j} - R_{i,j})^n} \right) + \varepsilon_{i,j} \quad (S2.5)$$

Then for the fractions $\frac{\frac{n}{m-n} \sigma_{i,j}^m}{(cutoff_{i,j} + \sigma_{i,j} - R_{i,j})^m}$, to divide $\sigma_{i,j}^m$ simultaneously for the denominator and numerator, and $\frac{\frac{m}{m-n} \sigma_{i,j}^n}{(cutoff_{i,j} + \sigma_{i,j} - R_{i,j})^n}$, to divide $\sigma_{i,j}^n$ for the denominator and numerator, we will get:

$$\varepsilon_{i,j} \left(\frac{\frac{n}{m-n}}{\left(1 + \frac{cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^m} - \frac{\frac{m}{m-n}}{\left(1 + \frac{cutoff_{i,j} - R_{i,j}}{\sigma_{i,j}}\right)^n} \right) + \varepsilon_{i,j} \quad (S2.6)$$

which is **Equation 2.5** on Page 19 when $R_{i,j} < cutoff_{i,j}$.

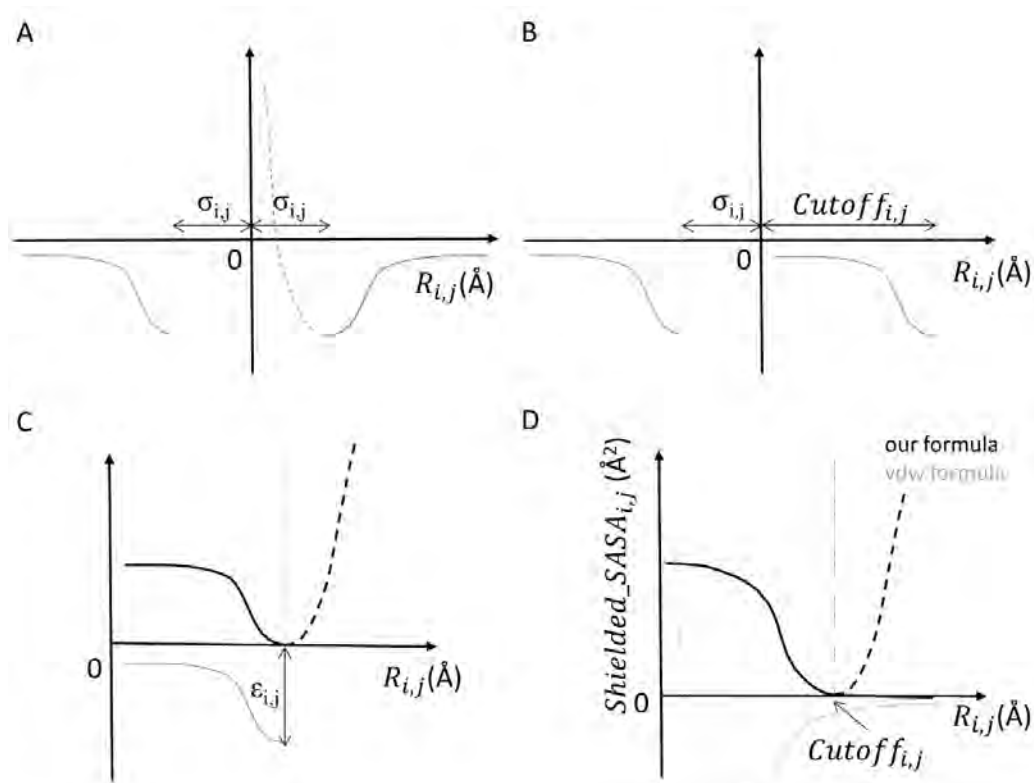


Figure S2.1: Transformation of our formula (Equation S6) from vdw function (Equation S1, more general form Equation S2) in schematic representations. A. starting from vdw function (only the beyond vdw radius part is kept, shown in solid gray line on the right side of the y axis), to reflect it by y-axis results in Equation S3; B. right shift it by $\sigma_{i,j} + cutoff_{i,j}$ results in Equation S4; C. up shift the curve by $\varepsilon_{i,j}$ results in Equation S5, which is the $(0, cutoff_{i,j})$; D. a comparison of our final formula and vdw formula.

Table S2.1: Defined 30 SASA types and their occurrences in the training and test sets

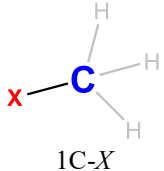
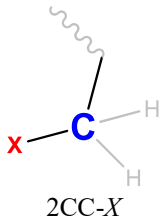
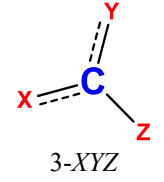
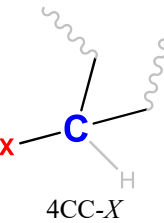
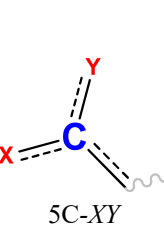

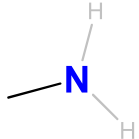
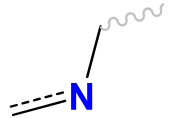
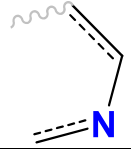
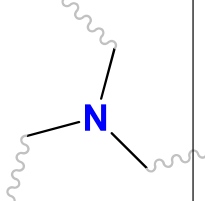
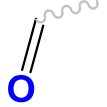
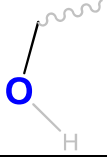
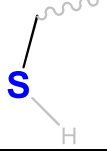
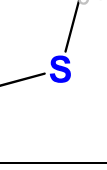
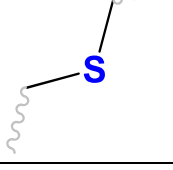

Element	Hybridization	Generic formula	SASA type	Locations	Atom No. in training set	Atom No. in test set	Atom radius
Carbon	sp ³	 1C-X	1CC	Ala side chain	80	444	1.7
			1CN	NME	10	0	
			1CS	Met side chain	10	18	
		 2CC-X	2CCC	Arg, Lys, Pro, Trp, Tyr, Phe, His side chain	220	914	
			2CCN	Arg, Lys, Gly, Pro side chain	40	220	
			2CCO	Ser side chain	10	36	
			2CCS	Cys, Met side chain	20	18	
	sp ²	 3-XYZ	3CC	Tyr, Phe, Trp side chain	130	404	
			3CCC	Thr, Phe, Trp side chain	40	101	
			3CCN	His, Trp side chain	40	24	
			3CCO	Tyr side chain	10	29	
			3CNN	His side chain	30	12	
	sp ³	 4CC-X	4CCC	Ile, Leu, Val side chain	30	155	
			4CCN	all backbone C α except Gly	210	787	
			4CCO	Thr side chain	10	46	
	sp ²	 5C-XY	5CCN ₁	Trp side chain	10	12	
			5CCN ₂	His side chain	30	12	
			5CNN	Arg side chain	10	44	
			5CNO	Backbone and Asn, Gln side chain carbonyl	240	915	
			5COO	Terminal carbonyl	20	133	
Nitrogen	sp ³		1NC1	Arg, Asn, Gln side chain	40	165	1.55

Table S2.1: — continued from previous page

			1NC2	Terminal amide, Lys	20	100	
	sp2, aliphatic		2NCC	Arg side chain, backbone amide	220	858	
	sp2, aromatic		3NCC	His side chain	70	36	
	sp3		4NCC	Pro backbone amide	10	24	
Oxygen	sp3		1OC1	Backbone and deprotonated carbonyl	280	1181	1.5
			1OC2	Ser, Thr side chain hydroxyl	20	82	
			1OC3	Tyr side chain hydroxyl	10	29	
Sulfur	sp3		1SC	reduced Cys	10	0	1.8
			2SCC	Met side chain, Cys in disulfide bonds	10	18	
Hydrogen	N/A		1H	all hydrogens	1780	6813	0*

*Zero radii are set for the Hydrogen atoms only for SASA calculations.

Table S2.2: Optimized (sigma and epsilon) and calculated (cutoff and max_SASA) parameters

SASA type	Cutoff (Å)	σ (Sigma)	ϵ (Epsilon)	Max SASA
1CC	3.1	4.370116	19.592480	89.5418522949
1CN		0.317054	7.148281	93.3032786658
1CS		1.208319	14.405034	101.172789338
2CCC		7.249659	18.793198	67.848137034
2CCN		5.492838	19.214204	78.9232674787
2CCO		1.568006	4.738056	62.3149039685
2CCS		3.856010	13.792617	72.0652500364
3CC		5.523807	13.179907	72.2402965204
3CCC		7.925637	0.700405	14.8528474421
3CCN		1.137401	4.914482	70.6469194565
3CCO		2.852471	0.690433	20.2921312847
3CNN		4.793021	17.786058	85.8858602953
4CCC		2.463345	1.586756	29.2869985239
4CCN		0.100000	0.328277	22.5608806495
4CCO		1.842881	1.600013	33.1334096093
5CCN1		3.532759	0.371097	16.5093987597
5CCN2		0.902828	0.021241	6.90712731848
5CNN		6.516442	3.681840	32.0850765017
5CNO		5.997082	0.739936	16.1244401843
5COO		9.776595	1.438099	16.6436516422
1NC1	2.95	3.008485	23.511977	94.0970695867
1NC2		4.290955	34.274575	95.1108847805
2NCC		3.296031	0.919202	22.8610751485
3NCC		5.589998	16.263937	64.7488801386
4NCC		1.0 ^β	0.000000 ^γ	0.180783832809
1OC1	2.9	6.764858	12.670634	58.3692979586
1OC2		2.827230	11.236117	69.8105657286
1OC3		2.827230	11.236117	80.1047057149

Table S2.2: — continued from previous page

1SC	3.2	2.520362	16.788985	105.113824567
2SCC		1.133725	5.828670	75.8598197695
1H	1.4 ^a	1.0 ^b	0. ^c	0.

^a Zero radii are set for the Hydrogen atoms, so the cutoff is always the probe of water radius 1.4 Å.

^b Sigma = 1.0 for hydrogen and 4NCC are to make sure the denominator is not 0 in Equation 6.

^c Epsilon = 0 enforces zero contribution in shield SASA for all hydrogen and 4NCC involved atom pairs.

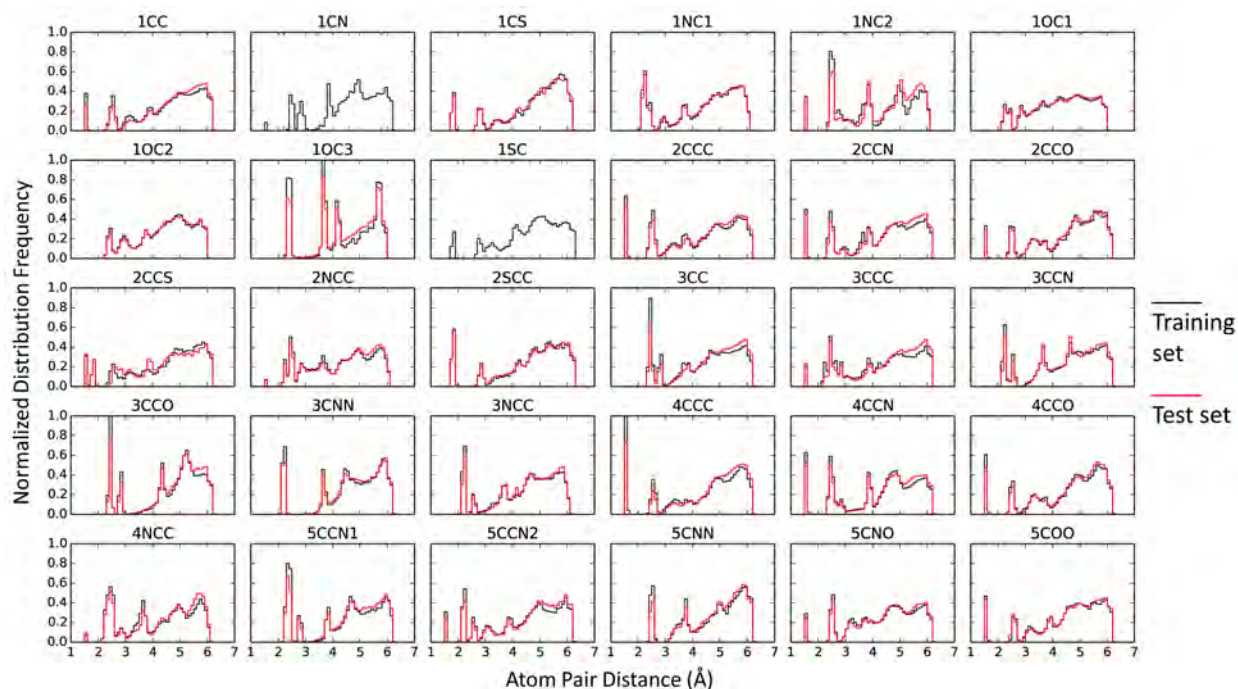


Figure S2.2: Distribution of pairwise atom distances within corresponding cutoffs for each SASA type atoms for training set peptides and test set proteins

Table S2.3: Sequences and conformational features in scrambled peptide training set







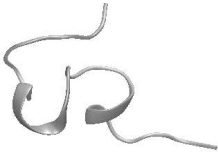



scrambled sequence index	Sequence	Secondary Structure	largest cluster (percentage)
1	RAH ^{δe} TH ^δ GYKMDNP EQIH ^e LFWCVS-NME	antiparallel, α -helix, coil	 (17.2%)
2	RWMCDVAGIH ^e ENL TPH ^{δe} SKH ^δ QYF-NME	α -helix, coil	 (13.8%)
3	ENLVAFPITWYQH ^δ H ^e RMCKDGS ^{δe} -NME	α -helix, coil	 (45.4%)
4	NVWPECH ^{δe} LQYDTI H ^e FH ^δ ASKRGM-NME	α -helix, coil	 (38.4%)
5	FMIH ^δ SEH ^{δe} CLWH ^e Q ANRKGTVDYP-NME	antiparallel, turn	 (11.7%)
6	FKH ^δ AH ^{δe} ECQH ^e RGLI VPSMYNTDW-NME	α -helix, coil	 (19.1%)
7	YIKQPSDFVWL ^e GTH ^e NAH ^δ EMCRH ^{δe} -NME	α -helix, coil	 (23.2%)
8	LDKH ^e AGH ^{δe} VSREFI H ^δ TWNQCMYP-NME	α -helix, coil	 (23.8%)

Table S2.3: — continued from previous page

9	FH ^δ RLQMDKEYNPS GAWIH ^{δϵ} TCVH ^ϵ -NME	antiparallel, turn	 (82.3%)
10	EDKLH ^ϵ ASRPH ^δ WYV H ^{δϵ} CFMTQNGI-NME	α -helix, turn, coil	 (15.8%)

Note: H^δ, H^ϵ, H^{δϵ} are Histidine that is protonated at N^δ, N^ϵ, or both N^δ and N^ϵ, respectively. This training set has been developed in the experimental state of pairwise SASA algorithm, at that time hydrogens were considered in the SASA calculations; but later as we decided to exclude hydrogens in SASA estimation, the protonation states do not make any difference.

Table S2.4: Temperature ladders for all REMD simulations.

System	Solvent model	REMD temperatures (K)
HC16	TIP3P	266.7, 270.2, 273.8, 277.4, 281.0, 284.7, 288.5, 292.3, 296.1, 300.0, 303.9, 307.9, 312.0, 316.1, 320.3, 324.5, 328.8, 333.1, 337.5, 341.9, 346.4, 351.0, 355.6, 360.3, 365.0, 369.8, 374.7, 379.6, 384.6, 389.7, 394.8, 400.0
	GB/SA (our method, LCPO)	279.5, 300.0, 321.9, 345.5, 370.8, 397.9
CLN025	GBNeck2, GB/SA (our method, LCPO)	252.3, 275.1, 300.0, 327.2, 356.8, 389.1
Trp-cage		247.7, 264.0, 281.4, 300.0, 319.8, 340.9, 363.3, 387.3
HP36	GBNeck2, GB/SA (our method, LCPO)	250.0, 262.2, 275.0, 288.4, 300.0, 317.3, 332.8, 349.0
Homeodomain	GBNeck2, GB/SA (our method)	288.7, 300.0, 311.7, 323.9, 336.6, 349.8, 363.5, 377.7, 392.4, 407.8, 423.8, 440.3

Table S2.5: The properties of the top 3 cluster representative structures from cluster analysis^a on HC16 GB simulations (300 K) and their occurrences in other simulations (300 K trajectories from GB/SA and TIP3P)

Clustering and occurrence Analysis		Cluster 1	Cluster 2	Cluster 3
Representative structure C α -RMSD (Å)		1.00	4.12	5.37
Representative structure ICOSA SASA (Å ²)		1686.4	1911.7	1813.9
Representative structure SASA-based (Y=7cal/mol·Å ²) nonpolar energy (kcal/mol)		11.8	13.4	12.7
Cluster population in GB (%)		57%	15%	5.3%
Occurrence ^b in each trajectory at 300K	GB	57.9%	15.0%	4.89%
	GB/SA: our method	91.4%	1.81%	2.95%
	GB/SA: LCPO	93.2%	1.37%	2.27%
	TIP3P	95.7%	0.109%	0.155%

^a For clustering analysis done on GB trajectories, 16000 frames in total are evenly obtained from the last halves of the two MD simulations starting from different initial structures, clustering criterion is pairwise RMSDs based on all C α atoms, using bottom-up aggregating average linkage algorithm, centroid distances < 2.0 Å.

^b The occurrence of certain cluster in GB/SA (Y=7cal/mol·Å²) or TIP3P trajectories was measured by the number of conformations that are < 2.0 Å (C α -RMSD) from the representative structure of this cluster, divided by the total frame number of the whole simulated trajectory at 300K.

Table S2.6: Cluster analysis for HP36 combined trajectory at 250 K and occurrences of the top 7 cluster representative structures in the four 300 K trajectories, respectively

Clustering and occurrence Analysis		Cluster 1	Cluster 2	Cluster 3	c4	c5	c6	c7
Representative structure C α -RMSD (Å)		2.401	7.677	3.031	4.696	6.169	5.474	7.209
Representative structure C α -RMSD on structured region 3-32 (Å)		1.565	6.873	2.569	3.829	5.416	4.545	6.233
Average C α -RMSD on structured region 3-32 (Å)		1.957 (0.543)	6.790 (0.136)	2.712 (0.310)	3.739 (0.267)	5.564 (0.226)	4.570 (0.148)	6.236 (0.151)
Average ICOSA SASA (Å ²)		3155.7 (105.8)	2914.9 (132.4)	3195.8 (113.3)	3365.7 (85.4)	3320.0 (132.1)	3376.7 (124.0)	3230.2 (110.6)
SASA-based (Y=7cal/mol·Å ²) nonpolar energy (kcal/mol)		22.1 (0.7)	20.4 (0.9)	22.4 (0.8)	23.6 (0.6)	23.2 (0.9)	23.6 (0.9)	22.6 (0.8)
Fraction (cluster population) in each trajectory at 250K (%)	GB	30.4 (10.4)	11.6 (5.82)	18.3 (0.15)	5.97 (5.95)	5.34 (5.34)	0.00 (0.00)	0.00 (0.00)
	GB/SA: our method	8.69 (1.25)	75.9 (0.21)	2.75 (0.05)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)
	GB (14sb)	35.7 (1.73)	4.31 (2.49)	4.36 (0.72)	4.72 (4.72)	3.38 (3.38)	8.92 (7.1)	1.67 (1.07)
	GB/SA: our method (14sb)	76.3 (2.51)	2.12 (1.8)	6.27 (0.19)	0.16 (0.16)	1.20 (1.20)	0.14 (0.04)	2.6 (0.06)
Occurrence [‡] in each trajectory at 300K (%)	GB	1.36 (0.13)	0.70 (0.43)	1.70 (0.20)	0.57 (0.12)	0.23 (0.23)	0.05 (0.05)	0.007 (0.007)
	GB/SA: our method	18.1 (0.33)	26.9 (1.52)	14.1 (0.59)	0.43 (0.01)	0.089 (0.081)	0.00 (0.00)	0.038 (0.034)
	GB (14sb)	2.81 (0.39)	0.44 (0.27)	2.16 (0.33)	0.74 (0.73)	0.22 (0.22)	1.61 (0.24)	0.30 (0.11)
	GB/SA: our method (14sb)	47.4 (0.25)	1.64 (1.34)	30.6 (0.18)	0.54 (0.43)	1.35 (1.32)	0.48 (0.18)	3.11 (0.023)

^a Clustering analysis was done on combined trajectory of the four (GB, GB/SA: our method, GB(14sb), GB/SA: our method (14sb)) methods. In total 40,000 frames (10,000 frames from each trajectory) are evenly obtained from 250K trajectories, clustering criterion is pairwise RMSDs based on structured region (residue 3 to 32 C α atoms), using bottom-up aggregating average linkage algorithm, centroid distances < 2.0 Å.

[‡] The occurrences are measured in the similar fashion as in Table S5, i.e. all frames < 2.0 Å (C α -RMSD in region 3-32) from the representative structure of this cluster, divided by the total frame number of the whole simulated trajectory at 300K.

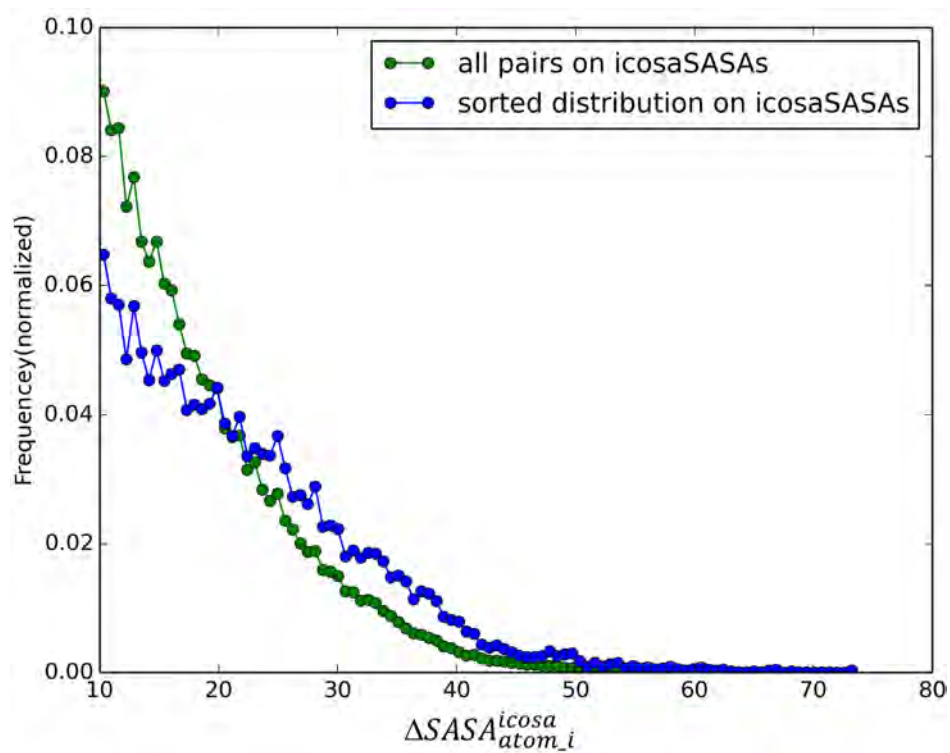


Figure S2.3: Normalized distribution of $\Delta SASA_{atom_i}^{icosa}$ including all frame pairs or sorted frame pairs for training set peptide atoms

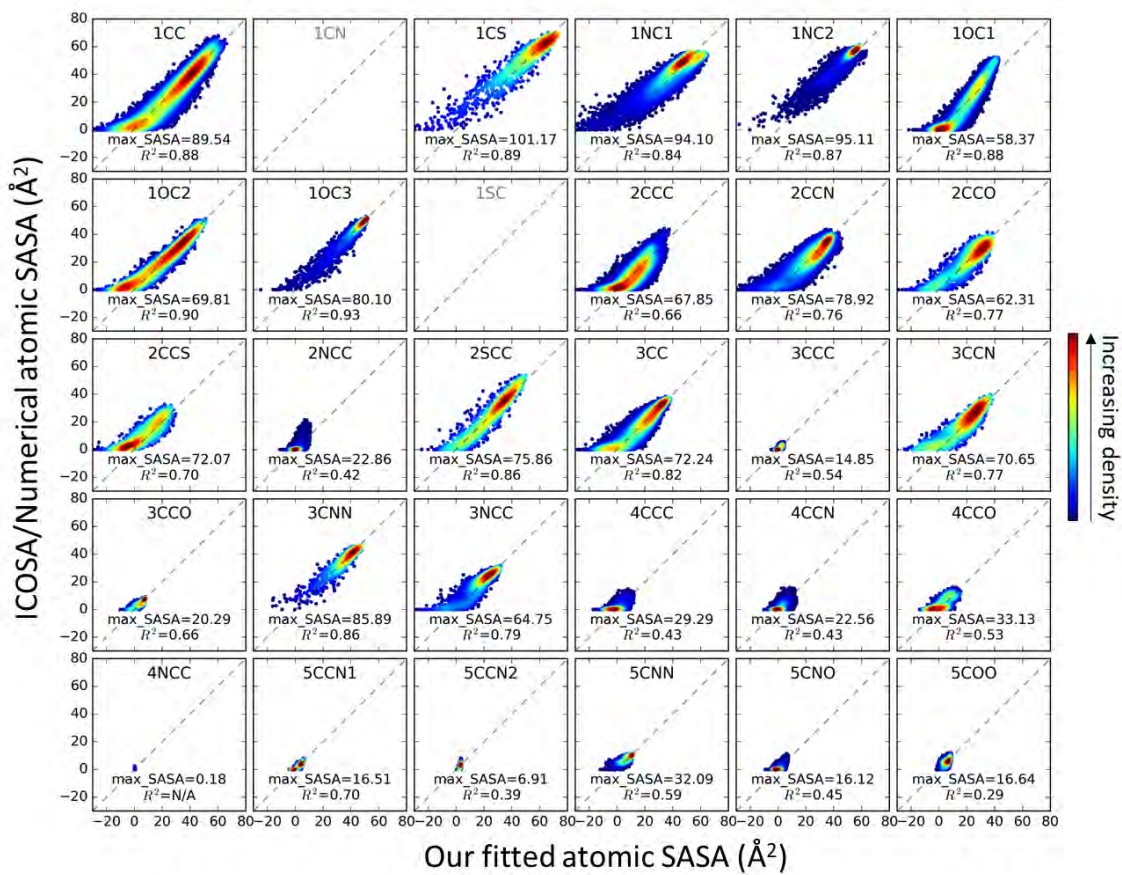


Figure S2.4: 2D histograms of fitted atomic SASA of each SASA type versus ICOSA numerical values for the test set. Perfect agreement is shown by the diagonal dashed lines. The color indicates the kernel density estimated using `scipy.gaussian_kde`.

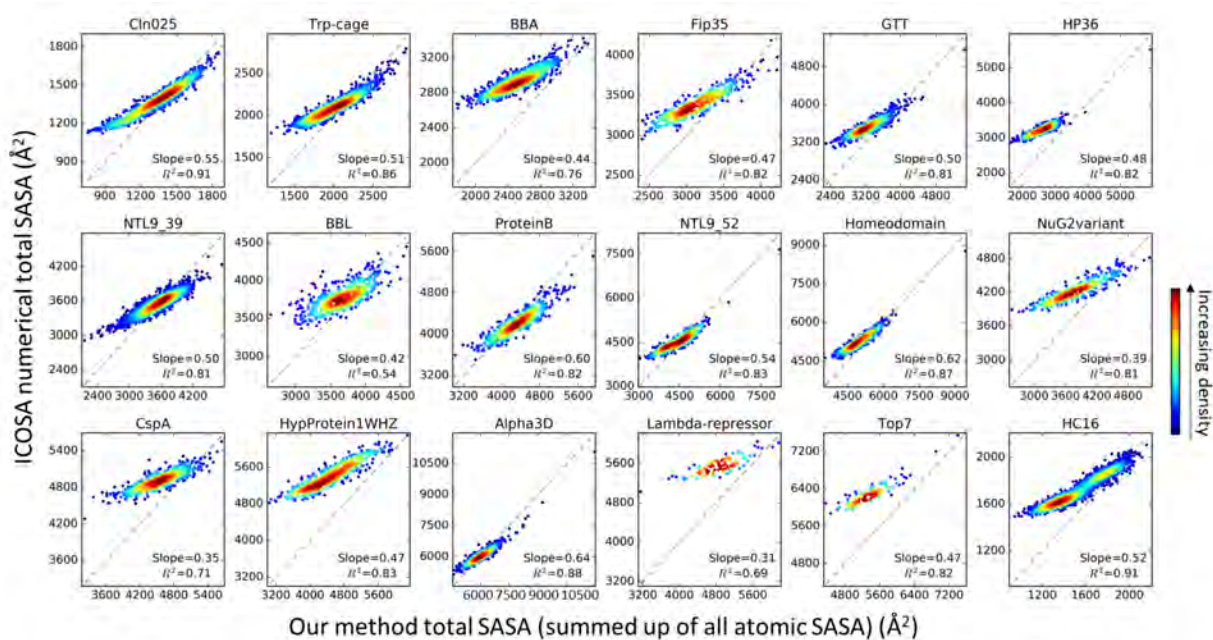


Figure S2.5: Deviation of sum of atomic SASA from the numerical SASA, represented in 2D histograms of total SASA versus ICOSA numerical values for the test set.

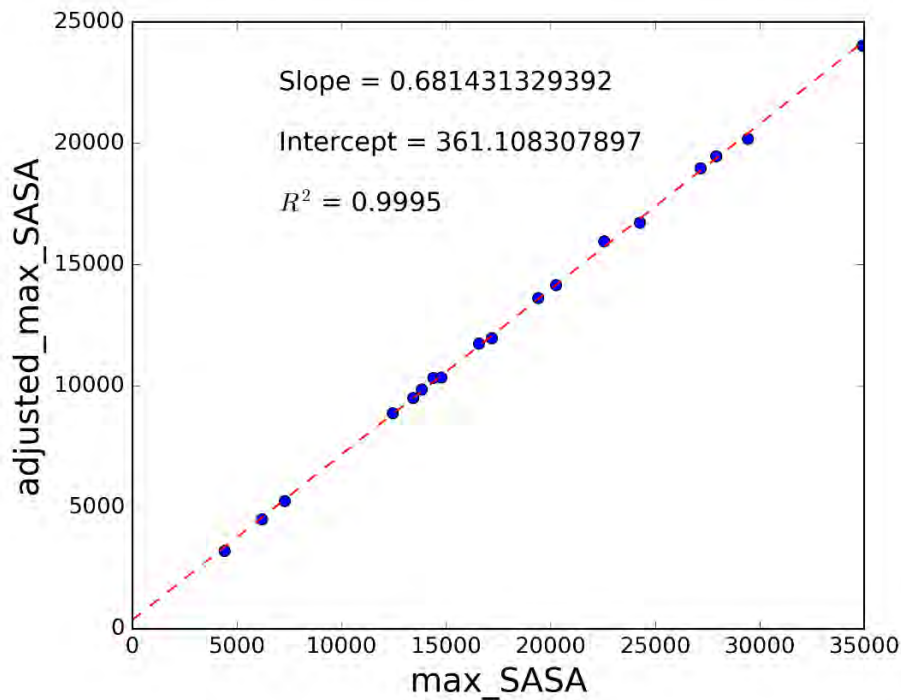


Figure S2.6: Transformation of max_SASA to adjusted_max_SASA by linear regression. Each data point corresponds to a protein in the test set.

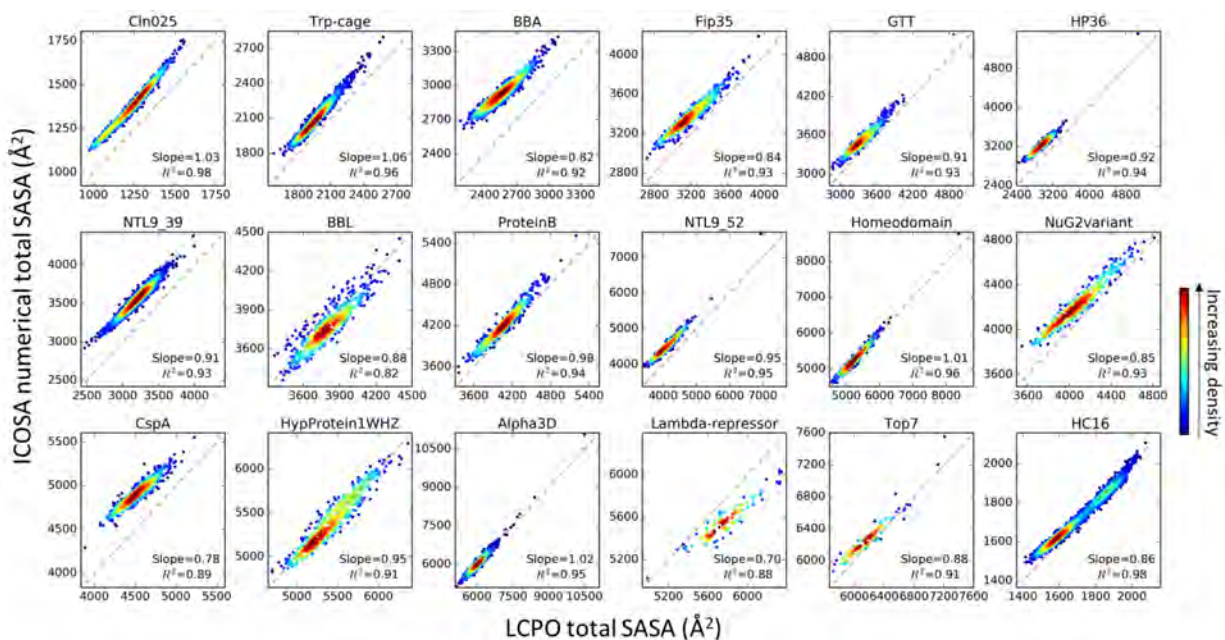


Figure S2.7: 2D histograms of LCPO fitted molecular SASA of each SASA type versus ICOSA numerical values for the test set.

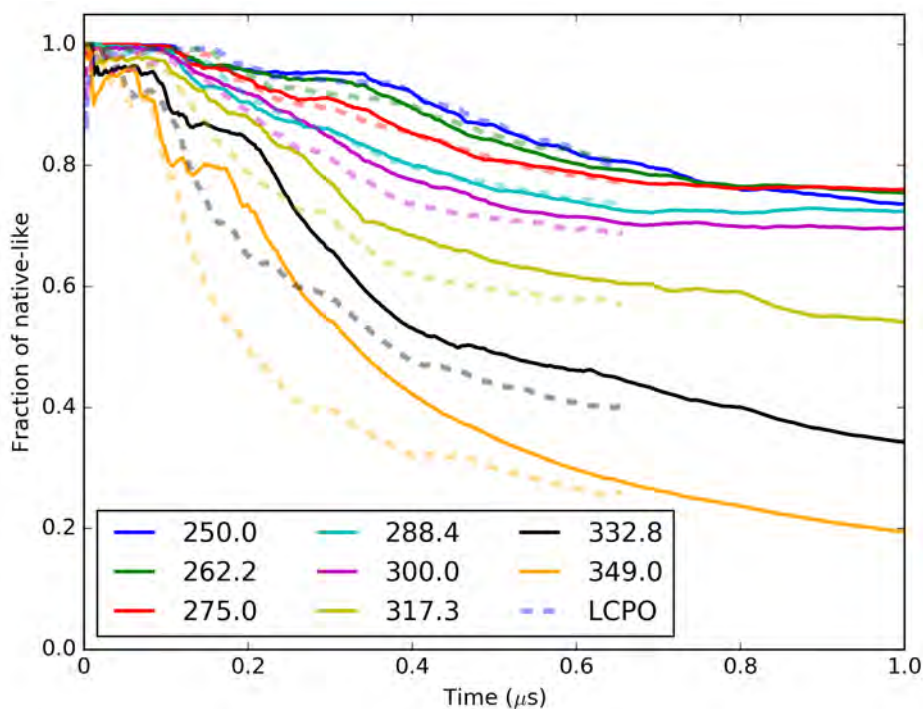


Figure S2.8: Unfolding of HP36 NMR structure observed in two GB/SA solvent simulations using ff14SBonlysc: our method (solid lines) and LCPO (dashed lines). The decrease of fraction of folded is calculated ($< 3.5 \text{ \AA}$ $\text{C}\alpha$ -RMSD excluding flexible termini) for each temperature replica throughout the REMD simulations. Only the first $1 \mu\text{s}$ of data is shown.

Chapter 3

Toolbox Development for Amino Acid Specificity Evaluations

3.1 Abstract

To validate the accuracy of computational model and their backbone amino acid specificity, we developed a novel secondary structure dihedral stability test toolbox and applied another helical propensity toolbox. In this chapter, we demonstrated the evaluation and comparison of two force fields on their capabilities in reproducing crystal structure dihedral angles and helical propensity measured from chemical shift, respectively. The findings cross validate our understanding in protein backbone parameter modifications and point out the necessities of improving amino acid specificity in the current model. The toolboxes and the same methodology could be applied to future studies.

3.2 Introduction

The question of whether current computational models are able to capture the amino acid specificity has been brought up and the significance has been addressed in Introduction 1.3.2. But for model validation and accuracy issue diagnosis, first and foremost, we need a set of toolboxes in which we could compare specific simulation results to benchmark measurements.

To develop and validate the protein modeling force fields and solvent models, quantities calculated from simulations are often trained to reproduce QM data or compared against experimental data collected at consistent conditions. In the earlier Amber force fields including ff94, ff96, ff99 and ff99SB, backbone parameters were fit against QM *ab initio* energy minima of amino acid fragments or short peptides, where QM energies were from gas-phase calculations. More recently, empirical comparison and fitting against experimental data have been applied to modify force field parameters, such as RSFF1, RSFF2, ff99SB*, ff99SBnmr, ff14SB, etc.

Crystallographic data provides a large, meanwhile growing, amount of structure information for proteins and their complexes. To validate protein modeling tools, model systems such as ubiquitin[20], toxin II protein crystal lattice[17] and others[21] have been tested in short simulations and used to evaluate the accuracy of corresponding force fields or solvent

models. But the evaluation conclusions, drawn from one or two proteins, might still lack transferability to other protein contexts. In contrast to the other extreme, large-scaled statistical approaches made the assumption that each amino acid would follow the standard energy profiles averaged from all of its kind over a large amount of structure distributions. For example, the top500 high-resolution crystal structure database, created and curated by the Richardson lab[117], had been used to derive statistical potential for force field potential energy evaluations[20]. Also protein coil libraries, originally initiated by Swindells *et al.*[118], have motivated the studies in intrinsic conformational propensities of amino acids and have been used for force field development[75, 68]. However, the evaluations compared against the PDB distribution[20] or coil library[75, 119] suffer from the lack of solid thermodynamic foundation or the limitations (inconsistent temperatures, crystal packing etc.) in data quality.

Therefore, it is worth trying to evaluate the reproducibility within contexts comparing against high quality crystallographic data: (1) to capture the experimental behavior of proteins in one context in simulations and test how well the force field/solvent model combination is in reproducing this one context; (2) to scale up to 30 proteins (i.e. 30 sequences of contexts) and collect the statistics for each amino acid. The secondary structure preferences of backbone dihedrals adopted in crystal structures are used as the criterion for deciding whether the experimental behavior is reproduced, although proteins are dynamic and crystal structures may only provide an averaged picture. We scale up the statistical analysis to tens of proteins after each of them has been compared with the corresponding crystal structure. This is different from directly compare against the PDB data base distributions, as each sequence is studied within its own context.

55 High Quality proteins (HiQ54+) data set was designed and used as benchmarks to improve Rosetta energy functions [120]. It consists of 55 non-redundant, single-chain and monomeric proteins from PDB through 2010. Without any significant errors such as bond-length/angle outliers, these crystal structures all have resolution smaller than 1.4 Å. Culling out the systems with ion or small molecule binders referring to several criterion described in Method, the selected 27 proteins are termed Hi27 data set. By working with the high quality crystal structures, the factors that are usually ascribed to protein instability in simulations, such as possible dimer interface, other necessary binders or crystal packing effect, could be excluded. The secondary structure balances achieved in our short simulations thus should be used to validate force field and solvent model accuracy. In addition, the effects of solvent model and backbone parameters are separated by altering just one thing at a time; we compare the secondary structure propensity change where just solvent model or force field backbone parameter set is changed. As to change solvent from explicit to implicit is too big of an alternation, in our analysis, explicit solvent results could work as the control.

Spectroscopic methods also provide a major resource of quantities that could be measured from aqueous state experiments and calculated from simulations. Scalar coupling is used in protein force field developments and testing[121, 23]; the NMR spin-spin coupling constants are related to protein backbone angle distributions by the Karplus equation[122]. $^{13}\text{C}\alpha$ chemical shifts from NMR measurements have also been employed to indicate the helicity of peptides[123, 124] and the secondary structure balance in different force fields[125, 71]. Backbone Amide NH Lipari-Szabo S^2 order parameters derived from NMR relaxation experiments are also useful quantities reflecting the backbone motions and dynamics[126, 20, 18, 24, 23].

Other approaches such as IR and Raman spectra data was used to determine the relative populations of the three major backbone conformations[127] and potentially could be referred to for force field validations.

For helical structure balance and amino acid specificity tests, we followed what was done by Best *et al.*[125] and then Perez *et al.*[71]. We compared the helical propensities calculated from simulations to a set of experimental measures derived from $^{13}\text{C}\alpha$ chemical shifts[124]. This set of experimental data provide helical propensity for all 20 amino acids. Although only helical propensities are calculated and compared to directly, it is an important data set in reflecting the intrinsic balance preferences and amino acid specificity of our computational models.

3.3 Methods

3.3.1 System setup

HiQ27 data set selection from HiQ54

27 out of the total 54 monomeric proteins were picked out as eligible systems for computational model evaluation. In the complete HiQ54 dataset, although all 54 proteins are of high quality ($< 1.4 \text{ \AA}$) and no multimer interfaces, several other criteria were checked before including certain protein into our toolbox: (1) no ions or small molecule binders, (2) no missing residues, (3) no non-standard amino acids, (4) no obvious crystal packing effect by visually examining the neighboring asymmetric units reconstructed using PyMOL[128]. After culling off the systems that conflict either criterion stated above, 27 proteins were selected into our toolbox, termed HiQ27. All the initial structures were then optimized by Reduce[129] (Asn/Gln/His flips corrected) and His protonation state predicted by H++[130]. **Table 3.1** summarizes the names, PDB codes and secondary structure features of all 27 proteins (sorted by numbers of amino acids).

Table 3.1: Protein crystal structures selected into HiQ27 data set

Protein Name	#PDB_ID	Ref	AA	Secondary Structure
protein G domain	2igd	[131]	61	alpha/beta (NuG2 Variant 1MI0 was simulated by us, pH=8, T=298K)
scytovirin lectin	2qsk	[132]	95	mostly coil
PDZ dom of NumB-bdg prot2	2vwr	[133]	95	alpha/beta
RNase Sa T76W	1t2i	[134]	96	alpha/beta
starch-bdg dom, glucoamylase	2vq4	[135]	106	mostly beta
A aceti thioredoxin	2i4a	[136]	107	alpha/beta
FK506-bdg prot 12	2ppp	[137]	107	alpha/beta
cardiac myosin-bdg	3cx2	[138]	107	antip.beta
Phox hom dom, P-inos-3-K C2-g	2wwe	[139]	111	alpha/beta tail

Continued on next page

Table 3.1 – continued from previous page

Protein Name	#PDB.ID	Ref	AA	Secondary Structure
shark single-dom antibody	2i24	[140]	113	antip.beta
sulfite red'ase DsrC	1sau	[141]	114	mostly alpha, one beta hairpin
FK506-bdg dom of FKBP38	3ey6	[142]	118	antip.beta
pak pilin, trunc	1x6x	[143]	120	alpha/beta
barley bowman-birk inhibitor	2fj8	[144]	120	mostly coil, some beta
P-lipase A2 homolog	1mc2	[145]	122	mostly alpha, one beta hairpin
RNase A	1kf5	[146]	124	alpha/beta mix (multiple pH structures are available)
bovine H protein	3klr	[147]	125	alpha/beta
bromodoamin 1 in BRD4	2oss	[148]	127	alpha
human lysozyme, synthetic	2nwd	[149]	130	mostly alpha, one hairpin
fish antifreeze	2zib	[150]	130	alpha/long beta
coactosin-like prot	1t3y	[151]	131	alpha/beta
microtubule end-bdg	3co1	[152]	132	alpha
adipocyte fatty-acid bdg prot	3q6l	[153]	132	beta barrel
M tuberc hyp prot Rv1873	2jek	[154]	140	alpha
tryparedoxin-I mut	1o8x	[155]	144	alpha/beta mix
Cel45A endocluconase	1wc2	[156]	180	beta barrel with alpha tails
cyclophilin B PPI dom	3ich	[157]	188	beta/alpha

(AAXAA)₃ for all 20 amino acids

As described in Perez *et al.*[71], in the (AAXAA)₃ sequence, X represents each of the 20 amino acids. The calculated helical propensity of each amino acid was compared to the experimental w_i assigned from ¹³C=O chemical shifts[124]. Different from what was measured in the experimental measurements, three substitutes of the same amino acid were repeated to increase the statistical significance of computed results[71].

3.3.2 Simulation details**HiQ27**

The initial structures were built in Amber LEaP from corresponding crystal structures listed in **Table 3.1** with more details in **Table S3.1**. For each system, two force fields: ff14SBonlysc, ff14SB and two solvent models: GBNeck2 and TIP3P were used for parameterizations. Three different combinations were employed: ff14SBonlysc+GBNeck2, ff14SB+GBNeck2, and ff14SB+TIP3P. For each system, disulfide bonds, termini and pro-

tonation states for Histidine residues were checked using H++ to be consistent with crystal pH values, the details are included in **Table S3.1**.

Equilibrations were carefully done following the lab wiki tutorial protocols; the details are include in Supporting Information on page 72. For the short MD production runs, 290K was used for all systems to reduce kinetic fluctuations and extend data collection prior to structural changes. For each force field and solvent model combination, two runs starting from the same equilibrated crystal structure were simulated, with different initial velocities ($ig = -1$). Langevin dynamics with a 1 ps^{-1} coupling constant and 2 fs time step were used. SHAKE algorithm for restraining all the hydrogen involved bonds were on at all times. All the production runs were of various lengths but over 50 ns. So we analyzed the first 50 ns of production runs for dihedral stability and inter-conversion ratios.

$(AAXAA)_3$

The initial structures were built in Amber LEaP from sequences. For every set of $(AAXAA)_3$ peptides consist of 20 amino acids, two force fields were used: ff14SB and ff14SBonlysc. For all the peptides, the termini residues were kept free, i.e. no cap residues were added to neutralize the charges. An additional set of simulations using ff14SBonlysc was done to compare capped termini with free ones. GBNeck2 implicit solvent was used in all simulations. Equilibrations were done for 1 ns ($250 \text{ ps} \times 4 \text{ steps}$) of simulations on Bell cluster (Simmerling Lab Computer Cluster). SHAKE algorithm for restraining all the hydrogen involved bonds were on at all times. T-REMD simulations¹ with 6 replicas at the exchange rate of every 1 ps were run for $1.75 \mu\text{s}$ on Bluewaters Super Computer. Langevin dynamics with a 1 ps^{-1} coupling constant was used. 4 fs time step and hydrogen mass re-partition[40] was used.

3.3.3 Dihedral stability and inter-conversion ratio calculations

The secondary structure preferences are inferred from backbone dihedral torsion angles. In this work, the definitions of secondary structure basins have been illustrated in **Figure 1.2C**. We defined five different secondary structure elements (SSEs), namely, beta, ppII, right-handed helix, left-handed helix and none of the above (outsider). A matrix scheme, termed **matrix of SSE**, $matrix_{SSE}$, was designed to keep track of the SSEs for each amino acid in a certain protein and collectively for the whole HiQ27 data set. Each row and column represent one SSE in $matrix_{SSE}$. For example, if we look into the crystal structure of protein RNase A (PDB code: 1KF5[146]), there are 12 Alanine residues, among them, 1 Alanine falls into outsider basin, 4 Alanine adopt beta, 2 in PPII, 5 in right-handed helix and 0 in left-handed helix. To represent the $matrix_{SSE}^{native}$ of Alanine in crystal structure of 1KF5, a matrix below:

¹Temperature ladders are 300.0, 325.0, 350.0, 375.0, 400.0, 425.0 K.

$$matrix_{SSE}^{native} = \begin{pmatrix} & outsider & beta & ppII & alpha & left-alpha \\ o & \mathbf{1} & 0 & 0 & 0 & 0 \\ b & 0 & \mathbf{4} & 0 & 0 & 0 \\ p & 0 & 0 & \mathbf{2} & 0 & 0 \\ a & 0 & 0 & 0 & \mathbf{5} & 0 \\ l & 0 & 0 & 0 & 0 & \mathbf{0} \end{pmatrix} \quad (3.1)$$

records the number of dihedrals staying in the same SSE or converting into other SSEs. The columns indicate the SSE in crystal structure (in the order of: outsider, beta, PPII, right-handed alpha and left-handed alpha) and the rows indicate the SSE observed in another conformation after simulations (in the same order, with only the first letters for simplification).

When the conformation of crystal structure is considered in Matrix (3.1), only the diagonal values are non-zero; diagonal value for left-handed basin is zero because there is no Alanine dihedral adopts left-handed conformation in the crystal structure. When other conformations are considered, Alanine dihedrals that start as certain SSE will end up in the same or other SSEs. For instance, in a new conformation, for all 12 Alanine residues, the 4 originally started in beta as seen in Matrix (3.1), now have 2 beta dihedrals left, shown in red in Matrix (3.2) below. The 2 dihedrals that convert out of beta: one is found in ppII and the other in alpha. Similarly, the values for other SSEs change accordingly. So we update this $matrix_{SSE}^{newconf}$ to be:

$$matrix_{SSE}^{newconf} = \begin{pmatrix} & outsider & beta & ppII & alpha & left-alpha \\ o & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} & 0 \\ b & 0 & \mathbf{2} & 0 & 0 & 0 \\ p & 0 & \mathbf{1} & \mathbf{1} & \mathbf{2} & 0 \\ a & 0 & \mathbf{1} & 0 & \mathbf{2} & 0 \\ l & 0 & 0 & 0 & \mathbf{1} & \mathbf{0} \end{pmatrix} \quad (3.2)$$

which records the SSE results for this new conformation. To be noted that the numbers in each column add up to be the same as the diagonal numbers in Matrix (3.1).

To count for all the frames from short MD simulations, two runs of first 50 ns of production run were analyzed by taking 500 frames of conformational snapshots. Similarly, when all of the conformations in a whole trajectory are examined, the numbers in each column of $matrix_{SSE}^{Nconf}$ will also add up to be the Alanine residues found in crystal structure multiplying by the number of frames. For example, for the native structure of 1KF5, Alanine $matrix_{SSE}$ is given as Matrix (3.1), the values in each column will add up to be **500**(outsider: o), **2000** (beta: b), **1000**(ppII: p), **2500**(alpha: a) and **0** (left-handed alpha: l) for each SSE, respectively. The stability ratio of certain SSE is then calculated via dividing the number on the diagonal by the sum value. The inter-conversion ratio of some certain SSE turning into other SSEs is also calculated using the same fashion. For example, if a $matrix_{SSE}^{Nconf}$ has

the values as below:

$$matrix_{SSE}^{N_{conf}} = \begin{pmatrix} & \textit{outsider} & \textit{beta} & \textit{ppII} & \textit{alpha} & \textit{left-alpha} \\ o & 225 & 147 & 143 & 453 & 0 \\ b & 64 & 1770 & 59 & 310 & 0 \\ p & 211 & 11 & 699 & 419 & 0 \\ a & 0 & 18 & 198 & 1318 & 0 \\ l & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3.3)$$

The stability ratios shown on the diagonal in red and the inter-conversion ratios in blue are calculated accordingly as below:

$$matrix_{SSE}^{ratio} = \begin{pmatrix} & \textit{outsider} & \textit{beta} & \textit{ppII} & \textit{alpha} & \textit{left-alpha} \\ o & \frac{225}{500} & \frac{147}{2000} & \frac{143}{1000} & \frac{453}{2500} & 0 \\ b & \frac{64}{500} & \frac{1770}{2000} & \frac{59}{1000} & \frac{310}{2500} & 0 \\ p & \frac{211}{500} & \frac{11}{2000} & \frac{699}{1000} & \frac{419}{2500} & 0 \\ a & 0 & \frac{18}{2000} & \frac{198}{1000} & \frac{1318}{2500} & 0 \\ l & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3.4)$$

Which is equivalent to:

$$matrix_{SSE}^{ratio} = \begin{pmatrix} & \textit{outsider} & \textit{beta} & \textit{ppII} & \textit{alpha} & \textit{left-alpha} \\ o & 0.45 & 0.0735 & 0.143 & 0.1812 & 0 \\ b & 0.128 & 0.885 & 0.059 & 0.124 & 0 \\ p & 0.422 & 0.0055 & 0.699 & 0.1676 & 0 \\ a & 0 & 0.036 & 0.099 & 0.5272 & 0 \\ l & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3.5)$$

Note that there are two alpha basin definitions used as illustrated in **Figure 1.2C**, the more stringent definition with narrower range gives **Matrix 3.5**, while a different matrix is calculated as below, consistent with the definitions in the previous studies of Wickstrom, Maier and Simmerling *et al.* [19, 23]:

$$matrix_{SSE}^{ratio} = \begin{pmatrix} & \textit{outsider} & \textit{beta} & \textit{ppII} & \textit{alpha} & \textit{left-alpha} \\ o & 0.45 & 0.0256 & 0.085 & 0.0356 & 0 \\ b & 0.128 & 0.885 & 0.059 & 0.124 & 0 \\ p & 0.422 & 0.0055 & 0.699 & 0.1676 & 0 \\ a & 0 & 0.086 & 0.157 & 0.6728 & 0 \\ l & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3.6)$$

In the same fashion, for each protein in the HiQ27 data set, a set of matrices for each 20 amino acid was calculated. The stability ratios for outside, beta, ppII and alpha basins were extracted for each amino acid and each protein system. The box-and-whisker plot was used to deal with the statistical fluctuations among the 27 protein systems. The boxes show the inter-quartiles of ratio ranges. The horizontal lines represent the median of ratio data. The whiskers extending to the most extreme, non-outlier data points, have caps at two ends. The outlier data points are illustrated as well.

3.3.4 Helical propensity calculations

Lifson-Roig Model[158, 159] was used for experimental[124] and computations[71, 125]. In this work, we followed the same model in which protein torsion angles were decided to be either in helix or random coil, with either 1 or 0 designated in represented matrices, respectively. The secondary structure basin defining a helix-forming region (ϕ, ψ) in $(-100^\circ$ to -30° , -67° to $7^\circ)$ has been illustrated by the smaller rectangle for alpha basin in **Figure 1.2C**. Whether a residue is at the start/end of helix or within helix is further differentiated in the model. There are three states in coil-helix transition, namely, coil, start/end of helix and within a helix. Their relative weights are 1, v_i and w_i , respectively. v_i could be understood as the equilibrium constant for a residue in a coil conformation to nucleate into a helix, and w_i is the equilibrium constant to extend an existing helical segment. The calculations to estimate v_i and w_i for the substitute residue X of interest in each peptide were done in a genetic algorithm optimizer written in Python and provided by Alberto Perez. The calculation process is as follows, (1) to analyze the ϕ and ψ for all residues in the simulated structural ensembles, (2) to extract the numbers of residues in three states, (3) to maximize a log-likelihood function:

$$L = \sum_i N_{w,i} \ln(w_i) + \sum_i N_{v,i} \ln(v_i) - N_{conf} \ln(Z) \quad (3.7)$$

$$\text{where } Z = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \times \prod_{i=1}^N M_i \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (3.8)$$

$$M_i = \begin{pmatrix} w_i & v_i & 0 \\ 0 & 0 & 1 \\ v_i & v_i & 1 \end{pmatrix} \quad (3.9)$$

with the populations of helical starts/ends represented as $N_{v,i}$ and in helix segments as $N_{w,i}$, total numbers of conformation represented as N_{conf} . v_i and w_i for each amino acid are the variables subject to be optimized.

3.4 Results and Discussions

3.4.1 Diverse amino acid specific dihedral stabilities in different models

HiQ27 Crystal structures go through larger conformational changes in implicit solvent

The short MD simulations for all 27 HiQ crystal structures in three force fields/solvent combinations have been carried out on GPU implementation of Amber 15. To increase the statistical significance of data collected, two runs of simulations were used and compared. Before the in-detailed SSE stability/conversion tests, the overall $C\alpha$ -RMSD measurements against the crystal structures, respectively, are plotted in **Figure 3.1** for ff14SB+TIP3P, **Figure 3.2** for ff14SBonlysc+GBNeck2 and **Figure 3.3** for ff14SB+GBNeck2. Within the first 100 ns of MD simulations, all of systems show $< 3 \text{ \AA}$ RMSD fluctuation with respect to the crystal structures in explicit solvent results. However, in implicit solvent simulations, several proteins (2FJ8, 2QSK, 2OSS, 3CO1) go through large conformational changes.

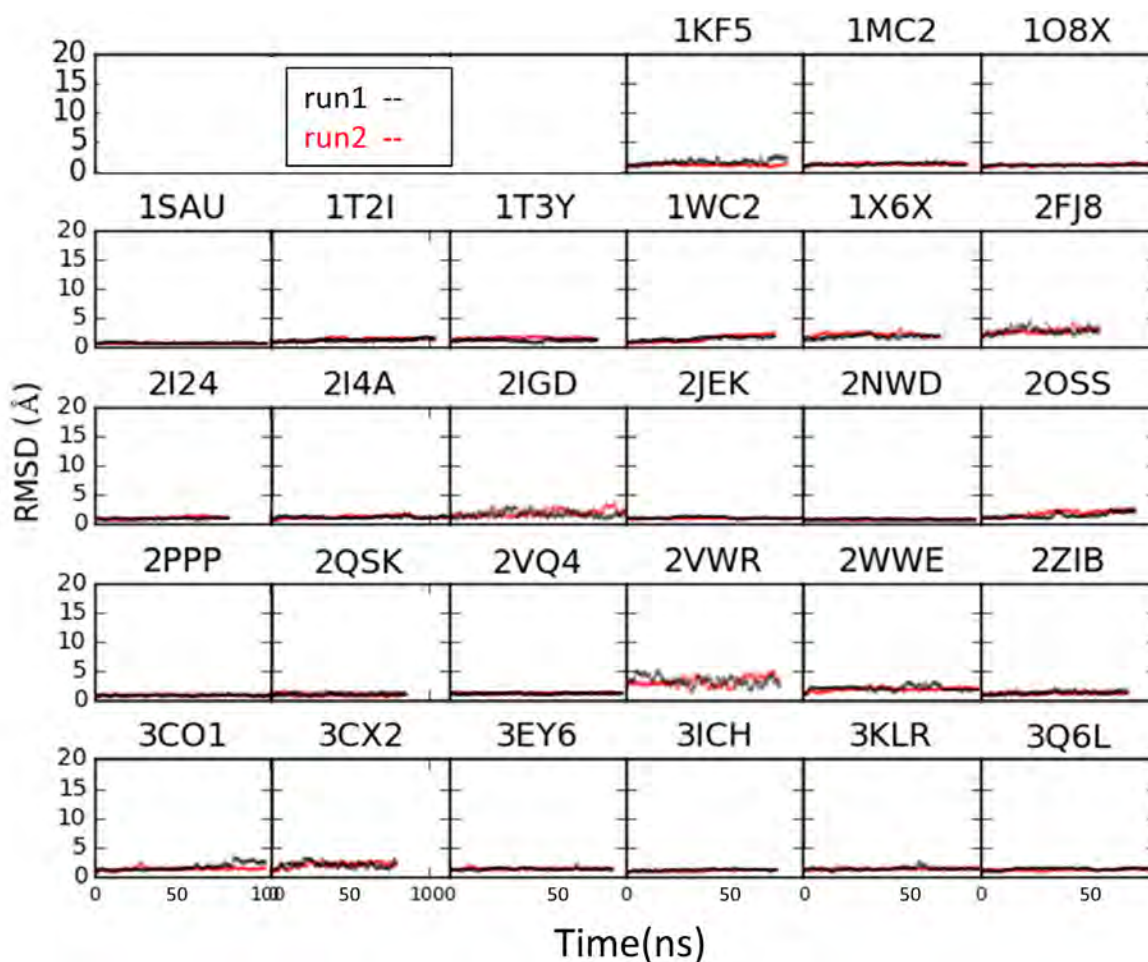


Figure 3.1: Two runs of short MD simulations for HiQ27 proteins using ff14SB and TIP3P at 290K

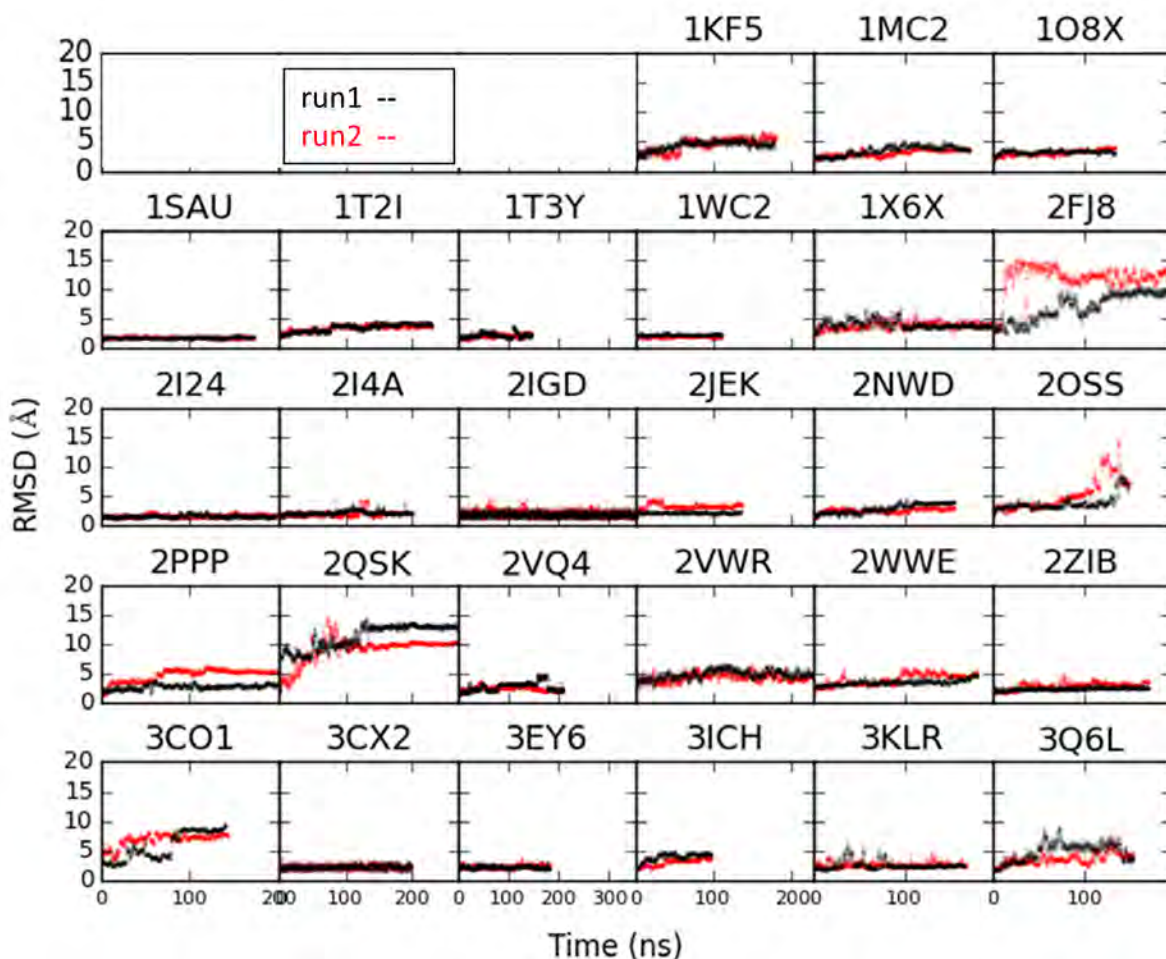


Figure 3.2: Two runs of short MD simulations for HiQ27 proteins using ff14SBonlysc and GBNeck2 at 290K

Interestingly, the main components in crystal structures of 2FJ8 and 2QSK are coils. The coil-enriched crystal structures of 2FJ8 has crystallized waters possibly providing stabilization contributions, although crystallized waters are also prevalent for other systems. As we are only simulating the protein molecule and there are no explicit water molecules included in implicit solvent simulations, it is within expectation that large conformational changes and unfolding of initial structures are observed. In contrast, explicit solvent simulations with crystallized waters show very stable $C\alpha$ -RMSD for 2FJ8 as seen in **Figure 3.1**. The system of 2QSK is even more challenging, this antiviral lectin scytovirin (SVN) protein has previously been reported in NMR and Mass-spectrum studies in which a structure of around 5 Å away from this crystal structure (PDB code 2jmv[160]) and a different disulfide bonding pattern was solved. So whether this crystal structure is a reasonable comparison is in doubt as the solution structure of this protein may well be a different structure. But as the goal here is to collect data from numerous protein systems for computational model accuracy, it is reasonable to not include 2QSK into SSE stability calculations instead of diving deeper into which structure should be compare to.

The other two crystal structures of 2OSS and 3CO1 are mainly composed of helices.

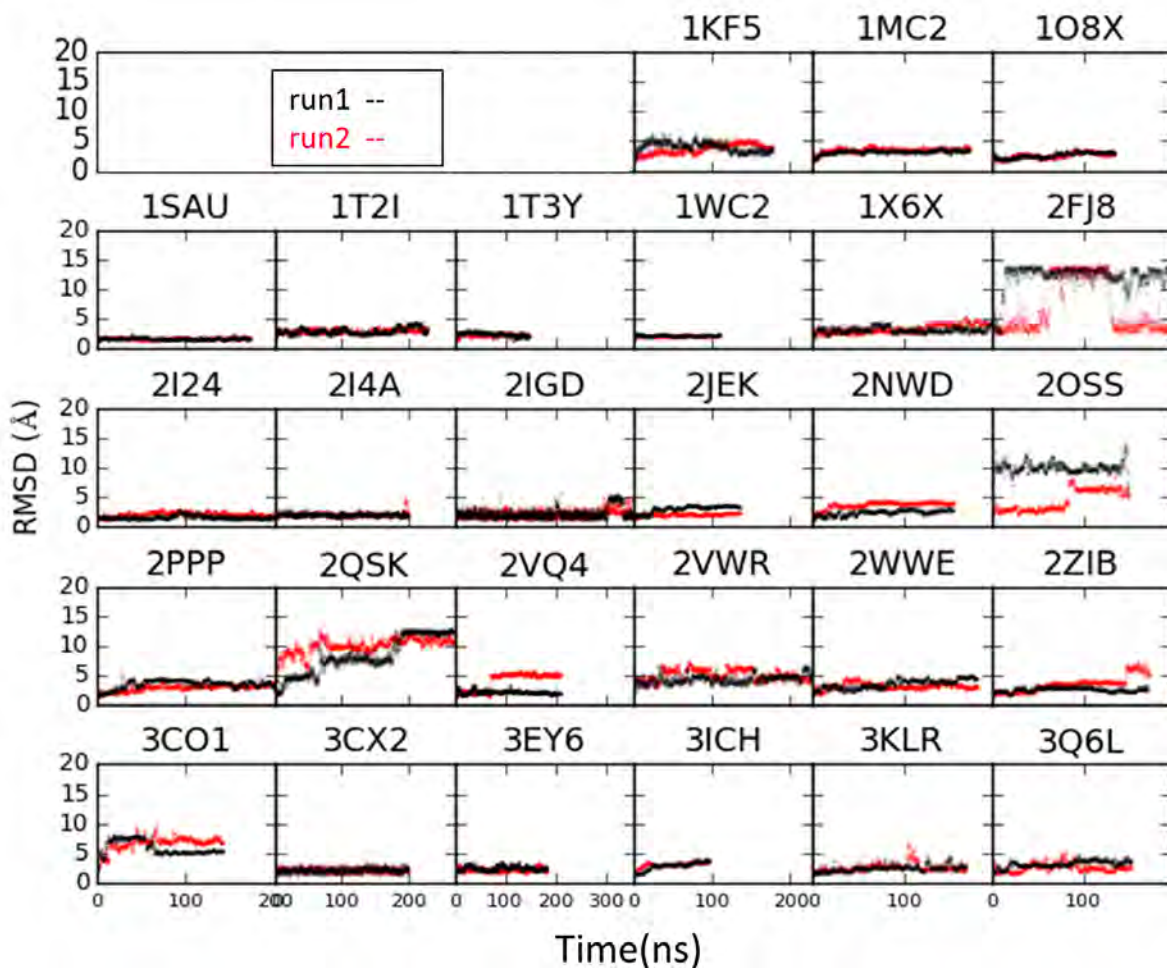


Figure 3.3: Two runs of short MD simulations for HiQ27 proteins using ff14SB and GBNeck2 at 290K

Although it is not salient here when ff14SBonlysc results are compared with ff14SB, we have seen ff14SB in stabilizing helical structures in HP36 and Homeodomain more effectively than ff14SBonlysc in **Chapter 2**. In these two proteins, these two force fields along with implicit solvent GBNeck2, do not show distinguishable stability differences, which indicate that it is not backbone parameter difference that causes the helical structures in crystals to unfold, but they also point to the crystal assembly or crystallized water related situations. Even though they add to larger fluctuations to the SSE changes, we still include them into the SSE stability and inter-conversion analysis as they are only two data points and would not statistically alter overall conclusion drawn on the larger scale of data set.

Four secondary structure element basin dihedrals have different stability ratios

Because proteins are dynamic molecules in solvent, the backbone dihedral torsion angles do not stay rigidly at one angle or in certain structural basin at all times, instead dihedrals as the only degrees of freedom in protein backbones, are dynamically changed as characterized in **Figure 3.4** throughout the simulations, even though very small overall fluctuation is

measured in all $C\alpha$ -RMSD as seen from **Figure 3.1** to **Figure 3.3**.

Different dihedrals originated in four secondary structures are found to possess diverse stabilities in general. For the alpha basin, if the definition follows the bigger rectangle as shown in **Figure 1.2C**, the stability ratios for all amino acids considered are very close to 100%. However, for beta region, most stability ratios are only larger than 80% for the simulation results of TIP3P solvent. When it comes to the ppII basin, around 80-90% of stability ratios are achieved in TIP3P solvent simulations, but it is very sensitive to the alternation of computational models and amino acid types. For outsider region, most of the TIP3P solvent results are below 60% and the ratios are also sensitive to different models and amino acid types. This observation is counter-intuitive in the beginning, but it is reasonable when the SSE definitions and the energetic point of view is taken; the larger area of SSE definition likely lead to larger probability of stability and inward-conversion rate, meanwhile the high energy barrier between ppII and left-handed alpha region likely expel the dihedrals to fluctuate towards the lower energetic regions for example the inter-conversion of beta, outsider and ppII basins.

SSE stability ratios possess amino acid specific features

Different amino acids have specific side chains and physicochemical properties, thus it is expected that regardless of the computational model used for simulations, different amino acids serve the structural roles distinctively as observed in HiQ27 data set. For example, as seen in **Figure 3.4B**, among the 19 amino acids, Methionine keeps the highest o2o and lowest p2p ratio, meaning the dihedral angles that start in coil basin are likely not to turn into secondary structures in the 50 ns of simulations, while the dihedrals in ppII region, even for explicit solvent, stay in the original places less than 60% of the time. But alpha and beta basins do not show specific feature for Methionine compared to other amino acids in general. Instead, Trptophan seems to achieve the highest ppII stability. But for Proline, its alpha basin stability is distinct from the rest amino acids, as it is a cyclic residue and restrained on the backbone thus it never samples beta nor outsider regions. Even though all the other amino acids have relatively high alpha stability, for Serine and Tyrosine, much more alpha dihedrals migrate out of the alpha basins in the short 50 ns simulations in all three force field/solvent combinations. All the specificities observed from the SSE stability ratio box plot are possibly due to specific dihedral spatial constraints determined by the side chains of different amino acids. Even though it is the backbone parameter that is assumed to dictate the secondary structure propensities in proteins, the coupling of side chain and backbone is not so easily separated. Another idea that has been explored by previous groups is amino acid specific force fields [161, 72, 68]. We also think it is a reasonable alternative to the current backbone parameter which is only trained based on Alanine and applied to the rest amino acids (except Proline and Glycine). More investigations into whether current models could reproduce the amino acid specific helical propensities have been shown in the Section 3.4.2.

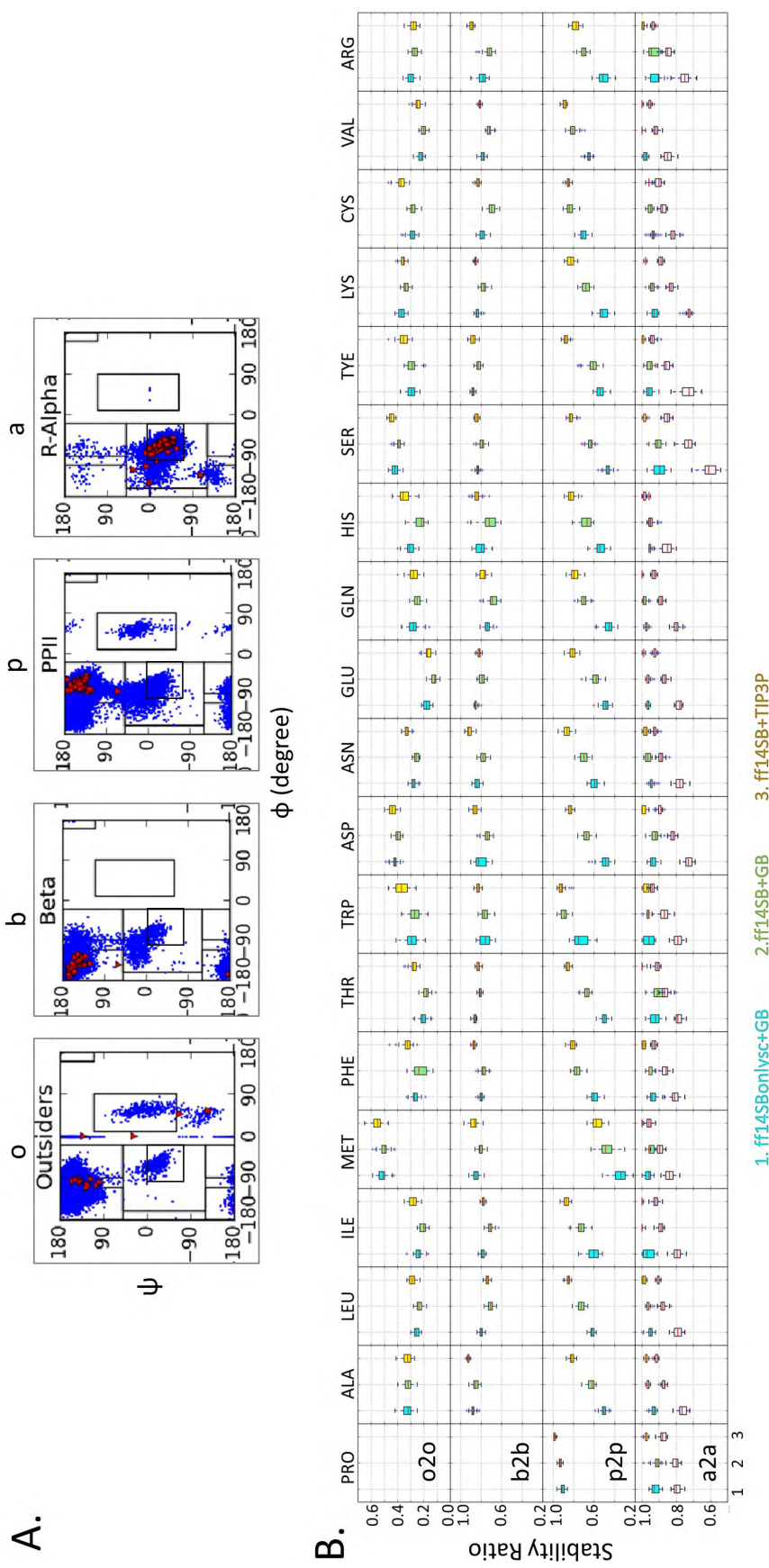


Figure 3.4: Amino acid specific stability ratios for 4 secondary structure basin dihedrals. A. Examples of the migration of monitored dihedral torsion angles starting from the respective SSE basins (denoted as red triangles) to all the other bins (blue data points) in 50 ns of simulations; B. The stability ratios of 19 amino acids for four different secondary structure basins calculated from the 50 ns of simulations using 1:ff14SBonlysc+GBNeck2, 2:ff14SB+GBNeck2 and 3:ff14SB+TIP3P. Glycine is not shown as it has different SSE definitions from the other 19 amino acids. O2o: outsider staying in outsider, b2b: beta staying in beta, p2p: ppII staying in ppII and two different definitions of a2a: alpha staying in alpha basin

ff14SB backbone modifications increase alpha and ppII stabilities

To split up the effects of force field and solvent model, we show the same analysis carried out for three different computational models, varying one thing at a time. The three model combinations studied are 1:ff14SBonlysc + GBNeck2, 2:ff14SB + GBNeck2 and 3:ff14SB + TIP3P. The model 1 and 2 only differ in the backbone parameters and model 2 and 3 only differ in the solvent models. It is expected that same length of simulations in explicit and implicit solvent do not result in comparable kinetics, so it is difficult to compare the model 2 and 3 quantitatively, but only a trend of more stable simulations in explicit solvent is concluded. Comparing stability ratios of model 1 and 2, the observed stability shift from beta to ppII and alpha regions is indeed consistent with the energy profile of ϕ dihedrals as shown in **Figure 3.5** below. When a modified energy function is applied in ff14SB, the -90° to -30° degrees of ϕ dihedrals, which correspond to the alpha and ppII region as seen in **Figure 1.2C** and **3.4A**, possess lower energy thus more stability in MD simulations. In **Figure 3.4B**, comparing model 1 and 2, when ff14SB is used in model 2, across all amino acids, the ppII stability is enhanced, for amino acids such as Phenylalanine and Cysteine, the ppII stability of model 2 is comparable with using explicit solvent of model 3. For the more stringent alpha region definition, the alpha stability ratios for model 2 compared to model 1 have increased in all cases except for Proline; because the backbone modification is not applied to Proline. With the positive comparison (all other amino acids) and control (Proline), it is safe to conclude that the changes observed in stability ratio from model 1 and 2 are consistent with the designed modifications.

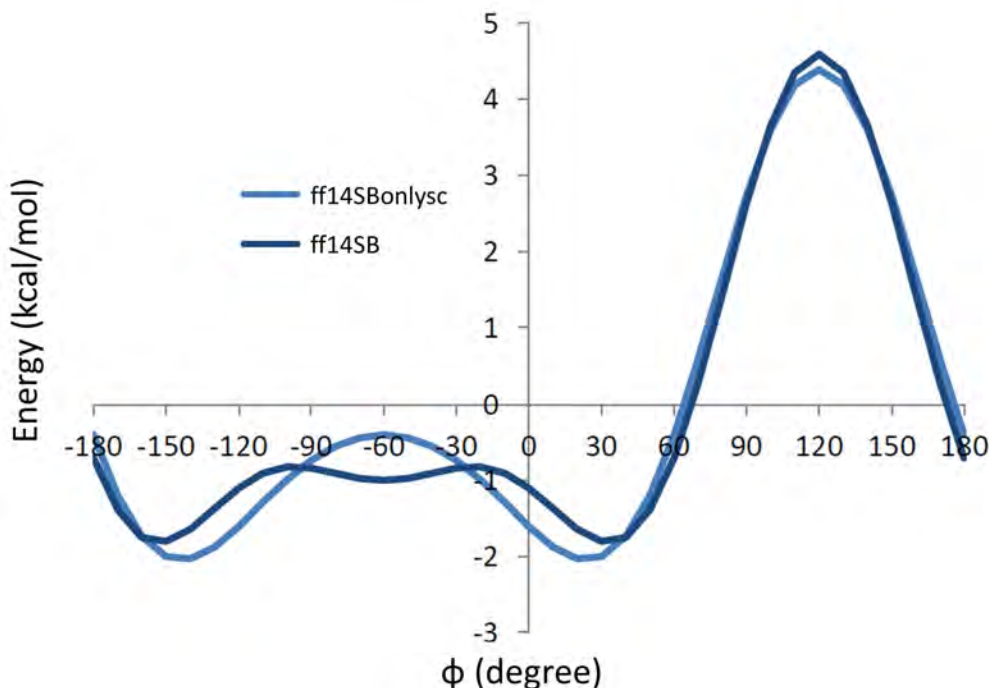


Figure 3.5: Backbone dihedral energy function for ff14SBonlysc (light blue) and ff14SB (dark blue). This figure is adapted from Figure 13 in Carmenza Martinez’s PhD dissertation.

The findings from the SSE stability test are addressed from the aspects described above,

but those are not all the conclusions we could draw, the inter-conversions of different SSE basins should also complement the stories. However, since it is a data set of high dimensions (20 amino acids \times 500 frames of snapshots \times 2 runs for 3 models), we have not established very helpful ways of representing and visualizing the inter-conversion data. Therefore, all the analysis and matrices described in Method of this chapter have been archived in hard drives. If anyone who became interested in the future, further studies and analysis could be followed upon from the current data set and results.

3.4.2 Helical propensities indicate the necessity of amino acid specific backbone parameters

For the second part of this chapter, we present the application of another toolbox, in which amino acid specific helical propensities are estimated by the helix extension w_i properties for all 20 amino acids. The experimental helical propensities are derived from $^{13}\text{C}=\text{O}$ chemical shifts [124]. The simulated helical propensities are calculated based on the Lifson-Roig model[158, 159] extracted from the $(\text{AAXAA})_3$ simulations, using ff12SB + GBNeck2, ff14SB + GBNeck2 and ff14SBonlysc + GBNeck2. The computed vs experimental helical propensities for these three model combinations are shown in **Figure 3.6**.

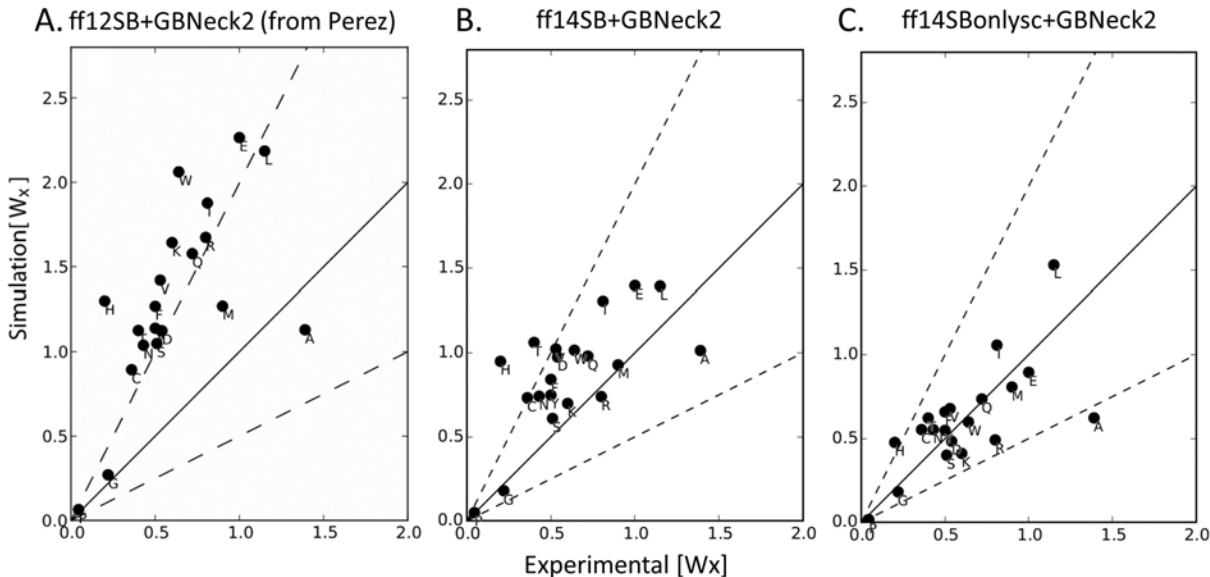
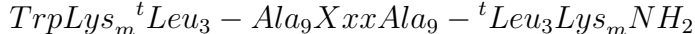


Figure 3.6: A. Computational vs. experimental helical propensities for ff12SB with GBNeck2, which is obtained from Perez[71]; B. ff14SB with GBNeck2; C. ff14SBonlysc with GBNeck2. Dotted lines indicate $k_B T \ln(2)$ error in free energy with respect to experiment. The solid line indicates perfect agreement computation and experiment

In experimental assignments, Alanine is the most helical-prone amino acid followed by Leucine. However, in all simulated models, Alanine helical propensity is largely underestimated. In the model of ff12SB+GBNeck2 (**Figure 3.6A**), Alanine is not helical enough, falling even below half of other amino acids; Proline and Glycine agree with experimental

values, while all the rest amino acids bias towards overestimated helical propensities. In the model of ff14SB+GBNeck2 (**Figure 3.6B**), Alanine is again not helical enough, but it is not as low relatively to other amino acids; Proline, Glycine, Arginine and Methionine agree well with experimental values; all others except Histidine and Threonine are slightly overestimating the helical propensities; while Histidine and Threonine bias the helical propensity by more than doubling the experimental values. As for the model of ff14SBonlysc+GBNeck2 (**Figure 3.6C**), except Histidine is overestimated and Alanine is underestimated, all the other amino acids fall into the $k_B T \ln(2)$ error range. Compared ff14SB with ff12SB, all helical propensities shift to smaller values by about the same value, which end up with relatively the same order. These findings suggest that amino acid specific backbone parameters are needed to fine-tune the relative helical propensities, as the backbone parameter for all (except Glycine and Proline) will shift all amino acids together without fixing the relative wrong orders.

The inconsistency between simulated and experimental systems, which should not be neglected, hinders the reliability of quantitatively comparing the simulation results with experimental values. According to this experiment, a series of peptides were synthesized and analyzed in D_2O (pD was controlled to be 7.0) at 25°C. The peptides shared the following sequence:



in which ${}^t\text{Leu} \equiv \text{tert-Leucine}$, $m = 6$ or 8 , $\text{Xxx} \equiv$ one of 20 amino acids. This sequence was used because it is experimentally tractable and retains detectable helicity over the aqueous temperature range 2-60°C. The polylysine caps were used to increase the solubility by preventing aggregation. But the simulated systems do not contain the charged termini and also with 3 Xxx substitute to increase the statistical significance.

However, this test case provides well-rounded helical propensity data for all 20 amino acids and a reliable relative comparison of simulations and experiments. Although only helical propensities are investigated in this case, it is still part of the useful toolbox for computational model evaluation and issue diagnosis. Therefore, it has been adapted for force field development test, which would provide more helpful insight into the future force field modifications.

3.4.3 Cross-validation of stability test and helical propensity test

In SSE alpha dihedral stability test, it is found that Serine, in both ff14SBonlysc and ff14SB in GBNeck2, performs the worst among all 19 amino acids in stabilizing in the original Ramachandran basin of the same definition. Alanine actually does better than Aspartic acid, Tyrosine, Lysine and Arginine when the medians are compared, and has completely higher stability ratio than Serine as seen in alpha region stability ratios (a2a row of **Figure 3.4B**). Although in helical propensity test, Alanine becomes the most problematic amino acid, which is underestimated the most in both force fields in **Figure 3.6**, Aspartic acid, Tyrosine, Lysine and Arginine are also of relatively lower helical propensities, which is consistent with SSE stability test.

As the alpha region instability of Alanine is only observed in the helical propensity test but not in the dihedral stability test, three possible explanations are discussed and

could possibly provide more insight into the future toolbox design. (1) Alanine instability in α -helical propensity is a sequence dependent problem, meaning in the special case of $(AAXAA)_3$, low alpha propensity of Alanine is exemplified but it may not be carried over in other systems, such as in the data set of HiQ27. In this sense, more investigations are needed to validate the conclusions made from this chapter. (2) The two toolboxes actually have tested different properties of secondary structure stabilities. In the dihedral stability test, it is the breakdown of secondary structures that is monitored, while extension parameter w_i in helical propensity test is measuring the α -helical dihedral formation once the previous dihedral is already in alpha. It is possible that with current backbone parameters, the ability for Alanine to stay in the pre-formed alpha region is not problematic compared to the rest amino acids, but the ability to convert back from other basins is the center of the issue. In the energetic perspective, the depth of the alpha basin right now is deep enough so that once alpha dihedrals fall into it, it is not exceptionally easy to get out. But the depth of other basins are relatively deeper than they should be thus it is difficult to sample back into the alpha basin. For other amino acids, the relative depths among all the secondary structure basins are more balanced. (3) In terms of the stability test, a modified approach to collect the statistics of dihedral basin conversion might be more reasonable. Instead of always referring back to the starting dihedral basin in the crystal structures, it makes more sense to reset the initial basins to the previous frame(s), which might be the real conversion ratio we are going after.

3.5 Conclusions

In this chapter, two benchmark data sets (SSE dihedral stability and helical propensity toolboxes) are employed and their usages of evaluating the accuracy of computational models are demonstrated. With respect to experimental quantities, the amino acid backbone specificities simulated from two Amber force fields in implicit solvent are compared and further analyzed.

In the developed dihedral stability toolbox, 27 high quality proteins previously used for Rosetta energy function training have been carefully set up, equilibrated and simulated for 50 ns. Fluctuations of structures are compared across two different force fields. Explicit and implicit solvents are also kept as single variable for comparison. As expected, larger conformational changes were observed in implicit solvent simulations. When all the dihedral angles are compared with their original dihedral secondary structure basin in crystal structures, different secondary structure dihedrals show diverse stability: alpha basin dihedrals in average possess higher stability than beta and ppII regions, outsider basin has the lowest stability. The comparison across 19 amino acids (except Gly) manifests specificity in each of the secondary structure stabilities, even without amino acid specific parameters. The different set of ff14SB backbone parameters compared to ff14SBonlysc has also shown more stability in alpha and ppII basin, which is consistent with the modified energy profile.

We also applied helical propensity toolbox to study if current computational models (ff14SB with GBNeck2 and ff14SBonlysc with GBNeck2) could reproduce experimental trend of helix extension propensity for all 20 amino acids. Compared with ff12SB that largely overestimates the helical propensity of all amino acids except Alanine is slightly underestimated,

ff14SB reproduces the experimental trend better overall and ff14SBonlysc performs the best. However, as the same backbone parameters were used (except Glycine and Proline), the improvement observed in other amino acids results in worsening the Alanine and underestimating its helical propensity by even more. Therefore, we concluded that amino acid specific backbone parameter might be the key to resolve this issue.

Lastly, the conclusions drawn from the two toolbox tests separately are cross-compared and discussed. Two tests agree on the low alpha stability residues Serine, Aspartic acid, Tyrosine, Lysine and Arginine. The discrepancy displayed in Alanine residue could be ascribed to three causes and provide further investigation directions.

3.6 Supporting Information

Table S3.1: Protonation states and other structural features for proteins in HiQ27 data set

#PDB_ID	Crystal pH	Structural features
1kf5	7.1	3 HIE: 12,105,119 1 HIP: 48 SS-bonds: 26-84, 40-95, 58-110, 65-72
1mc2	5.8	1 HIE: 110 1 HIP: 47 solvent: IPA 7 SS-bonds
1o8x	8.2	2 HIE
1sau	9.5	2 HIE 1 SS bond
1t2i	7.2	1 HIE: 85 1 HIP: 53 1 SS-bond
1t3y	8	1 HIE
1wc2	5.5	9 HIP 6 SS bonds solvent: ACT, PEG
1x6x	8.2	1 SS-bond
2fj8	5.64	10 SS-bonds
2i24	7.4	2 SS-bonds solvent ion: Cl
2i4a	4.6	1 HIP 1 SS-bond solvent: BME (hbonding with Lys51 sidechain)
2igd	4.8	alternate rotamers

Continued on next page

Table S3.1 – continued from previous page

#PDB.ID	Crystal pH	Structural features
2jek	6.5	1 HIP: 74 2 HIE 31, 66 solvent: SO4, GOL
2nwd	4.9	1 HIP 4 SS-bonds
2oss	7.5	1 HIE solvent: EDO
2ppp	7	3 HIE
2qsk	8	1 HIP (!) 5 SS-bonds solvent: Cl, GOL
2vq4	-	just alternate rotamers
2vwr	4.2	2 HIP
2wwe	5.5	6 HIP 1 HIE: 70
2zib	5.4	7 HIP 5 SS-bonds solvent: SO4
3co1	7.4	5 HIE
3cx2	6.9	4 HIE: 55,74 2 HIP: 59,106
3ey6	7.5	2 HIE
3ich	6	1 HEP: 96 2 HIE
3klr	3	1 HIP solvent: SO4, GOL
3q6l	7.4	1 HIS

Explicit solvent equilibration protocol

If **TIP3P** was the solvent, periodic boundary condition was always on. The equilibrium steps followed the lab wiki tutorial for TIP3P equilibration 9 steps:

(1) 10,000 steps of energy minimization for water and hydrogen atoms restrained on crystallized water oxygen atoms and heavy atoms of protein with $100 \text{ kcal}/(\text{mol}\cdot\text{\AA}^2)$ force constant. E.g. 1KF5 has 227 crystallized water molecules, so the restraint mask for this system is `' :130-356@O, !:WAT & !@H='`;

(2) in 100 ps (1 fs time step) of MD simulation time, to heat the system to targeted temperature (290 K in HiQ27 simulations), with $100 \text{ kcal}/(\text{mol}\cdot\text{\AA}^2)$ force constant restrained on the same atoms as in step 1;

- (3) 100 ps (1 fs time step) of constant pressure constant temperature with the same restrained as in step 2;
- (4) 250 ps (1 fs time step) of constant pressure constant temperature with weaker (10 kcal/(mol·Å²) force constant) restraints;
- (5) 10,000 steps of energy minimization with 10 kcal/(mol·Å²) force constant restraints on backbone and crystallized water oxygen ' @CA,N,C | :130-356@O ';
- (6) 100 ps (1 fs time step) constant pressure, constant temperature MD simulations with restrained backbone with the same weak restraints as in step 5;
- (7) 100 ps (1 fs time step) of constant pressure, constant temperature MD simulations with 1 kcal/(mol·Å²) force constant restraints on backbone atoms;
- (8) 100 ps (1 fs time step) of constant pressure, constant temperature MD simulations with 0.1 kcal/(mol·Å²) force constant restraints on backbone atoms;
- (9) 500 ps (2 fs time step) unrestrained MD simulations to finish the equilibration.

Implicit solvent equilibration protocol

If **GBNeck2** was the solvent, there was no periodic boundary conditions but equilibrium was done in similar fashion, except there are no crystallized water or neutralizing Na⁺/Cl⁻ ions. The 7 steps of equilibration are as follows:

- (1) 1,000 steps of energy minimization restrained on all heavy atoms with 10 kcal/(mol·Å²) force constant;
- (2) 500 ps (1 fs time step) of MD simulation time, to heat the system from 100K to targeted temperature (290 K in HiQ27 simulations), with the same restraints as in step 1;
- (3) 1,000 steps of energy minimization restrained on all backbone atoms with 10 kcal/(mol·Å²) force constant;
- (4) 500 ps (1 fs time step) of MD simulation and heating, with the same restraints as in step 3;
- (5) 500 ps (1 fs time step) of MD simulation with 1.0 kcal/(mol·Å²) force constant restraints on backbone atoms;
- (6) 500 ps (1 fs time step) of MD simulation with 0.1 kcal/(mol·Å²) force constant restraints;
- (7) 500 ps (2 fs time step) of unrestrained MD simulation to finish the equilibration.

Chapter 4

Strategic Refinement of Homology-Modeled and GB-Folded Protein Structures

4.1 Abstract

The refinement of protein structures has recently benefited from more accurate force fields and short MD simulations with backbone restraints. Strategic refinement trials of structures from homology modeling and *ab initio* folding using GB solvent are reported here. In the GB-folded data set, explicit solvent simulations stabilize the starting structures but do not significantly refine them. In the CASP data set, implicit solvent simulations starting from both crystal structures and structures generated by homology modeling indicate our ability to refine near-native structures, but improvements are on the regional scale. Both the best RMSD conformation in the trajectory and clustering analysis are used to identify improvements in the refinements. While there are cases where we could not improve the structure further, we are also able to rectify wrongly positioned loops within a nanosecond of simulation time and thus to refine the overall structure.

4.2 Introduction

During the last 6 years, the numbers of sequences in UniprotKB/TrEMBL database increased by a factor of 8, from 13 million to over 109 million[162], while the numbers of protein structures deposited in the Protein Data Bank (PDB) only doubled, from around 70,000 to around 138,000[163]. Therefore, when experimental structures of biomolecules are not available, computational structure predictions meet the demands and offer possible working models, thus they serve as indispensable tools. Current methods, especially template-based modeling, are successful in predicting overall folds for proteins. In light of evolution theory, if one structure of the same protein family has been solved experimentally, the structures of the other members could be rebuilt because the spatial arrangement of 3D structure is usually conserved within the same protein family[164]. More than two decades of Critical Assessment of Techniques for Protein Structure Prediction (CASP) since 1994

have documented and promoted these progresses[165, 166].

However, the $C\alpha$ -RMSD of those predicted structures of large proteins are typically greater than 4 Å[166, 107, 24, 167]. These structures need further refinement in order to reach the accuracy for interaction design, or virtual screening. Meanwhile, the experimental structural information coming out of NMR, SAXS, cryo-EM, and sometimes crystallography studies is of low resolution thus requires to be refined. Although the scope and accuracy of comparative modeling have been increasing, and *ab initio* structure predictions have been demonstrated on diverse sets of fast-folding proteins, to predict structures closer to the native structure than to the original templates still proves very challenging[168].

Protein refinement category has been set up since 2006 (CASP 7) to blindly evaluate the state-of-the-art protein refinement methodologies[165]. The methods employed for this endeavor include Molecular Dynamics, fragment and knowledge-based approaches, elastic network models, and hydrogen bond network optimization etc.[168]. Two factors are used to determine the refinement progress: (1) the ability to sample around the native structure efficiently (sampling & accuracy); (2) a scoring function that can correctly identify the native/near-native states (selection)[169].

The knowledge-based, or statistical potential, is one of the two major commonly relied energy functions in the assessment of structural models. These potentials derive structural features such as torsion angles, solvent exposure, crystal environment, hydrogen bond geometry from the Protein Data Bank. Yang Zhang and co-workers have pioneered the field using iterative threading assembly refinement (I-TASSER) server [170, 171, 172], where in the refinement step, constraints from threading alignments and PDB structures are used in fragment assembly simulation and hydrogen bonding networks optimization. KoBaMIN web server does protein structure refinement based on minimization of potential of mean force that considers the solvent effect, side chain rotamer positions, etc.[173]. Jones Membrane potential is able to utilize the sequence, structure and lipid environment for refining correct orientations for membrane proteins[174]. 3Drefine refinement protocol includes iterative optimization of hydrogen bonding network and atomic-level minimization using knowledge-based force field[175]. Also, protein evolution information[176] has been applied to generate pairwise residue contacts, which could be employed as structural restraints. In terms of the sampling, however, low resolution initial conformation search and rigid body rearrangement of structural segments often suffer from conformational traps.

The physics-based potentials, especially referring to a variety of molecular dynamics methods, have endured trial and errors in structure refinement as well. Based on physicochemical principles, MD using mechanical forces, arguably, should provide the ultimate potential functions for protein structure modeling[177]. Furthermore, it is in theory more transferable for more general purposes, such as in studying protein misfolding and aggregation. Although D.E. Shaw and co-workers have pointed out that force field (Charmm 22* and modified TIP3P solvent) being inaccurate might be the major factor of unsuccessful refinement[178], to refine protein structures using MD have been shown to be plausible. For example, Michael Feig and co-workers have successfully applied backbone-restrained short MD simulations in Charmm22* and modified TIP3P explicit solvent since CASP 9[179, 180] and have protocolized PREFMD (Protein structure REFinement via Molecular Dynamics) and locPREFMD web servers to improve the model quality of template-based structures using MD and structural averaging. Yun-Dong Wu and co-workers have applied their RSFF2

and TIP3P MD simulations to refine 30 of the CASP 10 structures and achieved better averaging improvement, using even weaker restraints[119].

In other cases, combinatorial methodologies have employed potentials from both sides and complemented each other efficiently. For example, Zhang *et al.* used fragment-guided MD to sample conformations and refine protein structures to atomic-level, where the MD energy funnel has been biased by the distance information from PDB fragmental analogs[181]. Another interesting example is from Fan *et al.* where hydrophobicity of solvent was altered in MD simulations to change intramolecular hydrogen bonding, secondary structure breakdown and reformation and rearrange poorly packed regions by mimicking function of chaperones (Chaperone-Hamiltonian)[169]. Rosetta score function that based on physical and statistical terms are utilized in combination of MD to interactively refine globular and membrane proteins [182, 183, 184], referring or not referring to experimental constraints.

In this work, we are interested in protein refinement problem, as well as building benchmarks for understanding how implicit solvent performs on more real-world problem like protein structure refinement, when both accuracy and sampling are thought to be fueled by recent ff14SBonlysc side chain modifications[23] and GBNeck2 solvent[62].

Since protein folding to near-native structures has been shown accessible by employing implicit solvent (GBNeck2) with a combination of ff14SBonlysc force field and GPUs, we intend to further test the accuracy of our model on the protein refinement problem. To validate our physics-based methods, it is essential to try improving protein structures by no means of experimental restraints or statistical potentials. First part of this chapter is to refine the most populated structures as a follow-up of GB folding experiment, which are still 2-6 Å away from the native NMR structures. For the second part of this chapter, we applied and analyzed the capability of implicit solvent in refining structures from template-based conformations provided by CASP 11. We left the proteins unrestrained for accuracy diagnosis and indeed we observed issues that have been elaborated in the first two chapters, including the lack of nonpolar term in implicit solvation as investigated in **Chapter 2** and the backbone secondary structure preferences studied in **Chapter 3**.

4.3 Methods

4.3.1 Refinement targets

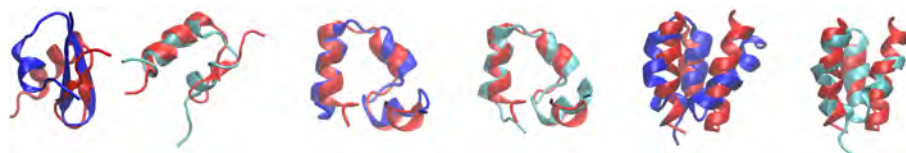
GB folded targets

We first applied ff14SB and TIP3P to refine the most populated cluster representatives sampled in the protein folding experiment using ff14SBonlysc and GBNeck2[24]. The information about all three proteins (BBA, HP36 and proteinB) has been listed in the Table below. The initial conformations were the representative structures from clustering analysis on 250K and 300K temperature trajectories, respectively, as described in this folding experiment[24].

Table 4.1: 3 GB folded targets and their structures

System	BBA		HP36		proteinB	
AA	27		36		47	
T_m (K)	no T_m reported		343		>363	
RMSD region	4-26 or 4-14,14-26		3-32		2-44	
$C\alpha$ -RMSD (Å)	4.6 ¹	5.1 ²	2.3 ¹	2.6 ²	4.2 ¹	3.3 ²

Structure



Note: ¹ 1st populated cluster from lowest temperature trajectories (BBA 244 K, HP36: 250 K, proteinB: 250 K). ² 1st populated cluster from 300 K trajectories. The initial structures of each protein (dark blue for lowest temperature cluster representatives and light blue for 300 K cluster representatives) are overlapped with native structure in red, respectively, based on the $C\alpha$ atoms in the RMSD regions.

CASP11 refinement targets

All 37 targets from the CASP11 refinement data set were examined for refinement. The detailed experimental information and structural features are listed in **Table S4.1**. Upon careful comparison of provided templates and experimental structures, we kept 30 proteins that (1) come with experimental structure in the downloaded database (see details in Note for **Table S4.1**), and (2) do not contain missing residues in the template (except only tail residues are missing). The selected 30 proteins and the quality of template structures are listed in **Table 4.2**. Ranging from size (#amino acid) 62 to 265, the templates provided by CASP11 are of various initial RMSD values and GDT-HA scores.

Table 4.2: 30 studied CASPR11 targets and the template model quality

target	AA	$C\alpha$ -RMSD of initial model	$C\alpha$ -RMSD of CASP #1 rank ¹	GDT-HA of initial model	GDT-HA of CASP #1 rank ¹
TR228	84	3.92	3.148	55.66	66.37
TR274 ²	194	6.80	5.367	29.10	28.55
TR280	96	4.03	3.006	59.37	71.88
TR759	62	4.23	2.116	45.16	62.10
TR760	201	3.14	3.115	57.71	58.70
TR762	257	3.07	2.162	70.82	72.76
TR765	76	2.58	2.245	59.09	72.73
TR768	143	2.61	3.634	64.69	72.90

Continued on next page

Table 4.2 – continued from previous page

target	AA	C α -RMSD of initial model	C α -RMSD of CASP #1 rank ¹	GDT-HA of initial model	GDT-HA of CASP #1 rank ¹
TR769	97	1.74	1.219	59.80	72.68
TR772	198	4.78	4.615	52.52	53.78
TR776 ³	219	3.22	3.183	64.27	68.61
TR780	95	2.73	2.505	54.47	60.00
TR782	110	1.93	1.757	65.23	74.32
TR783	243	3.26	2.679	58.02	64.09
TR786	217	3.62	3.716	49.08	54.15
TR792	80	1.99	1.478	57.81	75.31
TR803	134	5.97	7.241	34.33	38.25
TR811	251	1.45	1.191	73.51	76.59
TR816	68	2.53	1.154	51.84	74.27
TR817	265	1.81	1.699	66.32	71.04
TR821	255	2.45	1.634	49.02	63.63
TR822 ²	121	4.21	4.222	30.48	37.28
TR827	193	3.75	2.771	35.23	48.31
TR829	67	6.16	1.216	51.12	76.87
TR833	108	4.71	2.254	62.27	64.81
TR837	121	2.95	2.685	43.80	48.76
TR848	138	3.78	2.625	58.88	63.95
TR854	70	2.27	2.080	60.36	66.42
TR856	159	2.68	2.659	62.26	63.68
TR857	96	4.00	4.523	34.12	10.10

Note: ¹ CASP #1 rank is the best reported result among all the CASP11 Refinement submissions. ² System of missing tail residues in the template structure, so it differs from the native structure in the tail missing residues. ³ Ser 18 is phosphorylated in crystal structure and mutated to Ser in simulations.

4.3.2 Simulation details

GB folded targets

Two GB-folded structures and one native structure were loaded and built in LEP with ff14SB and TIP3P explicit solvent. For three structures in each protein, the same number of TIP3P waters were added to ensure the consistency (buffer size > 8 Å in all cases). Short equilibrium including minimization and short MD simulations were carried out to equilibrate the water densities and solute structures to 300 K (detailed protocol described on Page 72). 800 ns, 600 ns and 400 ns, respectively of TIP3P simulations were carried out for BBA, HP36 and ProteinB after equilibrations.

CASP 11 refinement targets

Different from GB folded targets where only explicit solvent was used with ff14SB, for CASP11 targets, we compared two different combinations of computational models (1) ff14SBonlysc + GBNeck2 and (2) ff14SB + TIP3P. In each case, one template structure and one native structure were set up and simulated using the same parameters except different starting coordinates. For TIP3P solvent, 8 Å of water buffer was added to both template and experimental structures. All the initial structures were built in LEaP and equilibrated following the same protocol as described on Page 72. For TIP3P solvent simulations, at least 40 ns of simulations were carried out in both native and refinement runs; for GBNeck2, at least 100 ns of simulations were carried out and analyzed.

REMD simulations using ff14SBonlysc and GBNeck2 were carried out for all except TR228, TR280, TR759, TR760, TR765, TR817, TR829 and TR833; these systems were excluded because MD simulations from native structures were stable at 300 K for these systems. Respectively, more than 300 ns of REMD simulations for the 22 systems were simulated from the template structures to enhance the sampling, with the highest temperature at 300K. The temperature ladders are included in **Table S4.2**. The lowest temperature replica, respectively, was analyzed for the percentages of structure refined and cluster analysis.

4.3.3 Evaluation Criterion

RMSD calculations were used to determine the refinement level of structures coming out of MD simulations. For GB folding proteins, we histogrammed the RMSD distributions of MD simulations to show to what extent the simulated structures were refined compared with the initial RMSD. For CASP11 refinement targets, Δ RMSD values were employed to measure the change of $C\alpha$ -RMSD against native structure in simulated structure with respect to the initial template model, a negative value indicates the template model is refined. In CASP, two other metrics are also helpful for global structure quality determination, namely, GDT-score and TM-score. In our analysis, all three metrics have been considered but only the data using RMSD are shown.

4.4 Results and Discussions

4.4.1 Refinements from GB predicted structures

The short MD simulations in ff14SB and TIP3P for all three proteins are initially from the top cluster representative structures predicted in microseconds of REMD using ff14SBonlysc and GB[24] starting from only sequence information. The RMSD histograms of those REMD simulations are compared with our MD simulations in this work, shown in **Figure A** of **4.1** for BBA, **S4.1** for HP36 and **S4.2** for ProteinB.

For BBA, the highest peak observed in 243.8 K trajectory in **Figure 4.1A** corresponds to the 4.6 Å structure shown in **Table 4.1** and the largest cluster (shown no apparent peak in **Figure 4.1A**) representative structure corresponds to the 5.1 Å structure shown in **Table 4.1**. When these two structures and native structure are solvated and simulated in ff14SB

and TIP3P, the RMSD histogram in **Figure 4.1B** indicates that MD simulated structures in terms of overall RMSD are not improved, as they all center around the original starting RMSD values. If we focus on the more local regions of BBA, however, we still see large improvements in the N-terminal hairpin structure and C-terminal helix. As seen in **Figure 4.1C**, when the hairpin is overlapped for the initial and native structures and the RMSD values are measured when this region is used in superimposition, RMSD distributions of simulations starting from both initial structures shift left towards smaller RMSD values. In **Figure 4.1D**, the helical region is refined throughout the whole MD simulations, as from both initial structures, the RMSD distributions measured with respect to the helical region in native structure, are almost all on the left of initial points. This result is interesting as it suggests that short MD simulations in explicit solvent are able to refine structures locally. But whether longer simulations are able to further refine the global structures are yet to be investigated.

For HP36, as the initial structures are originally less than 3 Å from the native structure, it is more challenging to get them refined. But we still see improvement throughout the whole short MD simulations as shown in **Figure S4.1B**; the RMSD distributions of refinement runs are even of narrower width compared with native run. One explanation for the immediate refinement of GB-folded HP36 structure is the necessity of nonpolar term in order to stabilize the HP36 native structure, which has been thoroughly investigated in **Chapter 2**. For ProteinB, the 3.3 Å structure is not as refined as the other 4.2 Å structure throughout the MD simulations as large RMSD distribution left-shift is only observed for the latter run not the former run.

As described in the Introduction, restrained MD simulations in explicit solvent have been employed by Mirjalili and Feig *et al.*[179, 180] in CASP competitions and have achieved overall success. But in our studies, instead of following their protocols, we tried a different force field and explicit solvent model (ff14SB and TIP3P vs. Charmm22* and modified TIP3P used by them) and found that short MD simulations using explicit solvent do not facilitate large conformational changes as was also concluded by them[179, 180]. Since our goal is not to perform better in CASP competition overall, we chose to try more aggressive refinement using unrestrained MD simulations in implicit solvent, rather than more conservative refinement using restrained MD simulations in explicit solvent. Therefore, in the rest part of this chapter, we show more practical refinement trials using unrestrained MD and implicit solvent, with CASP11 data set. Unrestrained explicit solvent simulations are also carried out for comparison.

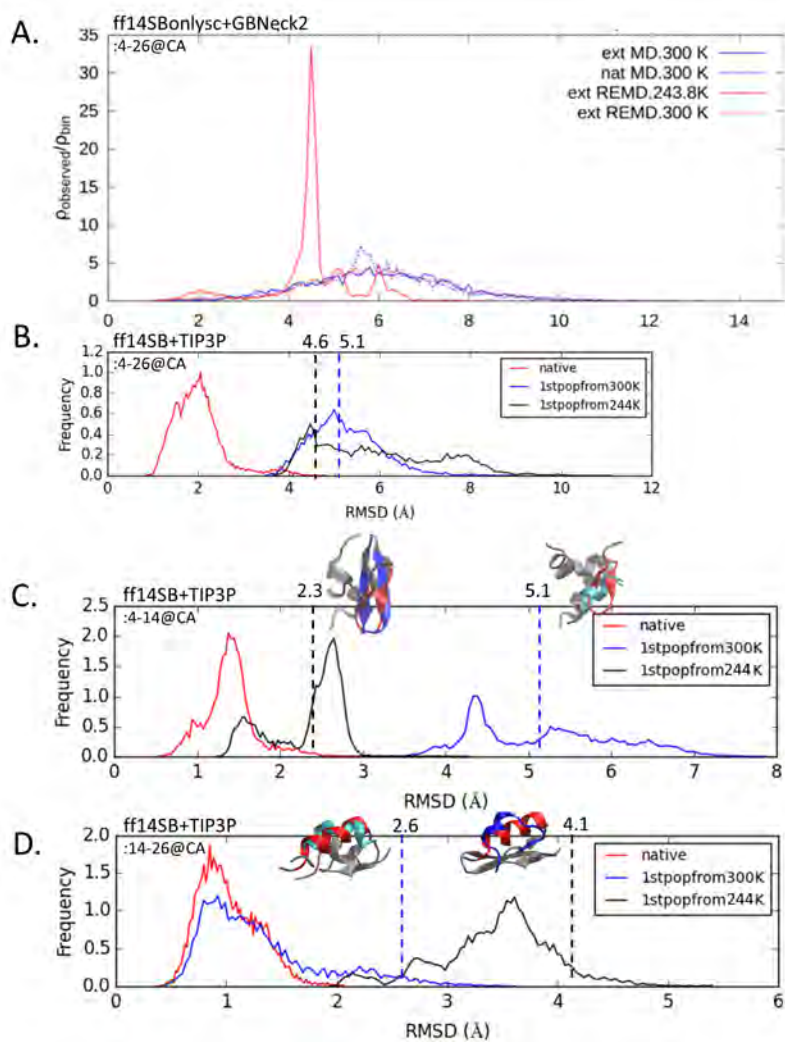


Figure 4.1: BBA GB-folded refinement, A. RMSD distributions of MD simulations at 300K from extended and native structures, as well as REMD simulations extracted from temperatures 243.8K and 300K from extended structure. This figure is adapted from the Supporting Info of the GB folding study[24]; B. RMSD distributions of MD simulations at 300K from native structure of BBA, and two top cluster representative structures from the REMD trajectories. The starting RMSD values are denoted on the upper x-axis and indicated as dashed lines; C. RMSD distributions of the same simulations as shown in B except the residue 4-14 $C\alpha$ -atoms are superimposed in RMSD measurements. The initial RMSD values and structures are illustrated, shown to overlap with native structure in red and excluded regions in gray; D. RMSD distributions measured with residue 14-26 $C\alpha$ -atoms superimposed.

4.4.2 Stability of CASP11 experimental structures in MD

Short MD simulations starting from experimental structures are first analyzed as controls. For explicit solvent results, as shown in **Figure 4.2**, all the MD simulations from native structures are more than 40 ns long and stay close to 1-2 Å RMSD from native struc-

tures. For implicit solvent results, as shown in **Figure 4.3**, are of longer simulations and larger fluctuations compared to the explicit solvent counterparts. Among the 30 proteins, 16 (TR274, TR280, TR759, TR760, TR765, TR769, TR772, TR776, TR780, TR782, TR817, TR829, TR833, TR848, TR856, TR857) of them stay at or below 5 Å in average RMSD, another 10 of them sample high RMSD structures back and forth, while 4 (TR228, TR803, TR827, TR854) of them rise to more than 10 Å away right at the beginning of simulations.

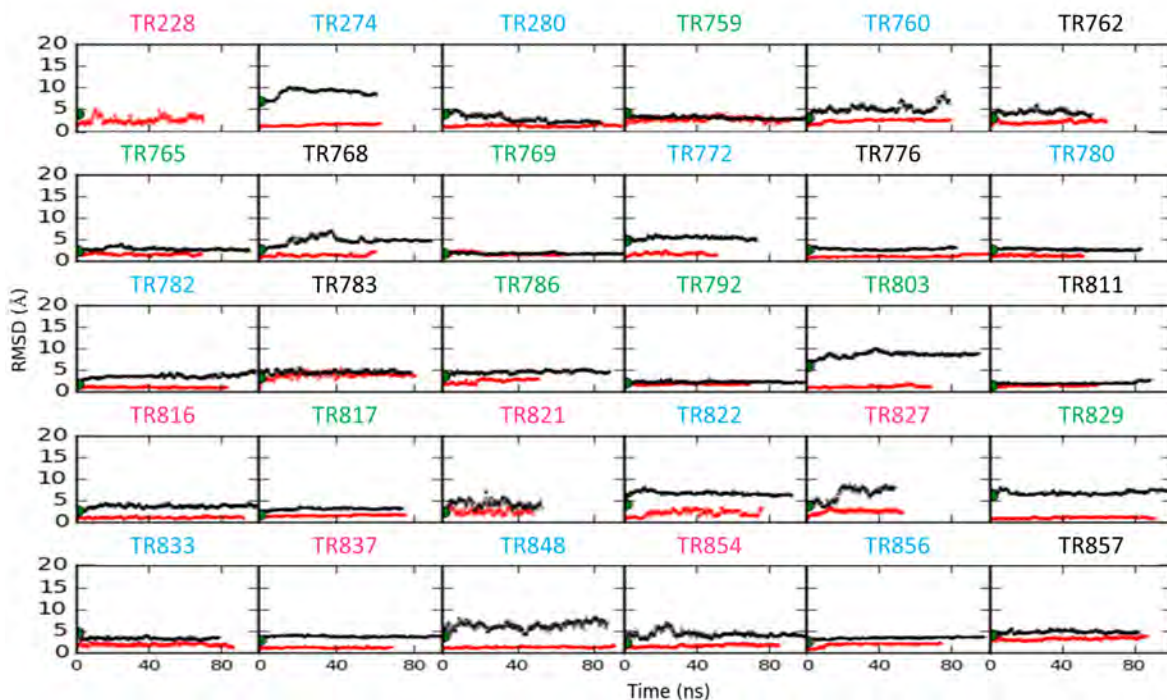


Figure 4.2: $C\alpha$ -RMSD against native structure for refinement and native TIP3P + ff14SB MD simulations. The refinement MD (black line) runs start from initial template structure and the native MD (red line) runs start from equilibrated experimental structures. The green dots at 0 time point in every subplot indicate the various initial RMSD values of template structures. The subplot titles are colored by secondary structures of this protein: mainly anti-parallel beta in blue, helix bundles in magenta, mix of helix & anti-parallel beta in green, parallel beta in black.

As we are not only interested in whether the accuracy of native simulations satisfy the refinement requirement but also evaluating the accuracy of our models at the same time, we take a step off the refinement results but to focus on why the accuracy in some proteins are worse than others. We first categorize all 30 proteins into four groups: anti-parallel beta structures, helix bundles, mix of helix & anti-parallel beta as well as parallel beta structures. Among the 16 stable native structures, 9 (TR274, TR280, TR760, TR772, TR780, TR782, TR833, TR848, TR856) of them are mainly anti-parallel beta structures, 5 (TR759, TR765, TR769, TR817, TR829) of them are mixed helix and anti-parallel beta structures, 2 (TR776, TR857) are parallel-beta, while none of them is helix bundle structure. Among the 4 proteins of > 10 Å deviations, 3 (TR228, TR827, TR854) of them are helix bundle structures, and 1 (TR803) is helix & beta mixture. There are also accuracy issues for the rest 10 proteins

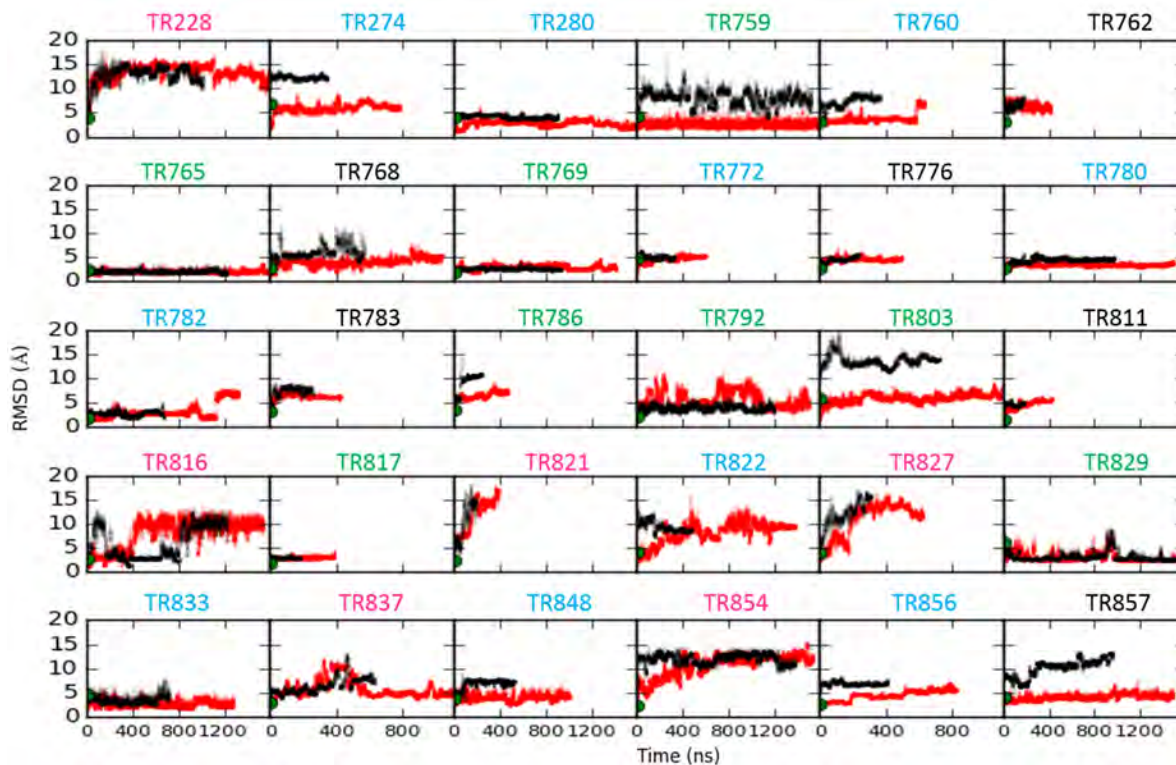


Figure 4.3: $C\alpha$ -RMSD against native structure for refinement and native GBNeck2 + ff14SBonlysc MD simulations. The plotting and denotation fashion are the same as **Figure 4.2**.

with our computational model, but we will focus on the most outstanding problems for now.

The fact that none of the helix bundle structures are stabilized in our simulations means our computational model is specially challenged by proteins abundant in helical structures. There are two aspects of this instability issue: (1) it could be the local helices that are not stable and transformed to other secondary structures; (2) it is also the tertiary structure of relative arrangement between the different helix bundles that go through a thermal unfolding. The local helical propensities point to the issues in backbone parameters, while the more global thermal instability is ascribed to the lack of nonpolar term in implicit solvent. As both issues have been addressed in details in **Chapter 3** and **Chapter 2** respectively, the observations instead of causes and modifications are to be elaborated in this chapter.

We then analyzed the secondary structure conservation for the native simulations. **Table S4.3** and **Figure 4.4** summarize the percentages of helical, extended and coil regions with respect to the percentages in experimental structures. The general trends in each secondary structure fractions are: for helical percentages, nearly one third of the proteins lose helical fractions by more than twice of the corresponding standard deviations; meanwhile more than half of the proteins gain coil percentages; for beta percentages, it is a mix of gain and loss, as the parallel-beta percentages decrease while anti-parallel beta percentages increase.

Two helix bundle proteins (TR854 and TR228) outstandingly loss their helical percentages compared to the fractions in crystal structures, after 400 ns of MD simulations in

ff14SBonlysc and GBNeck2. For TR854, 71% of helical fraction is lost to $47 \pm 9\%$, with the diminished percentages converted to coil structures (which rises from 29% to $48 \pm 9\%$). The overall RMSD changes indicate the loss of critical contacts or helical regions that are responsible for the stability of this protein. It also suggests that although helical structures are not folded all the time, they have not misfolded into other secondary structures. However, for TR228, 79% of helical percentage decreases to $33 \pm 9\%$, with extended percentage increases from 0% to $12 \pm 6\%$ and coil percentage from 21% to $55 \pm 7\%$. This is exceptionally large percentage of helical structures converting into the extended/beta in secondary structures, as it points out to not only the thermal instability but also backbone SSE propensity issues, which have been observed in the helical propensity test carried out in **Chapter 3**.

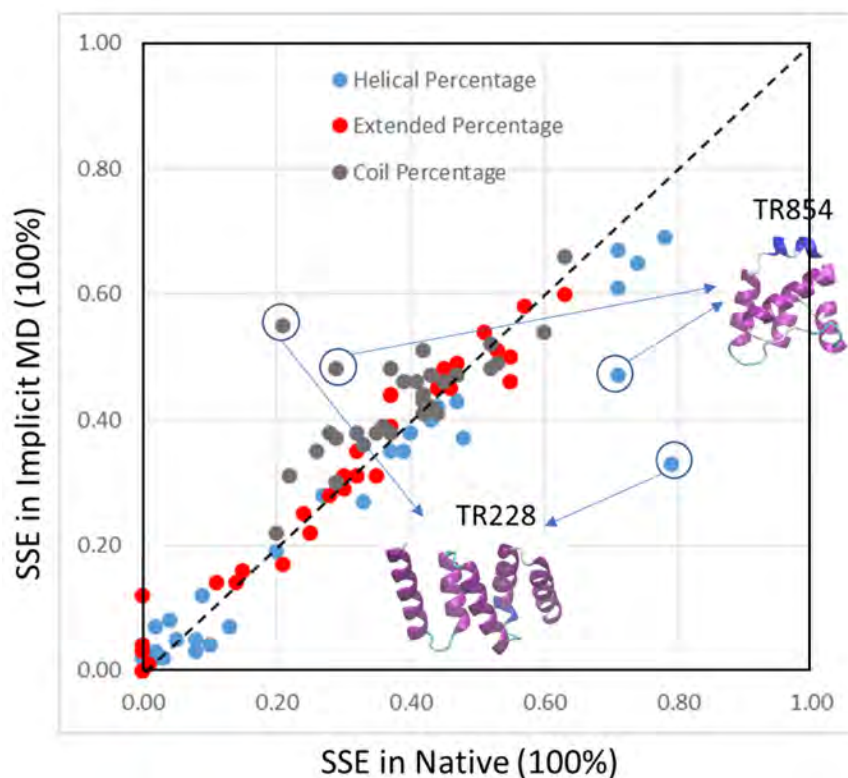


Figure 4.4: Secondary structure elements (SSE) lost indicates stability issue. The SSE (helical in blue, extended in red and coil in gray) percentages measured in DSSP[41] for crystal structures and the first 400 ns of ff14SBonlysc+GBNeck2 MD simulations. The diagonal dashed line indicates perfect agreement.

Even though we have observed inaccuracy in native simulations, we continued to run the refinement simulations from template structures and focus on the systems where we did well in stabilizing the native structures. In cases that sampling might be the issue, we additionally apply REMD as simulated annealing approach to enhance the sampling and accumulate the preferred structures to low temperature trajectories.

4.4.3 Best frame in simulations for CASP11

We summarize all the best sampled conformations for 30 systems with respect to the initial RMSD, calculated as ΔRMSD . The best results reported by CASP participants are listed as a comparison.

As shown in **Table 4.3**, among all the proteins, 23 (all except TR760, TR783, TR803, TR811, TR817, TR822 and TR854) of them have been refined by at least one of the simulations, if only the best frames are examined. If there was just one or two simulations get initial structure refined, we would consider them as anecdotal success as it may not be reproducible well enough, however, in this data set, there are 11 out 23 proteins where we are able to see consistent success comparing across all simulations using GB combined with ff14SBonlysc (MD and REMD) and TIP3P combined with ff14SB (MD at three temperatures). In 8 (TR280, TR765, TR772, TR776, TR780, TR786, TR816, TR833) out of these 11 systems (plus: TR759, TR827 and TR829), we are able to sample better than best-reported from CASP participants.

It is not an overall accomplishment in refinement for us nor a fair comparison for the CASP participants as the best-reported RMSD values are the results predicted without experimental structures available. However, it is noteworthy to see the great potential in implicit solvent simulations as they perform almost as well as explicit simulations, with originally thoughts of cons in accuracy. Among all the 23 cases that ever got refined structures in short simulations, except the 11 cases where better conformation appear in all simulations, half of the refined cases are only found in explicit simulations (TR274, TR821, TR837, TR848, TR856, TR857); in the other half, implicit simulations are also comparable (TR762, TR769, TR782, TR792), and could also take the lead (TR228, TR768) meaning only implicit simulations got refined structures. From our observation, the capability of implicit solvent refining template structures on CASP 11 data set is almost as competent as explicit solvent using the same amount of time and resources.

Although there are 16 of them stabilized below 5 Å in native simulations, the 24 systems that ever got refined do not entirely include the 16 stabilized proteins. For example, TR760 and TR817 have shown great stability in native simulations as seen in **Figure 4.3** are among the 4 systems that have not sampled better conformation than the initial ones. It is very likely that given longer simulation time, better structures could appear in the trajectories, however, the refinement have already become very challenging since both of them start at a very close to native structure RMSD (initial RMSD of TR760 template is 3.14 Å and TR817 is 1.81 Å). But good initial structure is also not the only deterministic factor contributing to the refinement difficulty, in systems such as TR765 (iRMSD = 2.58 Å), TR780 (iRMSD = 2.73 Å) and TR816 (iRMSD = 2.53 Å), even though the initial structures are pretty good, we are still able to sample better conformations in all simulations. Therefore, the factors that lead to the success rate of refinement are a mixture of starting point and secondary structure dependent.

In the next section, we are analyzing the refinement results from a more practical point of view, as in real CASP competition and protein refinement scenario, there is no prior knowledge of experimental structure as for us to pick out the best.

Table 4.3: Best Δ RMSD for #1 rank in CASP and GB/TIP3P MD/REMD simulations in this study
 Note: Δ RMSD measures the change of $C\alpha$ -RMSD against native structure in simulated structure with respect to the initial template model, a negative value indicates the template model is refined. ¹ simulations run at 300 K, ² simulations run at 310 K, ³ simulations run at 320 K

target	AA	$C\alpha$ -RMSD of initial model	Δ RMSD of #1 rank	Δ RMSD of GB REMD	GB MD ¹	TIP3P MD ¹	Δ RMSD of TIP3P MD ²	TIP3P MD ³
TR759	62	4.23	-2.114	-	-1.07	-1.85	-1.49	-1.77
TR829	67	6.16	-4.944	-	-4.19	-0.22	-0.37	-0.13
TR816	68	2.53	-1.376	-1.85	-1.76	-0.28	-0.40	-0.57
TR854	70	2.27	-0.190	0.37	1.33	0.27	0.38	0.12
TR765	76	2.58	-0.335	-	-1.64	-0.52	-0.59	-0.52
TR792	80	1.99	-0.512	0.06	-0.15	-0.24	-0.18	-0.30
TR228	84	3.92	-0.772	-	-1.12	-	-	-
TR780	95	2.73	-0.225	-0.60	-0.15	-0.51	-0.57	-0.37
TR280	96	4.03	-1.024	-	-1.06	-2.40	-0.95	-1.04
TR857	96	4.00	0.523	0.30	0.75	-0.12	-0.11	-0.20
TR769	97	1.74	-0.521	-0.33	0.06	-0.34	0.35	0.11
TR833	108	4.71	-2.456	-	-2.88	-2.12	-1.98	-1.63
TR782	110	1.93	-0.173	-0.61	-0.41	0.27	-0.09	0.03
TR822	121	4.21	0.012	0.36	1.06	1.50	0.74	0.48
TR837	121	2.95	-0.265	0.65	0.32	0.20	0.69	-0.03
TR803	134	5.97	1.271	0.60	1.72	0.67	0.58	0.68
TR848	138	3.78	-1.155	-1.44	-1.05	0.22	-0.11	-0.05
TR768	143	2.61	1.024	-0.55	0.61	0.03	0.21	0.13
TR856	159	2.68	-0.021	0.62	0.88	0.07	-0.03	-0.01
TR827	193	3.75	-0.979	-0.58	-0.09	-0.47	-0.84	-0.97
TR274	194	6.80	-1.433	1.33	2.00	-0.15	-0.22	-0.04
TR772	198	4.78	-0.165	-0.88	-0.93	-0.30	-0.78	-0.31
TR760	201	3.14	-0.025	-	0.56	0.52	0.84	0.31

Continued on next page

Table 4.3 – continued from previous page

target	AA	C_{α} -RMSD of initial model	#1 rank	GB REMD	GB MD ¹	TIP3P MD ¹	TIP3P MD ²	TIP3P MD ³
TR786	217	3.62	0.096	-0.18	-0.04	-0.17	-0.10	-0.13
TR776	219	3.22	-0.037	-1.09	-0.77	-0.94	-0.71	-0.90
TR783	243	3.26	-0.581	0.41	0.69	0.33	0.10	0.21
TR811	251	1.45	-0.259	0.79	0.97	0.11	0.43	0.14
TR821	255	2.45	-0.816	0.01	0.87	-0.24	-0.89	-1.03
TR762	257	3.07	-0.908	-0.05	0.37	-0.06	-0.67	0.52
TR817	265	1.81	-0.111	-	0.34	0.50	0.55	0.63
Avg.		3.41	-0.616	-0.12	-0.16	-0.16	-0.17	-0.17

4.4.4 Larger cluster size indicates higher confidence in refinement

We do RMSD-based cluster analysis on the simulations thus to blindly determine which structure or structural ensemble is more preferred in our physics-based models. **Figure 4.5** summarize all the MD results, the REMD results are also linked with the corresponding MD one for the improved cases.

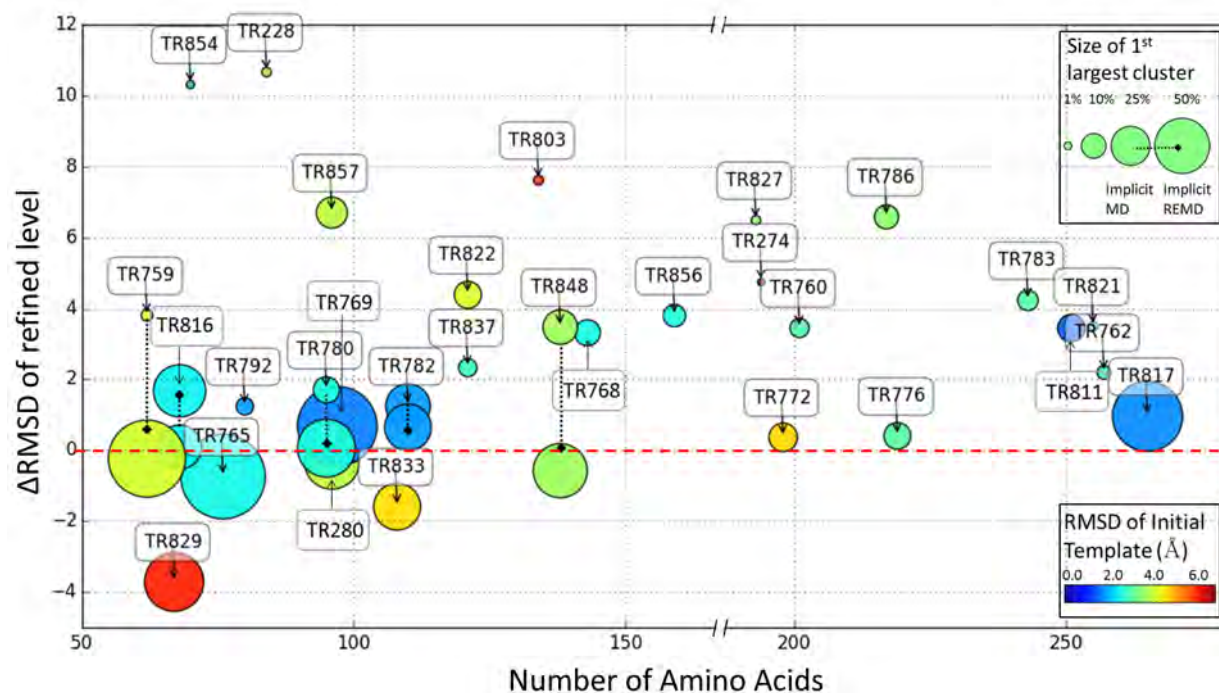


Figure 4.5: Refinement level for each protein considering the top cluster of refinement MD simulations. All 30 systems are displayed as the size of this protein (#AA) vs. $\Delta C\alpha$ -RMSD of refined level. RMSD is calculated from the representative structure for the top cluster, with respect to crystal structure ($RMSD_{crystal}$) or initial template structure ($RMSD_{template}$). Δ RMSD is calculated from the $RMSD_{crystal} - RMSD_{template}$. Δ RMSD below 0 (red dashed line) means this representative structure is refined as it is moved towards the crystal structure compared with the initial template; Δ RMSD above 0 means our refinement runs turn in the structures that are worse than initial template. Each filled circle is labeled with the target name. The size of circle means the size of that cluster (top right legend box). The cool-to-warm colors encode the small-to-large RMSD of initial template against crystal structure (bottom right legend box). MD results are linked to REMD results through black dashed lines with black diamond only on top of REMD circles; 5 systems that REMD improves/maintains the refinement are shown.

As indicated by the sizes of the filled circle close to or below the 0 Δ RMSD baseline, the proteins that achieve overall negative Δ RMSD (meaning would be blindly refined successfully) also possess larger cluster populations. In the cases where MD simulation does not frequently sample a stable conformation (meaning the 1st largest cluster has $< 20\%$ population), REMD simulations improve sampling and return more refined structures, at least not

getting worse, as seen in TR759, TR780, TR782 and TR848. Other cases such as TR829, TR816, TR765, TR769, TR280, TR833, TR817 (listed in the order of chain length), MD simulations are able to sample stable refined/close to refined structures with $> 50\%$ populations; all of these cases overlap well with the systems that could be stabilized in native simulations, which suggests that the stable/refined protein systems are reproducible thus not anecdotal successes.

Again we noted that the initial RMSD values of the templates are unrelated with the refinement confidence. We have discussed in the previous section 4.4.3 that the difficulty for already good template structures (with $< 4 \text{ \AA}$ RMSD) to get refined does not decrease using the best best frame as assessment criterion; here we also do not see apparent correlation between the refinement level and initial RMSD of templates using cluster representatives. However, we should also note that in CASP refinements, the strategy of participants are more conservative and focus on overall improvement with restrained in simulations, which also pitifully limit the potential in MD to rectify the $> 6 \text{ \AA}$ away structures going through large conformational change. As shown in **Figure 4.6A**, a long chain of termini is misplaced in the template structure provided, while in native structure, this termini is forming anti-parallel strand with another beta strand located at the other side of the tertiary structure. If restraints were applied to $C\alpha$ atoms as was done by others [180, 179, 119, 185], this large relocation of a whole termini would not be able to achieve towards a 3.82 \AA refinement.

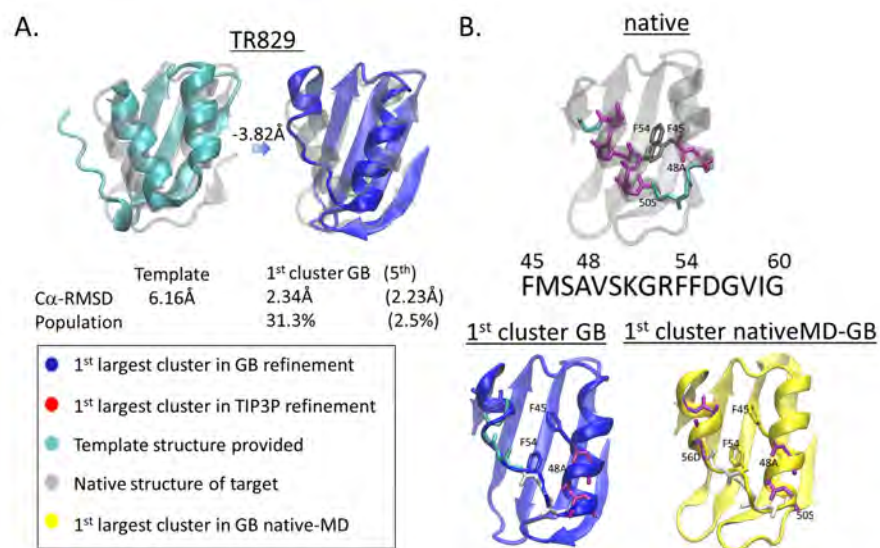


Figure 4.6: TR829 refined from GBNeck2 MD simulations. A. The template structure has gone through a loop relocation and got refined. The $\Delta RMSD$ calculated from $RMSD_{crystal} - RMSD_{template}$ is denoted on the conformation transition arrow. The $C\alpha$ -RMSD value and the population of two clusters (1st cluster and closest cluster) are denoted below the structures. All color codes used for **Figure 4.6**, **4.7** and **4.8** are listed in the box. Blue, red and yellow are for the largest cluster in GB refinement simulation, TIP3P refinement simulation, and GB native simulation, respectively; cyan is for the provided template structure; transparent gray is the native structure overlaid in the background. B. The native structure, sequence of a segment of interest and first cluster structures from both refinement and native simulations.

For the systems that are refined or close to refined determined by cluster analysis, structures overlapping with experimental conformations are shown in **Figure 4.7**. Compared with TIP3P solvent results, the top cluster representatives generated from GB simulations are of similar refinement level in TR280 and TR833. For TR280, two anti-parallel beta strands have been both rectified to the right position with $> 25\%$ of cluster populations, indicating high confidences. For TR833, anti-parallel beta structure with a helical tail has been refined in GB, while of slightly lower $\Delta RMSD$ achieved in TIP3P 1st cluster representative (-1.74 \AA for TIP3P and -1.69 \AA for GB), the helical tail was misrepresented in TIP3P. **Figure 4.7C** and **4.7D** illustrate the refined or close to refined structures observed in the top 3 clusters for TR765 and TR782, respectively. Although 1st clusters are not the best ones, given long simulations, the relative population would likely vary and converge better. Most of the helix bundle systems are observed to be stable in any preferred structure, thus TR821 in TIP3P is shown in 4.7E, which is consistent with the native simulation results in section 4.4.2.

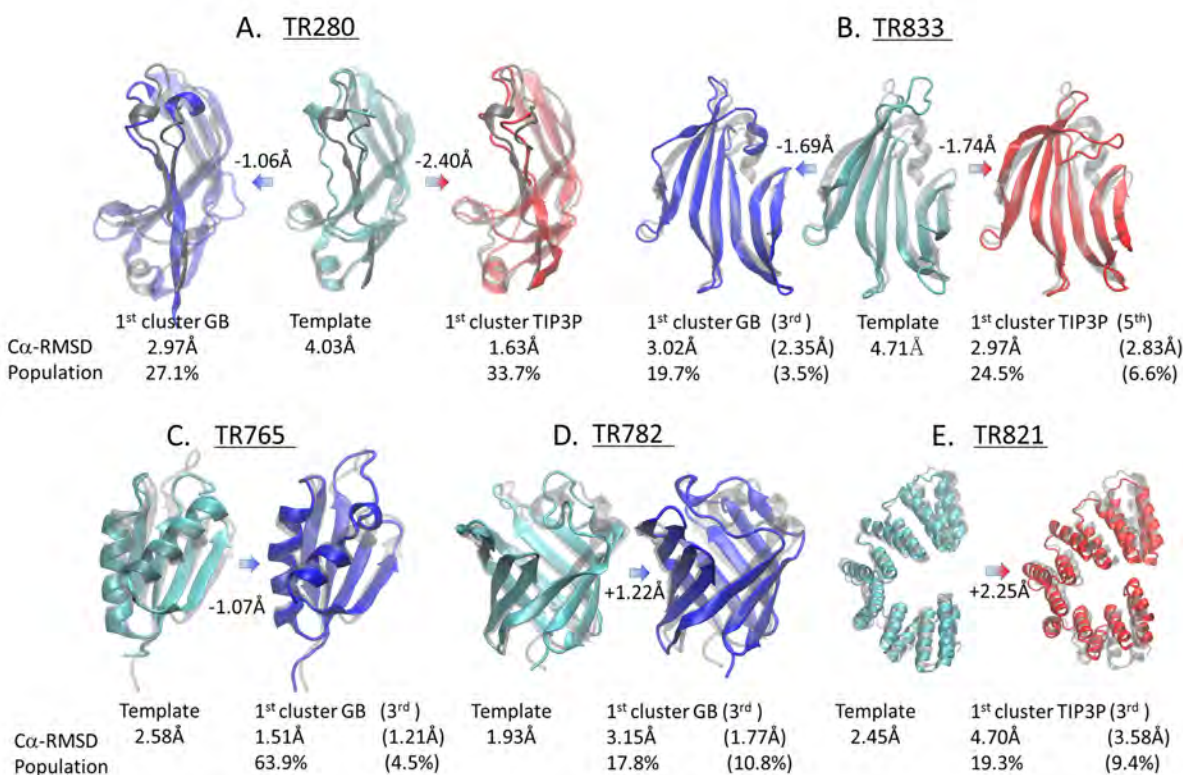


Figure 4.7: Other refined examples from GBNeck2 and TIP3P MD simulations for A. TR280, B. TR833, C. TR765, D. TR782, and E. TR821. The color codes and structural information ($\Delta RMSD$, $C\alpha$ -RMSD and cluster population) denotations are the same as those in **Figure 4.6**.

We also carried out REMD simulations for the systems that MD simulations from native structures were not stable at 300 K. As a more enhanced sampling technique, simulations

running at higher temperatures at the same time are able to jump over large energy barrier and anneal better structures within the same amount of wall clock time. For example, for TR768, only by using REMD-GB the better frames were ever sampled as seen in **Table 4.3**). However, REMD simulations results are not always better than MD ones. For example, as seen in **Figure 4.5**, the 1st cluster of MD simulation for TR816 has lower Δ RMSD than that of REMD simulations. If the largest cluster in REMD is examined, as seen in **Figure 4.8B**, a partially unfolded helical region shows up as a more preferred structure; **Figure 4.8E** also illustrate the point that as more occurrences are observed from the first several clusters, the percentage of frames got refined falls and stays low, until at the very end when the clusters with negative Δ RMSD reappear in the trajectories. This also points out the caveat of more sufficient sampling; if accuracy is doubt-worthy, more effective sampling will only shed light to the errors faster but is not going to resolve the issue. On the contrary, if it is truly the sampling hurdle with less accuracy issue, employing REMD will rescue the systems that predict worse structures from only MD simulations. For example, in TR780 and TR848 as shown in **Figure 4.8A** and **D**, **4.8C** and **F**, the most frequently sampled structures are refined conformations, while insufficient sampling in MD simulations were not able to sample and stabilize these structures as top clusters.

What is as important as getting global tertiary structure right, in high quality protein structure prediction, is to get atomic details right. Closer examinations on the most successful case TR829 alone shed light to the inaccuracy in using implicit solvent. As shown in **Figure 4.6B**, when two helices should form a helix-kink-helix motif, we observe a wrongly elongated helix followed by an extended loop, in both the 1st cluster of refinement and that of native simulation. The instability of helical regions are also observed in TR765 in **Figure 4.7C** and in TR816 in **Figure 4.8B**.

The possible causes, especially for TR829, could be ascribed to (1) the lack of nonpolar term in GB simulations, as the hydrophobic core formed by two nonpolar residue F45 and F54 is abolished in the preferred structures. It is promising that with the incorporation of nonpolar solvation developed in **Chapter 2**, improved stability of proteins would be observed; (2) the helical propensity of residue backbone is not as high as it is supposed to be, as was discussed in **Chapter 3**. A modified backbone parameter set might also be helpful in stabilizing both the global tertiary structure and the local atomic details.

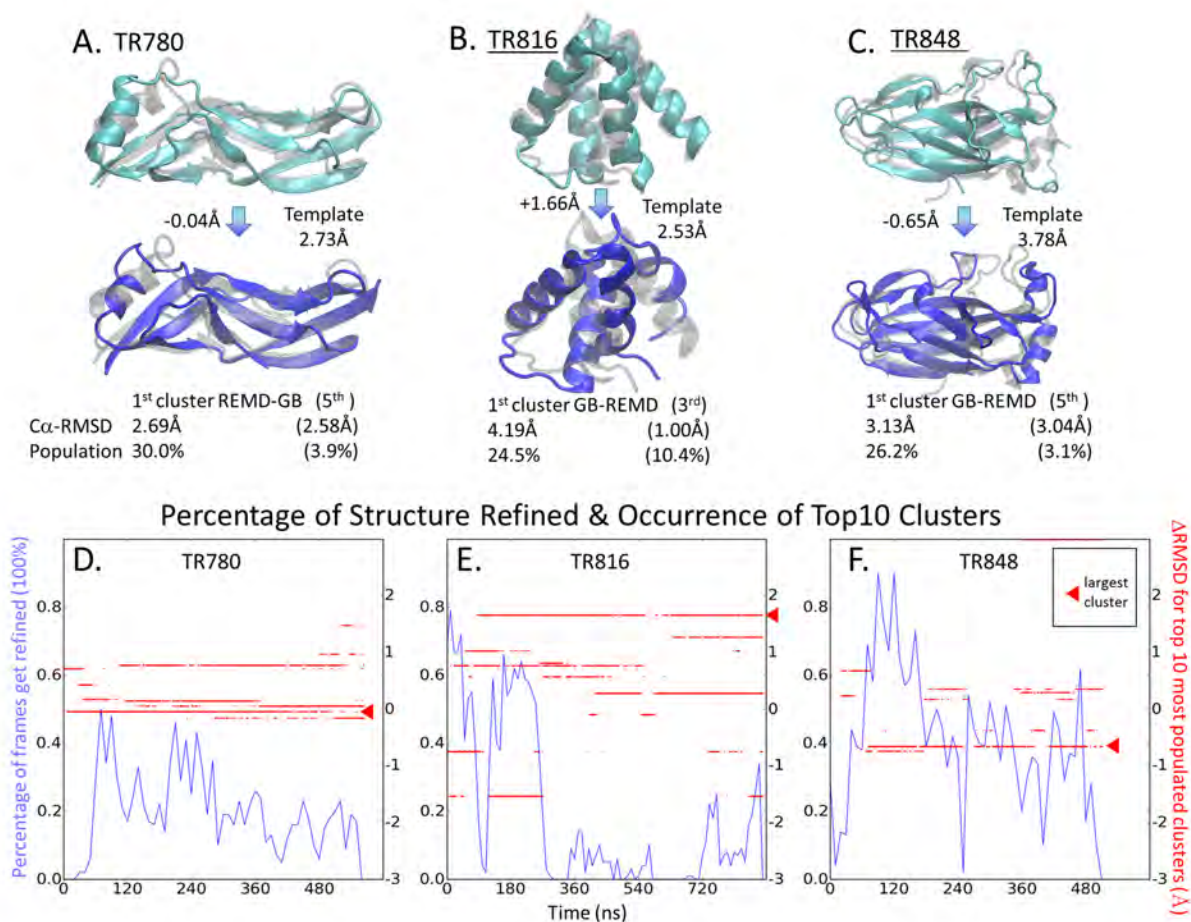


Figure 4.8: Refined examples from REMD simulations. Two sub-figures in a column are for one protein system, A and D for TR780, B and E for TR816, C and F for TR848. For each system, the top sub-figure shows the structures of template structure and 1st cluster structure overlaid with crystal structure, respectively. $\Delta RMSD$, $C\alpha$ -RMSD and cluster population are denoted. The bottom sub-figure is dual y-axis chart as a function of time (ns). The left y-axis is the percentages of structured refined (blue line), which is calculated by percentages of frames with $\Delta RMSD < 0$ using a sliding window of 10 ns. The right y-axis is the $\Delta RMSD$ for all top10 clusters (red dots); dots are present when the structures corresponding to these time points show up in one of the top 10 clusters, the largest population cluster is indicated by a red, left-pointed triangle sign.

4.5 Conclusions

In this chapter, two types of protein models are strategically refined using MD simulations. For the initial structures from GB-folded conformations, three proteins BBA, HP36 and proteinB are refined using explicit solvent. As the inefficiency in explicit solvent MD simulations limiting conformation sampling effectively, implicit solvent are used for CASP 11 refinement test. Different from applying restraints used in previous studies, we run short

MD simulations without restraints. Half of the 30 CASP11 proteins are stabilized in simulations starting from experimental structures. The systems that are not stable are analyzed for secondary structure compositions and are found mostly consisting of helical conformations, which point out the weakness of helical structures in the computational model. 23 out of 30 systems have sampled refined structure while cluster analysis results indicate a strong correlation of cluster size and refinement confidence. When global topology is refined, the local structural inaccuracy points to the future directions for model improvement.

4.6 Supporting Information

Table S4.1: Summary of all CASP11 experimental structures in refinement category

Note:

1. All the starting models are from http://www.predictioncenter.org/download_area/CASP11/targets/casp11.TR_startmodels.tgz, all the experimental structures are from http://www.predictioncenter.org/download_area/CASP11/targets/casp11.domains_official.release11242014.tgz, PDB_code are from <http://www.predictioncenter.org/casp11/targetlist.cgi>.
2. N/A in the pdb_code column indicates there is no related pdb_code recorded in the online CASP11 archive.
3. Except TR769 and TR857 are solved from NMR, all the rest experimental structures are solved in crystallography and the resolutions are shown in the unit of Å.

Target	AA	pdb_code	Resolution	Structure features
TR217	224	4wed (271-494 in pdb)	2.35	mix of helix, antip.beta
TR228	84	N/A		helix bundle
TR274	194	4qb7	2.74	antip.beta with one parallel
TR280	96	4qdy (135-230 in pdb)	2.74	antip.beta with little helix
TR283	168	N/A		parallel beta protected by helix (pbh)
TR759	62	4q28 (46-107 in pdb)	2.64	mix of helix, antip.beta
TR760	201	4pqx	2.39	antip.beta barrel + sandwich
TR762	257	4q5t	1.91	parallel pbh, antip.beta pbh mix
TR765	76	4pwu	2.45	mix of helix, antip.beta
TR768	143	4oju	2	parallel sheet (amyloid like)

Continued on next page

Table S4.1 – continued from previous page

Target	AA	pdb_code	Resolution	Structure features
TR769	97	2mq8	NMR	mix of helix, antip.beta
TR772	198	4qhz	2.13	antip.beta sandwich
TR774	155	4qb7	2.55	antip.beta with little helix
TR776	219	4q9a	2.86	parallel beta pbh
TR780	95	4qdy (10-134 in pdb)	2.74	antip.beta with little helix
TR782	110	4qrl	1.79	antip.beta barrel
TR783	243	N/A		parallel pbh and protected by antip.beta
TR786	217	4qvu	2.65	mix of helix, antip.beta
TR792	80	N/A		mix of helix, antip.beta
TR795	136	N/A		antip.beta, barrel/sandwich (no expl. In folder)
TR803	134	N/A		mix of helix, antip.beta (one helix very long)
TR810	243	N/A		parallel sheet barrel pbh
TR811	251	N/A		parallel sheet barrel pbh
TR816	68	N/A		helix bundle
TR817	265	4wed (36-270,495-524 in pdb)	2.35	mix of helix, antip.beta
TR821	255	4r7s	2.39	helix bundle (super helix)
TR822	121	N/A		antip.beta sandwich (starting is so off)
TR823	296	N/A		parallel beta barrel pbh
TR827	193	N/A		helix bundle
TR828	84	N/A		antip.beta (no expl. in folder)
TR829	67	4rgi	1.73	mix of helix+antip.beta
TR833	108	4r03	1.5	flat antip.beta
TR837	121	N/A		helix bundle
TR848	138	4r4g (34-171 in pdb)	2.62	antip.beta
TR854	70	4rn3 (24-93 in pdb)	2.15	helix bundle

Continued on next page

Table S4.1 – continued from previous page

Target	AA	pdb_code	Resolution	Structure features
TR856	159	N/A_whole		antip.beta sandwich
TR857	96	2mqc (6-101 in pdb)	NMR	parallel, antip.beta mix

Table S4.2: Temperature ladders for the CASP11 REMD simulations

target	REMD temperatures (K)
TR274	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR762	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR768	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR769	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR772	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR776	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR780	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR782	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR783	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR786	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR792	256.9, 265.0, 273.3, 281.9, 290.8, 300.0
TR803	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR811	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR816	256.9, 265.0, 273.3, 281.9, 290.8, 300.0
TR821	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR822	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR827	258.0, 262.4, 266.8, 271.3, 275.9, 280.6, 285.3, 290.1, 295.0, 300.0
TR837	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR848	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR854	256.9, 265.0, 273.3, 281.9, 290.8, 300.0
TR856	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0
TR857	256.5, 262.3, 268.2, 274.3, 280.5, 286.9, 293.4, 300.0

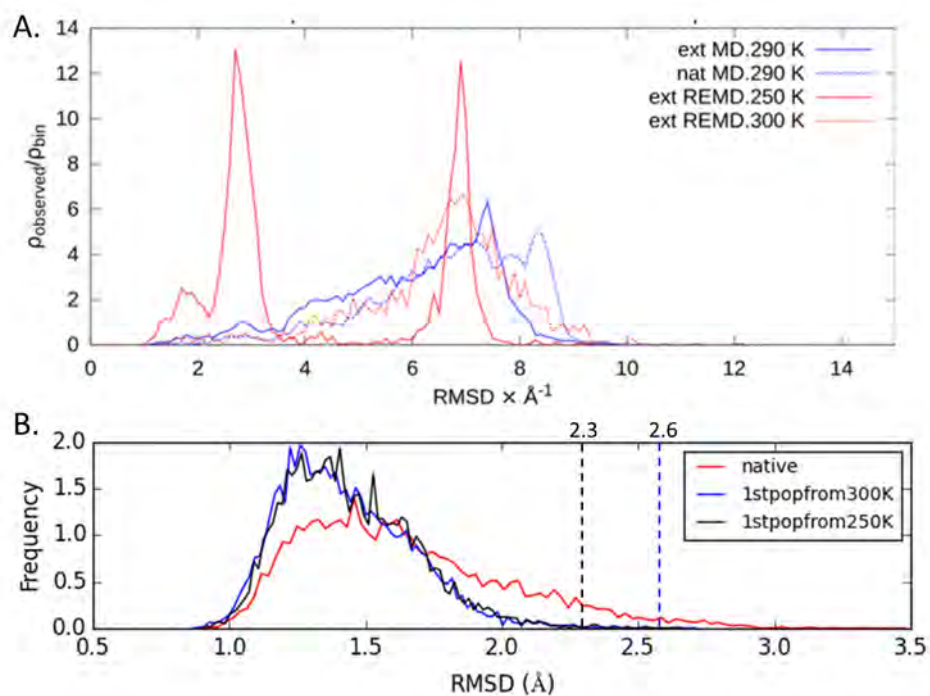


Figure S4.1: HP36 GB folding refinement, A. RMSD distributions of GB folding simulations; B. RMSD distributions of refinement simulations. The color codes are the same as **Figure 4.1**

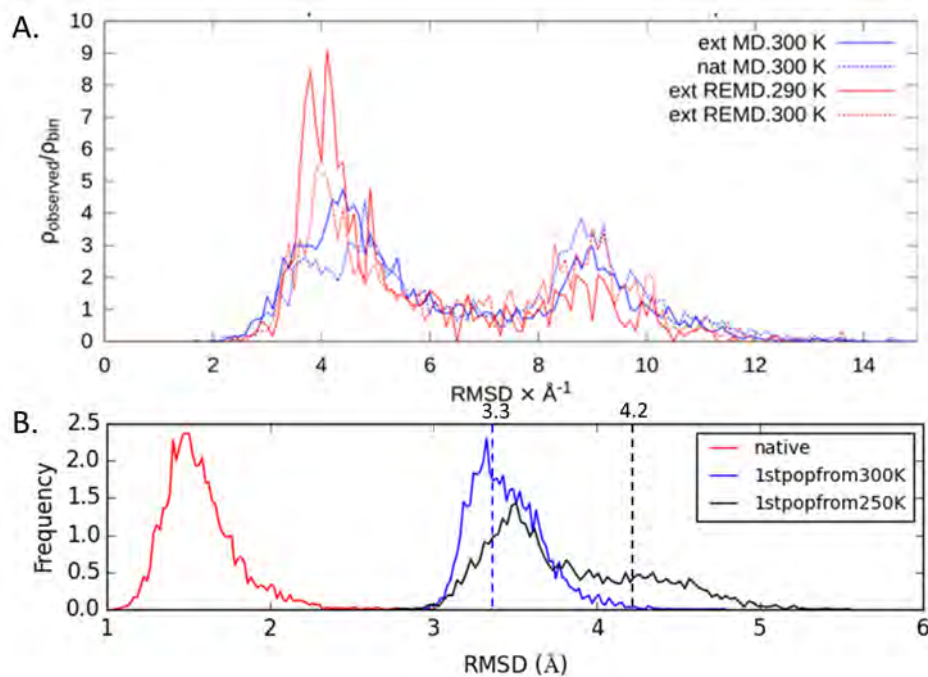


Figure S4.2: proteinB GB folding refinement, A. RMSD distributions of GB folding simulations; B. RMSD distributions of refinement simulations. The color codes are the same as **Figure 4.1**

Table S4.3: Secondary structures conserved in 400 ns of GB-MD simulations compared with the percentages in native structures

Target	AA	Helical percentage		Extended percentage		Coil percentage	
		native	MD-GB	native	MD-GB	native	MD-GB
TR228	84	0.79	0.33(0.09)	0.00	0.12(0.06)	0.21	0.55(0.07)
TR274	194	0.02	0.07(0.04)	0.46	0.45(0.02)	0.52	0.48(0.04)
TR280	96	0.05	0.05(0.03)	0.51	0.54(0.03)	0.44	0.41(0.04)
TR759	62	0.37	0.35(0.03)	0.30	0.29(0.04)	0.33	0.36(0.05)
TR760	201	0.08	0.05(0.02)	0.47	0.49(0.02)	0.45	0.46(0.03)
TR762	257	0.39	0.35(0.02)	0.24	0.25(0.01)	0.36	0.39(0.02)
TR765	76	0.32	0.31(0.03)	0.32	0.31(0.02)	0.35	0.38(0.04)
TR768	143	0.02	0.03(0.02)	0.35	0.31(0.02)	0.63	0.66(0.02)
TR769	97	0.43	0.40(0.04)	0.37	0.39(0.02)	0.20	0.22(0.04)
TR772	198	0.03	0.02(0.02)	0.37	0.44(0.03)	0.60	0.54(0.04)
TR776	219	0.47	0.43(0.03)	0.11	0.14(0.02)	0.42	0.43(0.03)
TR780	95	0.09	0.12(0.04)	0.55	0.50(0.03)	0.37	0.38(0.04)
TR782	110	0.10	0.04(0.03)	0.57	0.58(0.02)	0.32	0.38(0.03)
TR783	243	0.40	0.38(0.03)	0.21	0.17(0.01)	0.39	0.46(0.03)
TR786	217	0.27	0.28(0.02)	0.30	0.31(0.01)	0.42	0.41(0.03)
TR792	80	0.44	0.42(0.05)	0.14	0.14(0.02)	0.42	0.44(0.06)
TR803	134	0.33	0.27(0.04)	0.25	0.22(0.03)	0.42	0.51(0.05)
TR811	251	0.48	0.37(0.04)	0.15	0.16(0.01)	0.37	0.48(0.04)
TR816	68	0.71	0.67(0.07)	0.00	0.03(0.03)	0.29	0.30(0.07)
TR817	265	0.20	0.19(0.02)	0.28	0.28(0.01)	0.52	0.52(0.02)
TR821	255	0.78	0.69(0.04)	0.00	0.00(0.00)	0.22	0.31(0.04)
TR822	121	0.04	0.08(0.05)	0.55	0.46(0.05)	0.41	0.46(0.04)
TR827	193	0.71	0.61(0.03)	0.01	0.01(0.01)	0.28	0.38(0.04)
TR829	67	0.25	0.22(0.05)	0.32	0.35(0.02)	0.43	0.42(0.05)
TR833	108	0.08	0.03(0.02)	0.63	0.60(0.03)	0.29	0.37(0.04)
TR837	121	0.74	0.65(0.05)	0.00	0.00(0.00)	0.26	0.35(0.05)
TR848	138	0.13	0.07(0.02)	0.44	0.45(0.02)	0.43	0.47(0.03)
TR854	70	0.71	0.47(0.09)	0.00	0.04(0.03)	0.29	0.48(0.09)
TR856	159	0.02	0.02(0.02)	0.45	0.48(0.02)	0.53	0.49(0.03)
TR857	96	0.00	0.02(0.03)	0.53	0.51(0.03)	0.47	0.47(0.04)

Chapter 5

Study on Mechanism of IAPP Amyloid Fibril Initial Formation

5.1 Abstract

Islet Amyloid found in pancreas is correlated with the Type 2 Diabetes (T2D) Disease. The major protein component of Islet Amyloid is Islet Amyloid Polypeptide (IAPP) or amylin whose aggregation structure, mechanism and cytotoxicity *in vitro* and *in vivo* are still unclear. With growing knowledge in both theoretical approaches and experimental techniques in the field, we tried to validate the importance of α -helical intermediates in the mechanism of IAPP amyloid formation. In this project, based on the experimentally postulated mechanism of IAPP amyloid fibril initial formation, we applied the Molecular Dynamics (MD) simulation method to the IAPP monomers and dimers, which provide more understanding of the characteristics of monomeric IAPP and how IAPP molecules initially aggregate. These findings will provide insight and experience as to further research needs.

5.2 Introduction

Proteins that form fibrils usually belong to fibrous proteins. However, different from other fibrous proteins that commonly have a structural, supportive or motility role, amyloid-forming proteins are often found in organs and tissues as disease-related, abnormal fibrous, extracellular, and proteinaceous deposits[186]. The deposits are mainly composed of elongated fibers, with spines consisting of many-stranded β sheets, called amyloid fibrils. Amyloid-forming proteins have been identified and associated with a range of serious diseases, including amyloid- β peptide ($A\beta$) with Alzheimer's disease, prion protein (PrP) with the spongiform encephalopathy (e.g. Mad Cow disease) and islet amyloid polypeptide (IAPP) with type 2 diabetes (T2D)[187]. Because of the important roles amyloid fibrils play in human diseases, studies into the structures and properties of amyloid fibril and its formation are widely conducted.

Amyloid deposit found in the extracellular pancreases of patients is the hallmark of Type 2 Diabetes, T2D. IAPP, also known as Amylin, is the major proteinaceous component of islet amyloid. It is a 37-residue polypeptide pancreatic hormone, which has been found in

all mammals that have been studied so far[188]. This peptide is produced and stored in the β -islet cells of the pancreas. It is related to the pathology of T2D because $> 90\%$ of T2D patients exhibit amyloid plaques in their pancreas, and the severity of the disease correlates with the degree of plaque deposition[189].

Amyloid fibrils have been found to display the characteristic cross- β fiber diffraction pattern. This pattern was first observed by William Astbury in 1935[190]. With gradually advanced techniques and methods, such as solid-state NMR, model-building constrained by X-ray diffraction, people have known the most general common features among these fibrils. (1) In all amyloid fibrils, the strongest repeating feature is a set of β sheets that are parallel to the fibril axis, with extended strands nearly perpendicular to the axis, known as a protofilament. (2) The β sheets can be either parallel or anti-parallel. (3) The sheets are usually “in register”, meaning that strands align with each other such that identical side chains are on top of one another along the fibril axis, called a steric-zipper structure[187]. Like other amyloid species, we do not yet have full atomic structures for Islet amyloid fibrils, there are only models proposed for the fibril structures. **Figure 5.1** (Amyloid fibrils structure) shows a structural model of the full length IAPP fibril proposed by Eisenberg *et al.* denoted in a review article[188].

More and more evidence recently points the cytotoxicity of disease-related deposits to the oligomers of amyloid instead of the full length amyloid fibrils. In the case of Islet amyloid, it has been suggested that the toxic oligomers are the factors causing cell membrane disruptions[191]. Therefore, studies into the process of formation of amyloid fibrils is of great importance. One of the reasons is that the mechanism of how the monomeric peptides aggregate into insoluble and stable fibrils may provide possible therapeutic methods to avoid even reverse the fibril formation process, in order to eventually heal the related diseases.

The formation process is thought to share common features with the crystallization process, in which a slow lag phase is followed by a fast growth phase[191]. That conversion of soluble monomeric peptide to insoluble amyloid fibrils often involves partially unfolded intermediates [194]. Chemical cross linking studies reported that the production of mature fibrils can go through some metastable intermediates, including the very early species: dimers, trimers, tetramers, and higher order oligomers, but the results are sometimes conflicting with one another[188]. In this project, based on the knowledge of the early state intermediates up to date, we are trying to explore the structures and characteristics of monomeric and dimeric state of IAPP, in order to provide further insight for mechanism studies into fibril initial formation.

IAPP is soluble and intrinsically unfolded in its monomeric state, but unaggregated IAPP monomers do not adopt a classic random coil. The region of residues 5-20 of IAPP has been observed via NMR to adopt helical ϕ, ψ angles in solution, although at low level of persistency[195, 196]. Although early helical intermediates and their species have not been identified so far in IAPP amyloid formation, there is evidence that directly or indirectly implies that the α -helical regions of IAPP promote early dimerization. One study has induced a persistent helical structure of monomeric IAPP by negatively charged model membranes at a physiological pH[197, 198]. Within the negatively charged membrane environments in the presence of detergent micelles, the angle between the N- and C-terminus helices constrained to 85° , the structure of IAPP at a neutral pH was solved via NMR[192]. It has an overall kinked helix motif, with residues 7-17 and 21-28 in a helical conformation, with a 3_{10} helix

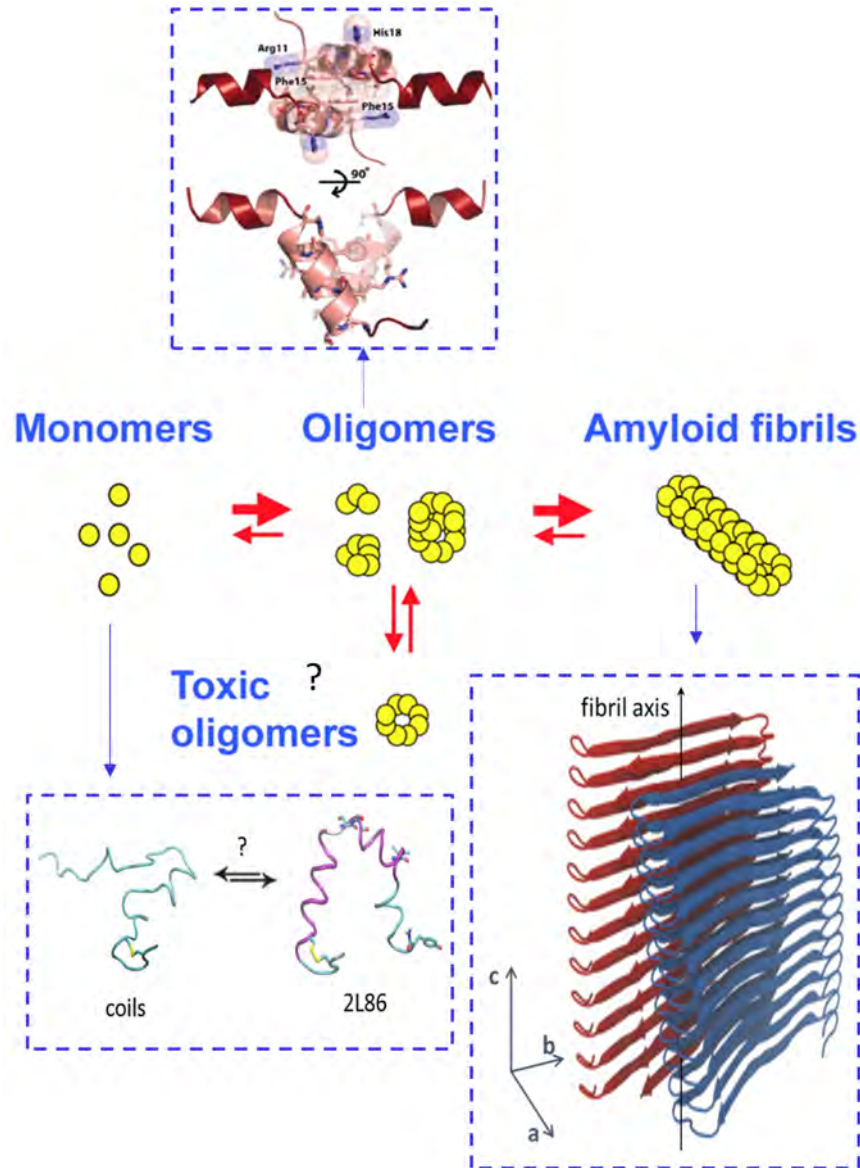


Figure 5.1: The proposed fibril formation mechanism analogous to the crystallization process: soluble monomers aggregate and form oligomer core, which is further elongated to full amyloid fibrils[191]. The experimentally proposed/characterized monomer (PDB code: 2L86[192]), oligomer (PDB code: 3G7V[193]) and amyloid fibril structures[188] are pointed to in blue dashed frames.

from Gly33-Asn35, and Ser19 and Ser20 locate in the kink region (monomer structure in **Figure 5.1**). Since we do not have IAPP monomeric structure in solution, this NMR structure could reasonably work as a reference structure for RMSD calculations, and provide full-length coordinates for building monomeric and oligomeric models concerning α -helical intermediates. In another study, the crystal structure of a C-terminal truncated fragment of IAPP fused to MBP[193] (a 370-residue maltose binding protein) can provide some insight to dimeric state of IAPP (oligomer structure in **Figure 5.1**). Each of the monomers

shows that IAPP can adopt an α -helical structure at residues 8-18 and 22-27, which have similar regions compared with NMR structure (residues 7-17, 21-28). Taken together, all the evidence has suggested the helical dimerization of IAPP may initialize fibril formation, which could be a rational intermediate on/off the pathway of amyloid fibril formation.

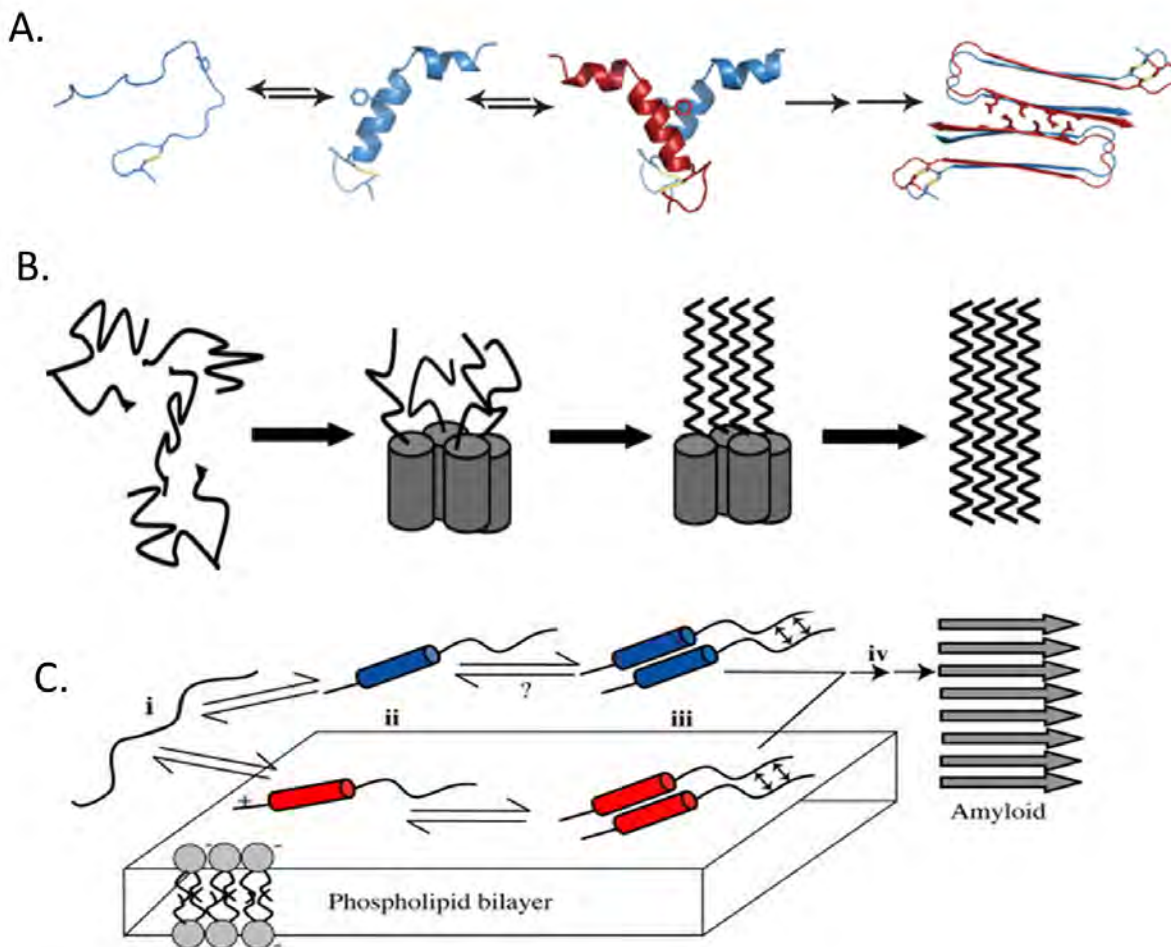


Figure 5.2: Schematic diagrams of how an α -helical intermediate might promote amyloid formation. A. disordered and α -helical monomers are shown in equilibrium, followed by dimerizations through N-terminal helices, which lead to fibril formation[193]; B. Initial oligomerization is driven by the thermodynamic linkage between self-association of α -helix regions, which generates a high local concentration of α -structure, promotes their stability and eventually leads to the formation of β -sheet-rich assemblies [199]. α -helices (N-termini) shown as cylinders, β -strands (C-termini) as zigzagged lines. C. in both solution and membrane environment, similar to the processes in A and B, α -helical regions in cylinders form parallel oligomers that facilitate amyloid nucleation and mature fiber formation[196]. All three diagrams are adapted from corresponding publications. They are schematic and not meant to imply a specific pathway of assembly.

Among all the hypotheses of oligomerization of pre-fibrillar structures, we see common grounds in the α -helical intermediates thus we hold the hypothesis that α -helical structures

are important for initiation of IAPP: the α -helical regions first associate to increase the local concentration, then a phase transition appears and propagates throughout the whole chain to finally form the fibrils. To be more specific, the initial aggregation of islet amyloid might be driven by the formation of an oligomeric helical intermediate with helical structures in the N-terminal region of IAPP. Once they have associated and interacted with each other, a high local concentration of amyloidogenic C-terminal segment is developed and leads to intermolecular β -sheet formation which then propagates through the sequence (**Figure 5.2**)[199].

In all mammals studied to date, IAPP is present and the sequences are largely conserved, however, not all IAPP species are found to form amyloid[200, 201]. Humans, non-human primates and cats form amyloid while rats and mice, which share the same 31 residues out of 37 with humans, do not form amyloid in their pancreas. Comparative experiments carried on rat IAPP (rIAPP) and human IAPP (hIAPP) have convinced, to some degree that the 6 different residues, especially the 3 Proline residues at position 25, 28 and 29 cause rIAPP not to form amyloid[201]. Proline is known as a cyclic amino acid which often acts as a secondary structural disrupter and also it is unfavorable in a β -sheet from the energetic point of view.

The only mutation naturally found in human at a low level is the substitution of Serine at position 20 by a Glycine residue (i.e. single mutant S20G). This variant, which has been shown to accelerate amyloid formation *in vitro*, has been suggested to cause a higher risk of T2D even though only to a slightly higher degree[202, 203]. When carried out in experimental studies, the acceleration degree of amyloid formation is notably higher than wild type IAPP and S19G variant (**Figure 5.3**)[204]. *In vitro* experimental kinetics measurements indicate that S20G variant greatly reduces the lag phase of amyloid formation, and leads to similar morphology of full amyloid fibrils. Studies so far explain S20G accelerating rate by an increased propensity to form a β -turn which resembles the β -sheet structure in the fibrils. Glycine are turn promoters and residue 20 that is involved in this mutation lies right in the turn region of the hairpin (residues 18–22)[205]. Other recent work has reported that in Alzheimer’s A β peptide amyloid formation acceleration, stabilization of the turn structures may be responsible[206], which could also possibly be true in the case of IAPP.

However, the experimental methods have their own limitations. Different experimental conditions lead to different, incomparable results. Part of the reason is the kinetics of IAPP aggregation can be very sensitive to very small changes in sample preparation, buffer composition, pH, even stirring frequencies during the measurement[188, 207]. In addition, mutagenesis approach in amyloid formation study makes things more complicated compared to the case of soluble globular proteins because the formation of different polymorphs and the difficulty of fibril stability determination[188]. For the crystal structures that experimentalists have obtained, they may also fail to convince people because the consequence of crystal packing leads to artifacts and those structures may not reliable models of *in vivo* structures. Furthermore, current standard spectroscopic methods lack both structural resolution and/or time resolution which leaves a gap of knowledge between monomer and fibril of IAPP during its formation process. In order to fill the understanding of how IAPP monomers of α -helical propensity aggregate into β -sheet-rich fibrils, we employ theoretical and computational methods, mainly molecular dynamics (MD) simulations.

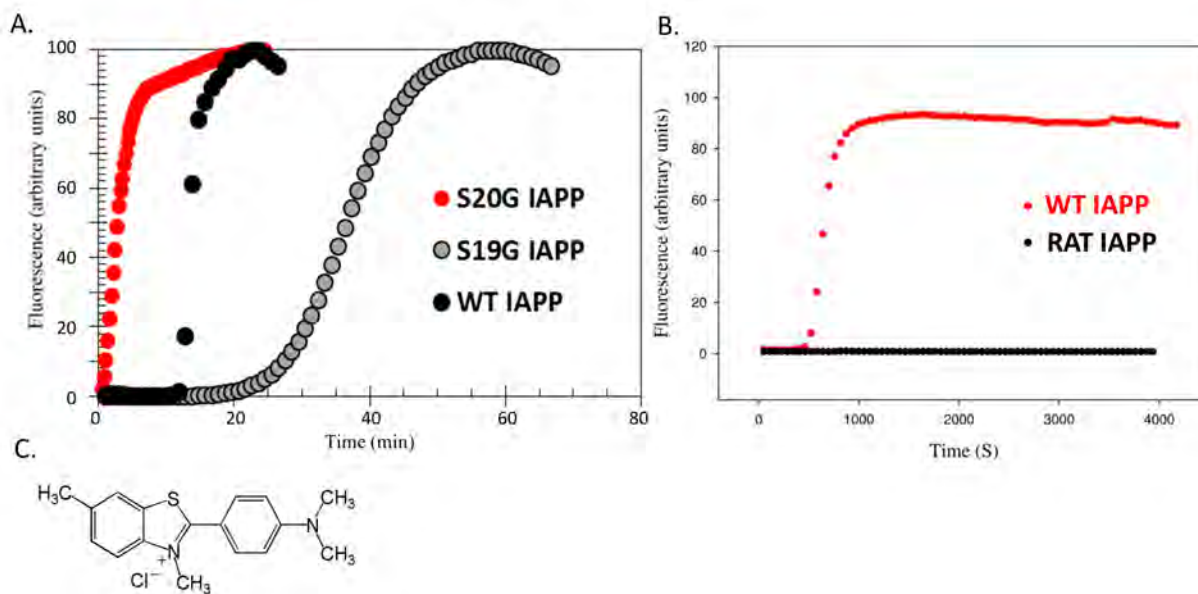


Figure 5.3: The fibril formations rates of IAPP and variants monitored by Thioflavin T fluorescence intensity. A. The fluorescence intensity changes as a function of time for WT IAPP (black dots), S20G (red dots) and S19G (gray dots). Data is from Rehana Akter in the Dr. Daniel Raleigh’s lab, which has been published in Rehana’s Master Thesis[204]. B. The fluorescence intensity changes as a function of time for WT hIAPP (red) and rat IAPP (black). Data from Ref[208]. C. Molecular Structure of thioflavin T. It can be used as reaction rate monitor because when it binds to β sheet-rich structures, such as amyloid fibrils in aggregation, the dye displays enhanced fluorescence and a characteristic red shift of its emission spectrum.

In previous MD studies, the various components on the pathway to form amyloid, from monomers[205, 209, 210, 211, 212], dimers[211, 213, 214], to fibrils[212, 215] of both hIAPP and rIAPP have been modeled in solution and also in membrane environment[216, 217]. Due to various methods and computational models used in simulations, the structural ensembles of aqueous IAPP monomer *in silico* behave differently. Besides the random coil of IAPP conformers that are consistently observed throughout all simulations, there are two other major conformers being identified: a compact helix-coil/hairpin structure and an extended β -hairpin structure. However, different results disagree on the relative ratio and the specific structures of each conformer. One of these findings disagree on the possible structural components of proposed IAPP dimeric states. To be more specific, when IAPP monomers dimerize, it is arguable whether dimers are initiated from helical monomers or hairpin structures; the question of which conformer do the IAPP dimers adopt leads to contradictory mechanisms[213, 214]. One of the mechanisms was proposed and supported by Shea and coworkers[213, 209]. Based on the assumption that β -hairpin structure sampled in monomer simulation is a possible direct amyloidogenic precursor to β -sheet-rich fibril formation, they proposed that dimerization of side-by-side assembly of β -hairpin monomers is on pathway to form β -sheet rich oligomers. However, this dimer structure (shown in **Figure 5.4B**) is associated through hydrogen bonding between the N-terminal residues of one β -hairpin

monomer and the C-terminal residues of another β -hairpin monomer; this N-terminus-C-terminus interface is inconsistent with the mature fibril structures as shown in **Figure 5.1**, the fibril model requires a in-register alignment of dimers (or oligomers) as shown in **Figure 5.4A**. In contrast, we have hypothesized the important roles of α -helical structures in the monomeric and dimerization states, which is consistent with the helix-coil structure regime[199, 193, 196, 192, 187].

To validate the importance of helical structures in IAPP dimerization, we explore two ways of building a dimer model. One starting point of an IAPP dimer model is the crystal structure obtained by Eisenberg and coworkers in 2009 (3G7V[193] shown as the oligomer structure in **Figure 5.1**, i.e. **Figure 5.4C**). It is an asymmetric unit that contains four fusions of C-terminal truncated hIAPP connecting to maltose binding protein (MBP). With the MBP chaperoning IAPP, a dimer structure was found to form by the N-terminal helices from two IAPP molecules packing against each other with key contacts being made near Phe 15, with 8–18 helices interacting at a 55° angle[193]. A second idea of IAPP dimer model is inspired by the putative intermediate model illustrated in **Figure 5.2C**, **Figure 5.4D** is adapted from it. Although never observed in solved structures nor from previous simulation results, this schematic dimer model embodies the association through α -helical regions followed by a parallel-sheet phase transition, which could be an important precursor of full-chain parallel-sheet structures as seen in the mature fibrils. These two ways of building a dimer model for understanding hIAPP low-order oligomerization are reasonable and worth exploration as they both emphasize the necessity of α -helical intermediates and it is through the associations of helical regions that processes relevant to oligomerization proceed. The first α -helical model (**Figure 5.4C**) is atomic-detailed and the second N-terminal α -helix with C-terminal extended model (**Figure 5.4D**) is more conceptual, which means they may not be exclusive of each other. However, we study them separately. The α -helical model assumes dimer interface consisting of two N-terminal α -helical regions at an angle of 55° , while the N-terminal α -helix with C-terminal extended model assumes the importance of N-terminal α -helices close in distance, as well as the C-terminal regions close as extended conformations, so that a parallel-sheet could be facilitated in the amyloidogenic C-terminal segments.

In this work, it is not our goal to settle the disagreement of monomer conformations and compositions by applying one or two more computational models. Instead, we ask the questions (1) whether MD simulations of IAPP and variants in monomers could validate the importance of α -helical conformers and (2) how that would facilitate our understanding of IAPP fibril initialization mechanism by further validating low-order oligomers (dimers) models. To carry out the investigations, we first focus on only WT hIAPP monomer characterizations. Then we present the structural and energetic differences found by simulations for IAPP monomer variants. Lastly, we demonstrate how we built dimer models to understand the oligomerization of WT hIAPP mapping to our hypothesis that α -helical structures are important for initiation of IAPP fibrils.

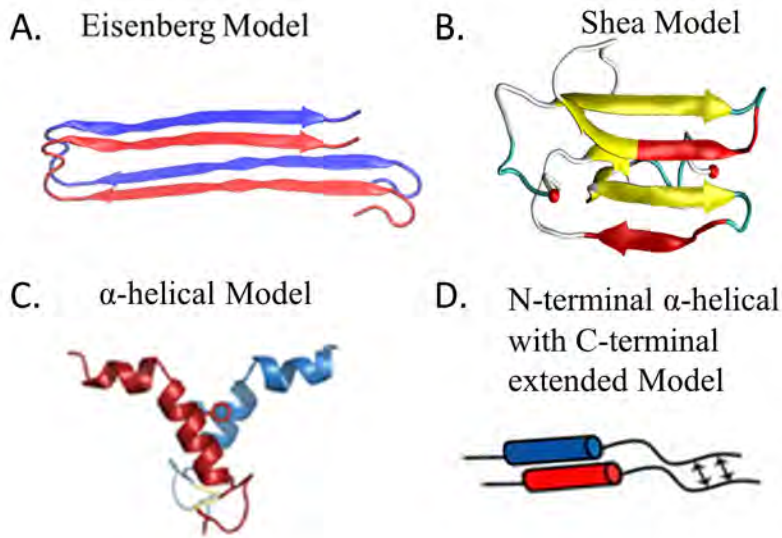


Figure 5.4: Four possible dimer models proposed by others (B and C) or hypothesized in this study (A and D). A. The Eisenberg dimer model isolated from the full fibril structure proposed by Eisenberg *et al.* as shown in **Figure 5.1**, with hydrogen bonding occurring between the strands-in-register, i.e. the N-terminal region residues form hydrogen bonds with N-terminal region only, the same with C-terminal regions. B. The Shea dimer model identified from MD simulations in Ref[213] with a N-terminal region to C-terminal region hydrogen bonding interface. C. The α -helical dimer model crystallized by Eisenberg and coworkers (PDB code: 3G7V[193]), observed with a N-terminal α -helices dimer interface. D. hypothesized N-terminal α -helical with C-terminal extended dimer model, it is a schematic model referring to **Figure 5.2C** which has not been observed in solved structures nor from previous simulations.

5.3 Methods

5.3.1 Hypothesis validation plan and system setup

IAPP monomer and its variants

We hold the overall hypothesis that α -helical structures are critical intermediates for IAPP fibril initialization. The first priority, before we are able to validate the roles of α -helical structures in fibril forming mechanism, is to sufficiently characterize the structural ensembles of hIAPP monomers in physiological solution at atomic details. With respect to the experimental kinetics data of IAPP and its variants, we further hypothesize that the microscopic occurrences of α -helical structures are relevant to their macroscopic fibril formation rates. The fibril-formation promoting variant (S20G) is likely to adopt the highest α -helical content in monomeric states, while the fibril-formation disrupting variants (rat IAPP, I26P) are less helical-prone; the fibril formation rates of other systems may also follow similar trend with the α -helical fraction abundances.

Human IAPP and its variants including rat IAPP, S20G, S19G, I26P, C25C7S, free C-

terminus hIAPP (C_{acid}), acetyl-truncated 8-37 and truncated 8-37 were studied in this work. The sequence information for all systems is listed in **Figure 5.5**. The initial extended structures were built in LEaP module of Amber version 12. In all the systems except rIAPP, His18 was in neutral state. All the peptides except C_{acid} were amidated at C-terminus. In all except C2SC7S and the two truncated systems, a disulfide bond was added between the sulfur atoms on Cys2 and Cys7 (:2@SG-:7@SG); to avoid energetic problems of forcefully adding disulfide bond at the beginning of extended structure, the two residues were left as CYX (residue name of Cys that is involved in disulfide bond in Amber) without adding a bond until 10 ns of simulations. Sulfur-sulfur distances for each frame of the 10 ns simulation were calculated and the shortest sulfur-sulfur distance conformation was then used to generate the new topology and to serve as the new initial coordinate with disulfide bond. NMR structure (2L86[192]) was used as the second initial structure for hIAPP with disulfide bond, amidated C-terminus and neutral His18 from the beginning of simulations. The point mutations for variants were done in the Swiss PDB Viewer software.

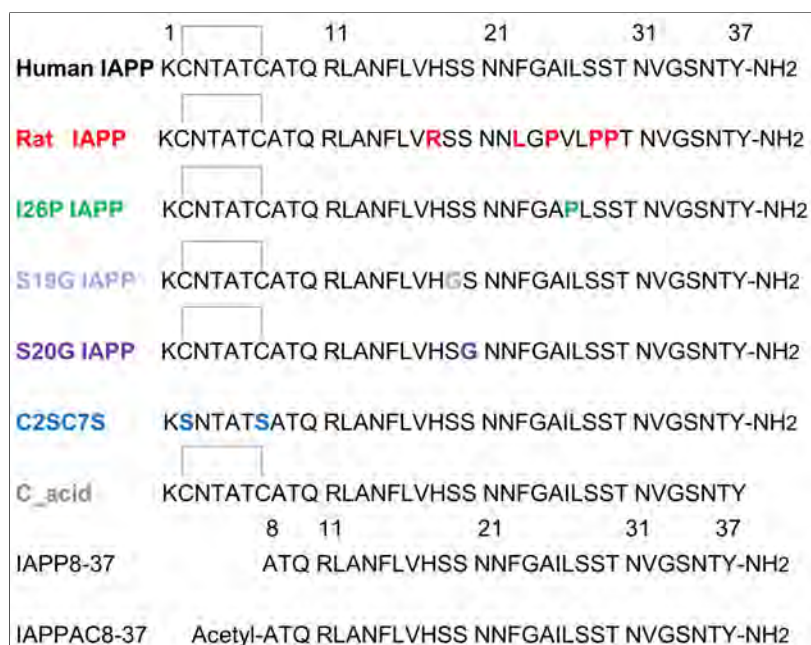


Figure 5.5: Sequences for WT, RAT, I26P, S20G, S19G, C2SC7S (experimentally mutated to Serine to mimic reduced Cysteine at C2 and C7), C_{acid} (C terminal with COO⁻ charged) and two versions of residue 1-7 truncated. All except C_{acid} are amidated at C-termini. Gray bridges indicate disulfide formation between the two connected Cysteine.

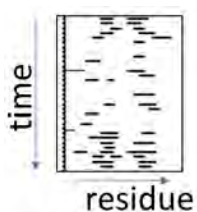


Figure 5.6: Schematic helix length visualization

DSSP analysis was used to get secondary structure types and regions throughout the simulations. Fractions of secondary structure of all systems were analyzed by ptraj implemented in Amber. Among all the secondary structures, α -helix, especially the transient nature of helices formed in IAPP WT was further quantified.

To restore an accurate picture of transient helical structures in our simulations, a visualization tool was developed as displayed in **Figure 5.6** where the formations of short helices could be quantified and visualized with designating short lines to the residues where α -helices are formed. Each row indicate one snapshot/time point of structure in the simulations, each residue is marked as a short line if this residue is assigned as α -helix in DSSP.

All the conformations sampled in our simulations provide useful structural ensembles for comparisons with experimental data, such as the calculated average radius of gyration (R_g) was compared to the values derived from NMR, Förster Resonance Energy Transfer (FRET), or small angle X-ray scattering (SAXS) data, the average distances of aromatic rings (Tyr, Phe, His and Trp) obtained from simulations was compared with fluorescence intensities measured in FRET. Some calculations were done but not shown in this chapter because no experimental data was obtained for comparison. For whoever gets interested in continuing this work, the data has been archived for your reference.

α -helical dimer models

The specific hypothesis in modeling the α -helical dimer was to test the validity of this α -helical dimer structure solved in crystallography (3G7V[193]) as a model for solution behavior. We also held a second hypothesis that if this dimer structure was a valid α -helical dimeric intermediate, the co-crystallized maltose binding proteins which chaperon the dimer formation would not be necessary for its thermodynamic stability.

As IAPP molecules in the crystal structure are C-terminal truncated fragment, the starting structure of full-length hIAPP α -helical dimer model was obtained by modeling two copies of the NMR structures (2L86[192]) onto the crystallized IAPP dimer segment(3G7V[193], **Figure 5.4C**). The superimposition of the two monomers were based on the heavy atoms of residue 8-17, which is the common helical region shared by these two experimental structures; in NMR structure, the α -helical region is 7-17, while the helical regions are 8-18 in the crystal dimer structure. It was assumed that the co-crystallized maltose binding proteins chaperon the dimer formation but are not necessary for its thermodynamic stability, thus the MBPs were deleted from the system.

To keep the dimer together, a distance restraint was put on each monomer to prevent the two monomers from flying away from one another. As IAPP monomers just transiently sample α -helical structures without detergent micelles or other helix promoting reagents, to thoroughly validate the role of N-terminal α -helices in this dimer model, we set up three different dimer systems:

(1) two helices restrained (**hlx**) in the dimer in which restraints were put on 7-17 residues to maintain the α -conformation of each monomer during the dimer simulations. All the backbone hydrogen bonds of residues 7-17 were restrained to the distances in the NMR

structure, with a force constant of 20 kcal/mol;

(2) two monomers unrestrained (**unr**) in the dimer in which no α -conformation restraints were imposed;

(3) hybrid restraints were applied (**hyb**) in the dimer where one 7-17 residue region of the dimer was applied with the helical restrained while the other monomer was left unrestrained.

Three measurements were employed to characterize the α -helical dimer systems in simulations. The first quantity is the distance between the N-termini of the two monomers within one dimer system, termed **N-N distance**, which measures the center of mass distance considering all the heavy atoms in the N-terminal residues 7 to 17. 10.0 Å is the N-N distance measured from the crystal structure (3G7V[193]) and also the starting N-N distance for all three systems. Another dihedral angle formed by four atoms on the dimer backbone was used to determine the relative angle formed by the two helix bundles. The four atoms are amide nitrogen in residue 14 on the first monomer (14@N1), carbonyl oxygen in residue 10 on the first monomer (10@O1), carbonyl oxygen in residue 10 on the second monomer (10@O2), and amide nitrogen in residue 14 on the second monomer (14@N2), which could be denoted as **dihedral N1-O1-O2-N2**. The corresponding dihedral angle adopted in crystal structure is 64°. Throughout the MD simulations, these two quantities were tracked for the three α -dimer systems. Furthermore, to validate the key contacts formed between the two Phe 15 residues in the α -dimer crystal structure, the **intermolecular interaction energies** throughout the MD simulations were calculated and decomposed to each residue by MM-GBSA (Molecular Mechanics-Generalized Born/Surface Area) module in the Amber package with GBNeck2 and surface tension of 5 cal/(mol·Å²). Numerical SASA (gbsa=2) was used to estimate the SASA for all the dimer molecules.

N-terminal α -helix with C-terminal extended dimer models

The other way of building dimer models followed the hypothesis that two in-register N-terminal α -helices with C-terminal extended monomers (as shown in **Figure 5.4D**) might be the precursor of mature fibril-like dimer formation. To validate this hypothesis, two groups of dimer structures, namely the **precursor group** and the **control group**, were generated to answer whether the in-register aligned dimers would be more likely to form parallel-sheet mature fibril-like structure (**Figure 5.7**).

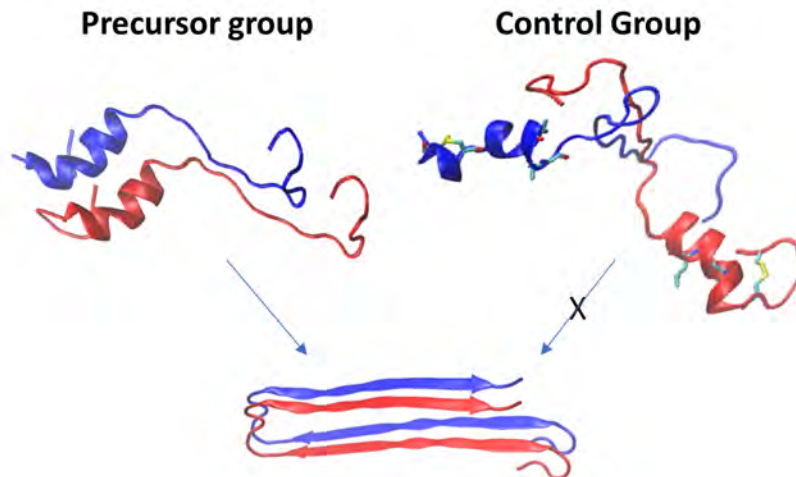


Figure 5.7: Hypothesized scheme in which two monomers that have N-terminal α -helix with C-terminal extended conformations (consistent with the schematic structure shown in **Figure 5.4D**) must be aligned in-register (precursor group) to lead to the mature fibril-like dimer (the same structure as shown in **Figure 5.4A**); while the dimer with α -helical regions spatially apart and extended regions misaligned (control group) is not the right precursor.

As illustrated above, for the hypothesis validation, we designed two groups of dimer molecules, whose monomeric states all adopt conformations that form an α -helix at the N-termini meanwhile the C-termini stay pretty extended. This structure is consistent with the schematic structure shown in **Figure 5.4D**). These dimer models were termed **α -helical N-terminus with extended C-terminus dimers**. In the precursor group, two monomers must be aligned in-register as to lead to the mature fibril-like dimer (the same structure as shown in **Figure 5.4A**); while for the dimers in the control group, even though each monomer maintains the same secondary structure features, if the α -helical regions were spatially apart and the extended regions were misaligned, the probability of going through a phase transition and forming fibril-like dimers would be significantly lower. Short MD simulations were carried out for dimers structures in both groups. Had the perfectly validated scenario taken place, for the dimers in the precursor group, thermodynamically stable dimer structures would associate through the N-terminal α -helices and form parallel sheet at the C-terminal regions first, and then gradually go through a phase transition to form fully parallel sheets. But for the control group, we would not see these associations or intermolecular hydrogen bonding formation, or phase transition in the same way. From a more realistic point of view, the findings as stated below would also partially convince us the validity of this dimer model, when the control group simulation results were compared to those in the precursor group, if (1) the N-termini started spatially apart would not directly interacting and thus weaker associations of helical regions would be observed (e.g. larger N-N distances); or (2) the C-termini started not adjacent would not form parallel-sheet and thus less parallel-sheet fraction would be observed.

As there was no relevant available experimental structure, we referred to the schematic structure of **Figure 5.4D** and filtered the hIAPP monomer trajectories for the structures that have N-terminal α -helix and C-terminal extended conformations. The criterion used

for the filtering are (1) helical length > 4 (1 turn of helix) in N-terminal residue 1-21 region, measured by DSSP and the analysis tool defined in **Figure 5.6**; (2) backbone RMSD < 2.0 Å of C-terminal residue 23-30 calculated against the fibril structure in **Figure 5.4A**. Among the 12000 frames in the WT hIAPP monomer trajectory, 63 structures that satisfied the above two criterion were further cut down to 7 structures (lost track of why it was these seven); these structures of diverse topologies were then used as initial monomer structures as shown in **Figure S5.1** for docking to build relevant dimer models.

To build thermodynamically stable and reasonable dimer models, the macromolecular docking program DOT 2.0 was used for the pairwise docking of selected monomer conformations. The docking protocol follows the tutorial written by Kevin Hauser http://simmerlinglab.org/wiki/index.php/Macromolecular_docking_with_DOT2.0. Unlike just docking two monomers, a pairwise docking process was carried out, where every 1 of the 7 monomer structures was docked to other 7 monomer structures, which generated 28 sets of dimer structures of various ranking scores (**Figure S5.3** illustrates the top ranked representative structures). All these 28 resulting highest scored dimer conformations were collected as the dimer structures in the control group. For the precursor group, as among the 28 there was only 1 dimer structure that has small N-N distance and in-register aligned C-termini, all the top 5 ranked conformations were manually checked and chosen by visualization, which resulted in the 12 dimer structures as shown in **Figure S5.2**. Dimer 11 and Dimer 12 were triplicated due to their large similarity with the hypothesized dimer in **Figure 5.4D**.

5.3.2 Simulation details

IAPP monomer and variants

All molecular dynamics (MD) simulations were fully unrestrained and carried out using pmemd.cuda of Amber. For hIAPP Wild Type and its variants, ff14SBonlysc and GBNeck2 were applied as force field and solvent model. For each system, two steps of minimization were applied first on all hydrogen atoms then on all side chain atoms, each step was followed by a heating step to raise the temperature gradually from 100 K to 300 K. In the equilibration, the force constants put on backbone atoms were reduced every 500 ps, respectively 2.0, 0.5, 0.1, 0 kcal/mol. SHAKE[39] was applied to constrain all bonds linking to hydrogen atoms. Each system was allowed to produce without restraints for 4 μ s. The temperature was set at 300 K which was the temperature of which fibril formation rates in **Figure 5.3** were measured.

We also carried out simulations for hIAPP WT using other force fields combined with explicit solvents using REMD starting from two conformations: (1) ff99SB+TIP3P, (2) ff14SBonlysc+TIP3P, and (3) ff14SB+TIP3P. For each of the two starting conformation, 8.0 Å of TIP3P[101] water molecules were added as a truncated octahedral periodic box. Equilibrations were done as described in the Supporting Information of **Chapter 3 on Page 72**. All the TIP3P REMD simulations were run in NVT ensemble; 8.0 Å was used as the non-bonded interaction cutoff; PME was used for long range electrostatics; Langevin dynamics with 1 ps^{-1} collision frequency was used; 48 replicas¹ were used and each replica was run

¹temperature ladders are 277.9, 279.7, 281.5, 283.3, 285.1, 286.9, 288.8, 290.6, 292.5, 294.3, 296.2, 298.1,

for 4 μ s. The 300.0 K trajectories were extracted and analyzed for the secondary structure composition. The error bars were calculated from two simulations.

Clustering was done on a combined trajectory of monomer WT, S20G and S19G, each system contributed 12000 frames of structures that were extracted from two runs of simulations. The hierarchical agglomerative average-linkage algorithm in ptraj program was used. The pairwise RMSD values were calculated based on the backbone atoms on residue 13 to 30, with an epsilon value of 3 Å, which produced 3601 clusters.

α -helical dimers

MD simulations for α -helical dimers were carried out using ff99SB and GBNeck2. For the starting structure of the dimer model, an equilibration was done in the same fashion as for the IAPP monomers. The corresponding α -helical restraints were added starting from the last equilibration step. During the simulations, a distance restraint was put on the two C α atoms of Cys2 on each monomer to keep the dimer together. This was imposed by a flat-well restraining function in which the force constant of 0 when the two atoms were < 60 Å while the force constant increased to 20.0 kcal/mol when the distance was beyond 60 Å. SHAKE[39] was applied to constrain all bonds linking to hydrogen atoms. The center of mass translation and rotation were removed every 500 MD steps (1 ps). Each of the three dimer model systems (hlx, hyb and unr) was allowed to run for 1.8 μ s.

α -helical N-terminus with extended C-terminus dimers

The dimer model structures resulted from pairwise docking were parameterized using ff14SBonlysc and GBNeck2. MD simulations for α -helical N-terminus with extended C-terminus dimers were carried out for at least 2 μ s after a similar equilibration process as described for the IAPP monomers. No restraints were applied to keep the dimers close in space nor to restrain the secondary structures.

5.4 Results and Discussions

5.4.1 Benchmark of four computation models in modeling hIAPP monomer

For the hIAPP WT MD simulations, RMSD values (data not shown) of peptide backbone were calculated with reference to the 1st frame of 2L86 NMR structure[192]. Compared to IAPP monomer simulations that have done before[209, 213, 218, 212], although the charge of monomers and water models vary from one to another, the intrinsically unfolded characteristics of IAPP monomer remains, as no obvious local or global minimum is observed in any of the trajectories.

Although no folded structures have been observed throughout the simulations, transient secondary structures are adopted. As no consensus for IDPs (including hIAPP) on which

300.0, 301.9, 303.8, 305.8, 307.7, 309.7, 311.7, 313.7, 315.7, 317.7, 319.7, 321.8, 323.8, 325.9, 328.0, 330.1, 332.2, 334.3, 336.4, 338.6, 340.7, 342.9, 345.1, 347.3, 349.5, 351.8, 354.0, 356.3, 358.5, 360.8, 363.1, 365.5, 367.8, 370.1, 372.5, 374.9 K

physical-based model reflects the actual secondary structure preferences, we first examine across all the combinations we have simulated. The results are summarized in **Figure 5.8**. The α -helical and anti-parallel sheet fractions are sensitive to the changes of force field/solvent model combinations. For α -helical fraction, we observe a reasonable agreement between ff14SBonlysc+GBNeck2 with ff14SB+TIP3P, with minor inconsistency on the residue 9 to 17. Compared to the nearly diminished α -helical fractions predicted in ff99SB+TIP3P and ff14SBonlysc+TIP3P, the difference between the former group is nearly negligible. Due to the formation of disulfide bonds between residue 2 and 7, turn structures are observed in all simulations, with a 60% to 80% turn fraction from residue 4 to 6 predicted by all simulations. For anti-parallel sheet fraction, the error bars are much larger compared to other secondary structures due to the slow kinetics of β -structures formation and breakdown. However, across four different force field/solvent model combinations, common regions centering at residue 9-11, 15-18, 26-31 and 35-36 are shared, as the two terminal regions form a short anti-parallel sheet and the two regions in the middle form a longer anti-parallel sheet. 3_{10} -helical fractions are low and of similar percentages among the four models.

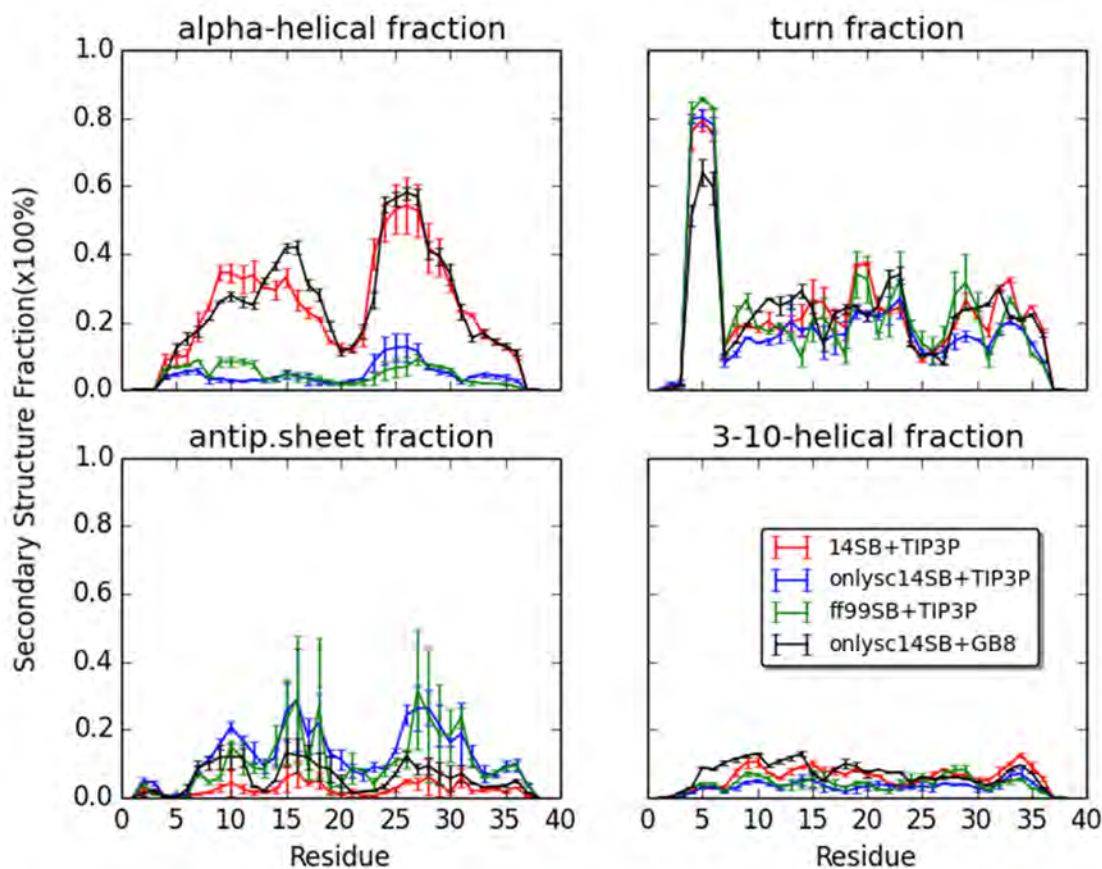


Figure 5.8: Four secondary structure fractions for hIAPP WT REMD simulations at 300K trajectories from four force field/solvent model combinations: ff14SB+TIP3P, ff14SBonlysc+TIP3P, ff99SB+TIP3P, ff14SBonlysc+GBNeck2.

Since protein folding to near-native structures has been shown accessible by employing implicit solvent (GB-Neck2) with a combination of ff14SBonlysc force field and GPUs[24],

additionally implicit solvent suffer less from convergence issue, also this model gives similar results to ff14SB+TIP3P of the secondary structure trend, it is reasonable to focus more on the trajectories generated from modified side chain parameters and implicit solvent, i.e. ff14SBonlysc and GBNeck2. Therefore, although helpful insight could also be drawn from the trajectories generated by other force field/solvent model combinations, we focus on the ff14SBonlysc and GBNeck2 combination for WT hIAPP and its variants.

5.4.2 Transient secondary structures in monomeric hIAPP

As α -helix structures are proposed to be relevant as the intermediate structures on the pathway of IAPP fibril formation, we further quantify the compositions of α -helix structures. Although up to 60% of α -helical fraction has been observed in **Figure 5.8A** for ff14SBonlysc with GBNeck2 simulations, whether the percentages indicate the appearances of α -helices simultaneously or not is not clear. In other words, it is possible that a few long α -helices are formed, or multiple short helices are sampled transiently throughout the whole simulations. As shown in **Figure 5.9**, the most often adopted α -helical length is 4, which is one helical turn. A second peak appears at the helix length of 7, which corresponds to two consecutive α -helical turns. With over 35% of one-turn helices but less than 13% of two-turn helices, the α -helical composition is partially revealed; not a few long helices but multiple transient and short helices are most often adopted by the hIAPP WT monomers in our simulations.

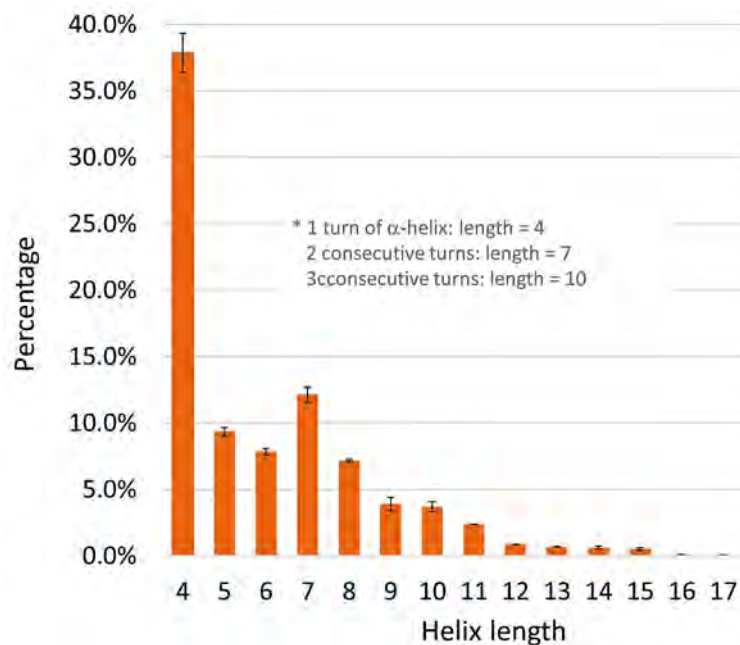


Figure 5.9: Percentages of helices lengths in hIAPP WT simulations with error bars calculated from the standard deviation between two trajectories. The regions of α -helix are assigned from DSSP and visualized using the tool developed as shown in **Figure 5.6**.

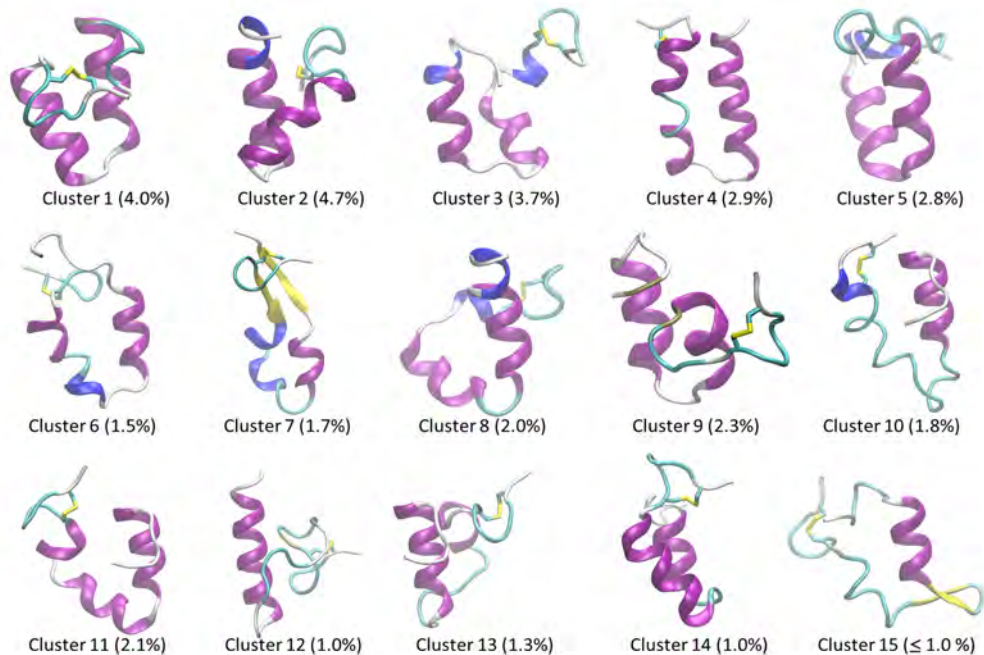


Figure 5.10: Clustering based on monomer WT, S20G and S19G combined trajectories, with reported percentages for WT. All the structures are the most representative structures within each cluster. Peptide structures are colored by secondary structure, purple indicates α -helix, blue stands for 3_{10} -helix, yellow for β -sheet, cyan for turn, and white for coil structure. The corresponding population of this structure is at bottom of each cluster.

Furthermore, when cluster representatives are displayed in **Figure 5.10**, a more clear cascade of α -helical structures are recovered from the trajectories. Among the top 15 clusters, 2 of them (cluster 7 and cluster 15) contain anti-parallel sheet structures, and the others are made of α -, 3_{10} -helix and turns.

With the largest cluster not exceeding a population size of 5%, it is ensured that disordered conformations are indeed reproduced in our simulations. The ability of our model to successfully predict the intrinsic disorder of IAPP is encouraging, as our models were trained against peptide energy minima and tested on native folded proteins, but the underlying physics shared between all proteins is the same and is possible to be reproduced by a model of good transferability, although over-compactness has been recognized as an issue in other studies[219, 220, 221].

5.4.3 Helical fractions and kinetic rates in IAPP variants

Despite $> 86\%$ sequence similarities among all full length IAPP variants and the mutations appear after residue 18, there are still helical fraction variances observed in the N-terminal 7-17 regions. As seen in **Figure 5.11A** where the point mutation in S19G and S20G are both around the kink region, S19G compared with S20G reduces α -helical fraction by 15%, which indicates that the N-terminal conformations are influenced by the point mutation. Although not statistically significant, the trend that is observed in decreasing

N-terminal α -helical fraction in the order of S20G, WT, and S19G, is actually consistent with the decrease of kinetic rate of fibril formation in the minutes to hours time scale.

But the α -helical fraction for WT, RAT, I26P and charged C-terminus (shown in **Figure 5.11B**) are nearly indistinguishable considering the residue 1 to 18 N-terminal regions. RAT sequence which starts to differ from residue 18 (H18R), is even shown to possess the highest α -helical fraction at F15 and L16 among all. Although A25, S28 and S29 substituted by three Proline lead to disruption of high helical structures compared to WT and charged C-termini peptides, Proline 29 initializes a new helical region which is sampled for 30% of fraction in the RAT simulations. The experimental findings for I26P are similar as RAT IAPP, which also abolish amyloid formation. Similar trend is also observed in the simulation; residues F23, G24 and A25 in I26P variant abolish α -helical formations entirely, with P26 initializing helical fraction $< 20\%$, with respective to much higher helical fraction in the same region in WT.

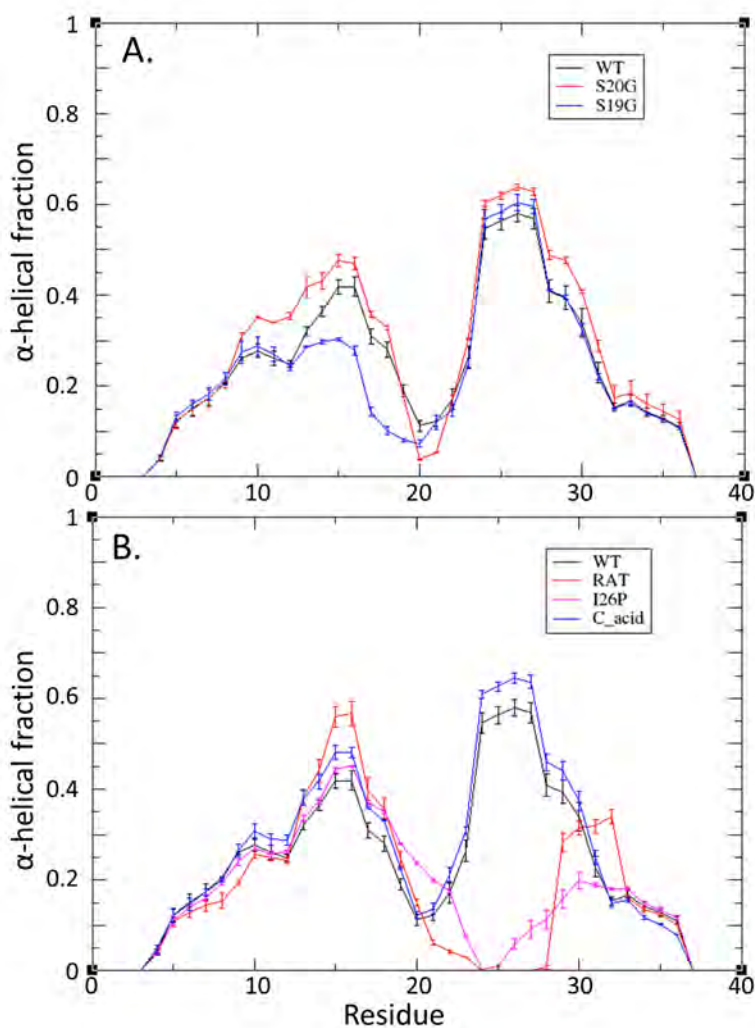


Figure 5.11: A. Fraction of α -helical content for WT, S20G and S19G. B. Fraction of α -helical content for WT, RAT, I26P and C_acid.

5.4.4 Dimer models proposed as the smallest oligomer

Simulated dimers deviate from the crystal conformation

Fused to maltose binding protein and solved in crystallography by Eisenberg and coworkers in 2009 (PDB code: 3G7V), this hIAPP dimer structure was proposed to be an important intermediate on the pathway of IAPP fibril formation[193]. In this crystal structure, a dimer interface is formed by two IAPP molecules packing against each other with key contacts being made near Phe 15, with 8–18 helices interacting at a 55° angle. To describe the spatial distance and relative angle in our defined ways (see Methods 5.3.1), a N-terminal helix-helix center of mass distance at 10.0 \AA and a N1-O1-O2-N2 dihedral of 64° are adopted in the crystal structure. This dimer structure at full-length was assumed to be thermodynamically stable intermediate thus MD simulations were carried out to valid its stability. Three systems were set up to test how stable the N-terminal helices would be when they were involved in a dimer interface. In the **hlx** system, both N-terminal regions (residue 7-17) were restrained as helices throughout the $1.8 \mu\text{s}$ of simulation, which ensures the two Phe 15 are locally exposed for making key contacts. Another set of **hyb** system was used to test if one monomer helix would promote the helical content on the other monomer. The **unr** system with both chains unrestrained was compared as a control.

In all three systems, the crystal dimer conformation is not sampled frequently enough to be considered as a thermodynamically stable structure, referring to the N-N distance and N1-O1-O2-N2 dihedral angle with respect to the numbers in crystal structure, shown in **Figure 5.12**. Although the N-N distance for both the hlx and hyb systems stay pretty close to 10 \AA , if the defined dihedral angle is examined, the hlx dimer never comes back to its starting angle at 64° , and the hyb dimer samples 64° shortly but does not stay or fluctuate around it. In the case of unr system, the two quantities designed to characterize the spatial relativity for helices are very noisy thus do not reflect the structural features in it.

Secondary structure fractions provide a better understanding of the structural differences in the three dimer systems. For the hyb system, it is observed that pre-formed helix promotes the α -helical stability; compared to the unr system, the monomer that is not restrained to be a helix has increased the helical content by nearly 20%, which is likely due to the intermolecular interactions with the other monomer with restrained helix. In the hlx system, interestingly, although the helical fraction of residue 8 to 16 stays close to 100%, the C-terminal helical content is not higher than that of the other two systems but slightly (10%) slower. For the unr system, the N-terminal helical fraction diminishes to 50% or less compared to 100% in the hlx system, more turn and anti-parallel sheet structures are observed (**Figure 5.13**).

The α -helical (monomer² and) dimer simulations were done using ff99SB and GBNeck2 (before the advent of ff14SBonlysc with better side chain parameters), which is different from the four models shown in **Figure 5.8**. With respect to the hIAPP monomer simulations using ff14SBonlysc+GBNeck2 and ff14SB+TIP3P, some common structural features are (1) a turn structure formed from residue 4 to 6 is always present due to the disulfide bond (Cys2-Cys7), and (2) the prevalent helical fractions are disrupted by Asn21 and Asn22. However, monomer and dimer simulations using ff99SB+GBNeck2 disagree with the other

²Data not shown but archived

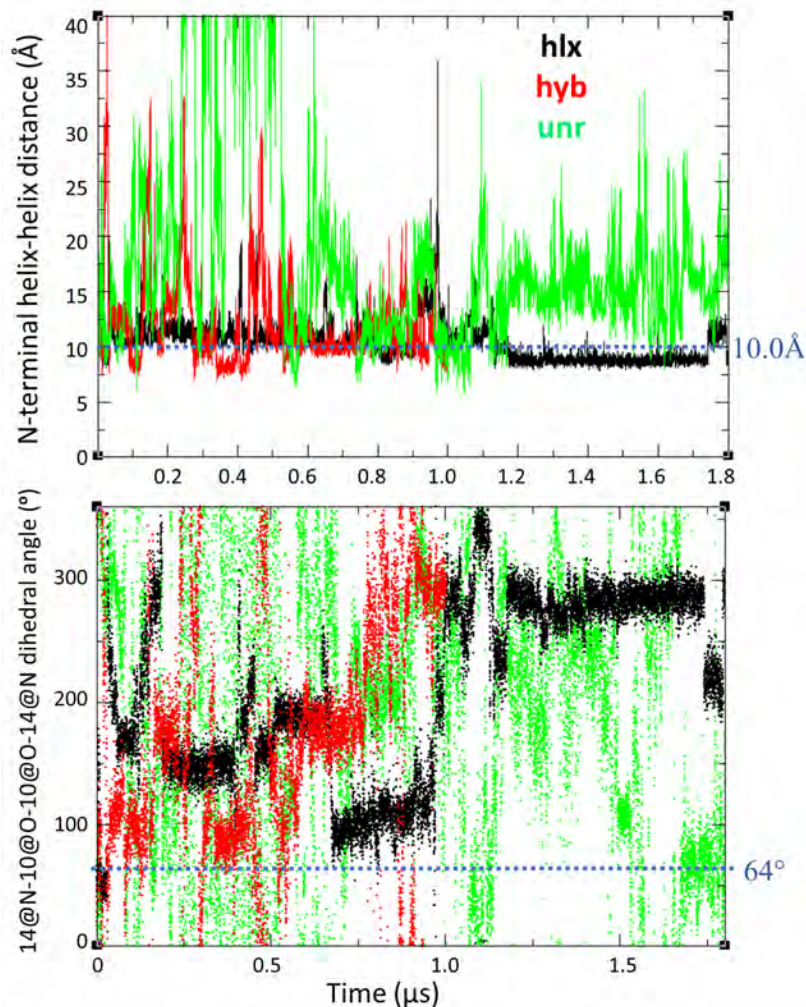


Figure 5.12: The spatial positions of N-terminal helices on the α -helical dimer. A. N-terminal helix-helix distance (termed N-N distance), measured by the center of mass distance of the heavy atoms in residue 7-17 on one monomer to the same set of atoms on the other monomer. The line in black is for the **hlx** system with both N-terminal helix restrained, the red line is for the **hyb** system with one system restrained and the green line is for the unrestrained **unr** system. B. The dihedral angle formed by the amide nitrogen on residue 14 and the carbonyl oxygen on residue 10 from both monomers (termed N1-O1-O2-N2 angle). The same color code was used in the dotted lines. The graphs were generated in Grace version 5.1.22.

two on whether the C-terminus or the N-terminus of IAPP is more helical-prone. Usually it is the N-terminus that is thought to be more helical-prone[195, 222] and the C-terminus is more amyloidogenic[223, 201].

Hydrophobic interactions in N-terminal helices

Although the dimer simulations do not validate the intermediate role of the crystal dimer structure, The N-N distances of around 8 to 12 \AA in the hlx and hyb systems still suggest

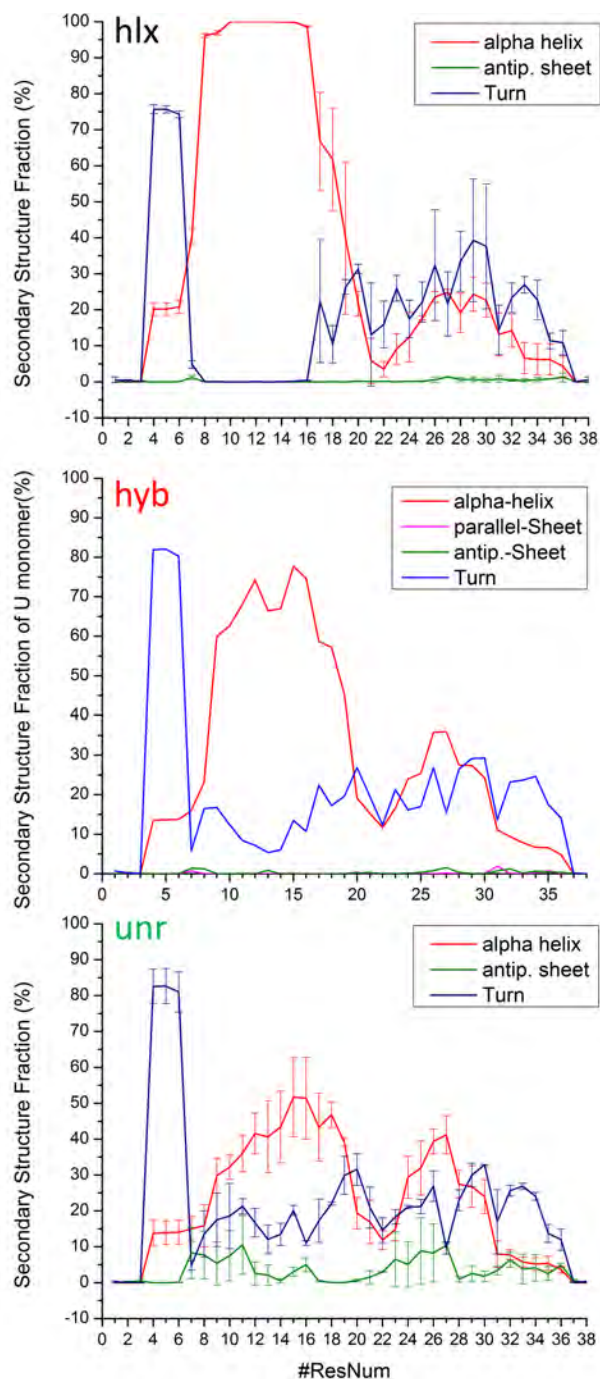


Figure 5.13: Fraction of secondary structure contents for the three systems of hIAPP in α -helical dimer simulations. The top panel is for **hlx** system where two N-terminal regions from residue 7-17 are restrained to helix, the error bars are calculated from the two monomers simulated as a dimer. The middle panel shows the secondary structure fractions for the unrestrained monomer in the **hyb** system. The bottom panel is for the **unr** system, the error bars are calculated from the two monomers.

that there are strong N-terminal interactions between two monomers within a dimer system.

To further elucidate how the monomers interact with each other in the simulations, we calculated the binding energy and decomposed to each residue. The three contact maps show different binding modes of the three systems (**Figure 5.14**). Notably, in all three systems, Phe15 from both monomers, which are suggested to have key contacts in crystal structure, are not interacting strongly during our dimer simulations. hlx and hyb systems share some common features, (1) N-termini of the two monomers make the strong favorable interactions; the residues Thr9, Leu12 and Leu16 seem to be the driving force of dimer formation in the simulations, (2) there are also favorable interactions between N-terminus of one monomer and C-terminus of the other monomer, (3) the C-terminus of each monomer do not seem to interact as strongly. One salient way the hyb system differs from hlx is that the C-terminal residues 20-37 of monomer without restraints apparently interact more frequently with the N-terminal residues 9-15 of the other monomer with N-terminal helix restraints. To be more specific, C-terminal residues Ile26, Leu27, Thr30, Asn31, Val32 have strong van de Waals and electrostatic interactions with Arg 11 located on N-terminus on the other monomer. The binding energy and residue contacts in the dimer simulations suggest that when N-terminal residues on one or both monomers are restrained to be α -helical, the two monomers show strong N-terminus-mediated interactions, while these contacts are not found in the unr system.

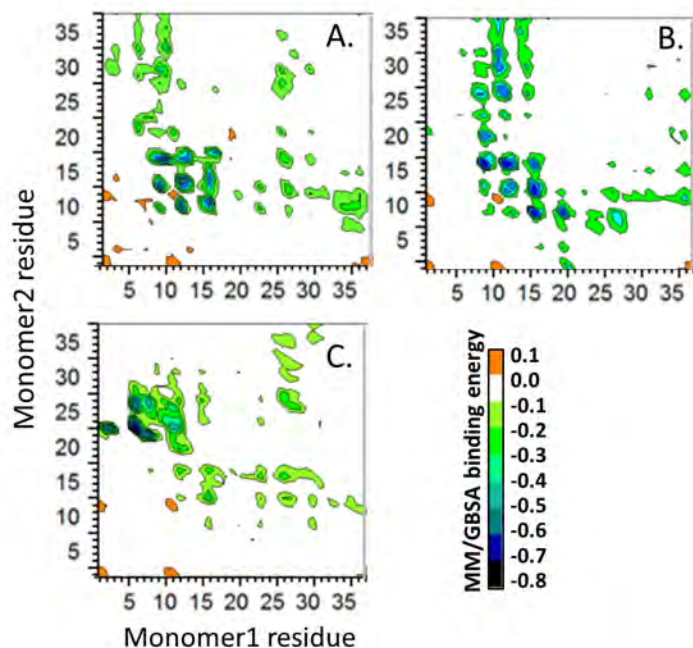


Figure 5.14: Energy decomposition map of three dimer systems. A. hlx system: two N-terminal helices restrained; B. hyb system: monomer 1 is N-terminal helix restrained and monomer 2 has no restraints; C. unr system: two unrestrained peptides. Graphs are generated in Origin 8.6.

Dimers in the precursor group have not lead to fibril-like dimers

The initial setup of the N-N-termini in space for the precursor group is more close and parallel than the control group (with average N-N terminal COM distance of $9.5 \pm 0.41\text{\AA}$ vs. $17.8 \pm 7.00\text{\AA}$ for the control group). Throughout the simulations, this feature has been kept in both groups: (1) in the precursor group, except in two runs the two monomers in a dimer have run apart and sometimes come back together, all the rest N-N distances are below 20\AA ; (2) the N-N distances in the control group are in average large, although it contains more simulation runs, and the distances fluctuate more(**Figure 5.15A** and **B**).

However, for secondary structures in the precursor and the control group, we do not see significant differences between these two groups. As seen in **Figure 5.15C** and **D**, one out of 16 precursor runs gain parallel sheet contents over time, but there is also one outstanding system in the control group too. As seen in **Figure 5.15E** and **F**, for α -helical fraction, the precursor group gains helical contents at the end of $2\ \mu\text{s}$ of simulations and the control group has noisier data and difficulty to directly compare against. In **Figure 5.15G** and **H**, for the anti-parallel sheet fractions, the precursor and the control group are getting more converged instead of diverged. 0-20% of anti-parallel sheet are maintained in the control group; even though the precursor group starts with nearly no content of anti-parallel sheet, at least three of runs develop $> 10\%$ of anti-parallel sheet fraction at the end, suggesting that the parallel sheet as observed in mature fibril structure is not favored in such a dimer model.

The ideal scenario of fibril-like dimer formation in the precursor group does not happen after $2\ \mu\text{s}$ of simulations, but revised expected outcomes which would differentiate the precursor group from the control group are partially observed. The N-N distances in the precursor group that are low throughout the simulations suggest some direct interactions and associations of the N-termini. But it also could be explained by insufficient sampling. The low parallel-sheet fractions observed in both groups show that parallel-sheet formations between the in-register aligned dimers are not more frequently observed, which suggests an invalid hypothesis, insufficient sampling or model inaccuracy.

Our models proposed at the dimeric level have not been validated but they might still be held when the concentrations of protein molecules reach a certain threshold as larger order of oligomers, given it is still an open question whether dimeric state of hIAPP is energetically stable as the smallest oligomer. Also the time scale of fibril formation rate as studied in the kinetic experiments is more than minutes, which is beyond the current computational achievable time scale of microsecond to millisecond. Therefore, the expectations not reached in the short time scale do not fully abolish the hypotheses nor validate them. Lastly, for the studies of intrinsically disordered proteins (IDPs, including IAPP), protocols to validate the all-atom computational models in reproducing the structural features of interest are in demand; that IDPs are currently simulated to be overly compact[219, 220, 221] suggests that their predicted secondary structure fractions may contain false positive indications, as to accurately describe a rugged energy landscape is still challenging for the current computational models.

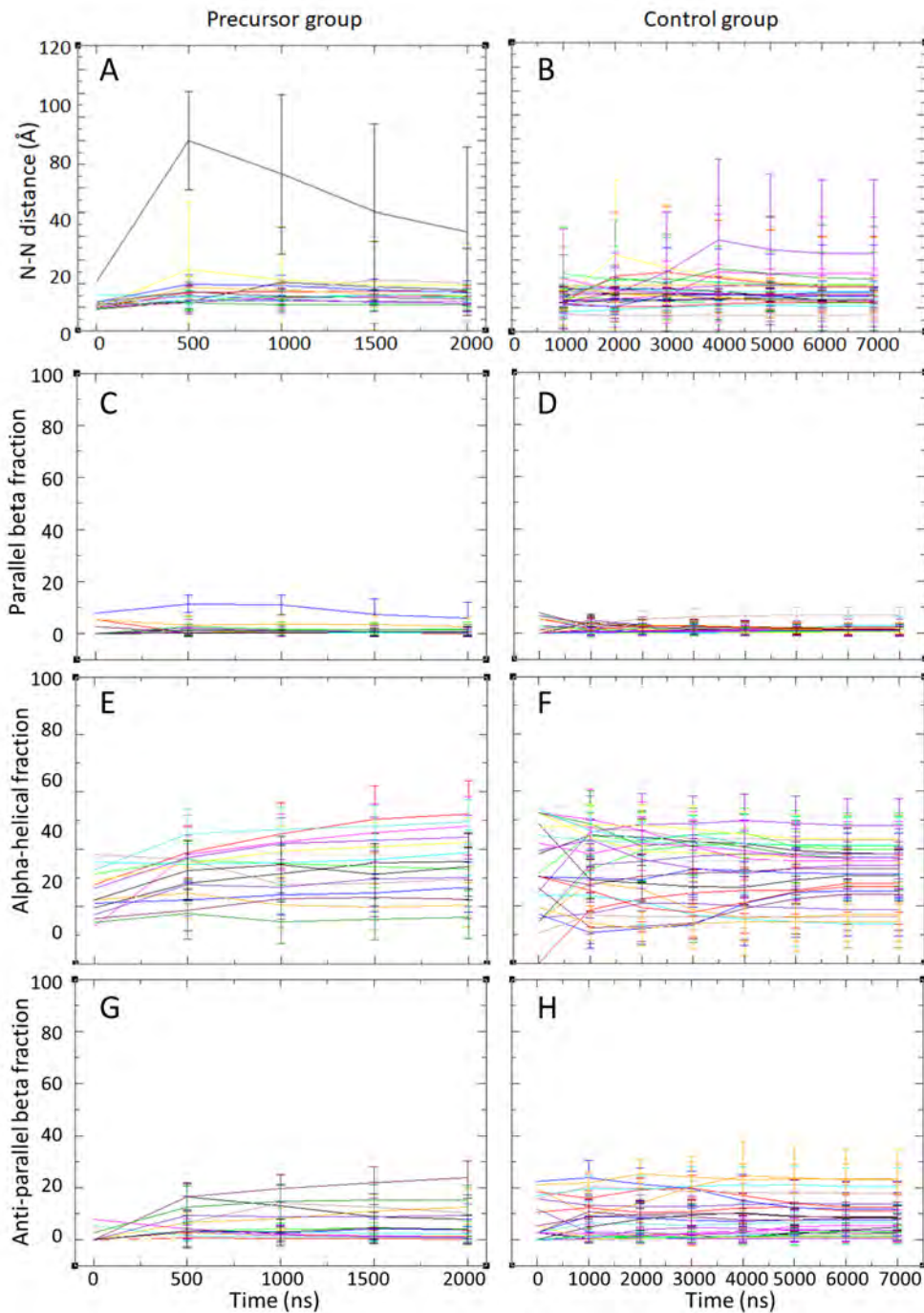


Figure 5.15: A and B. N-N distance over time for 16 runs of simulations in the precursor group and 28 runs of simulations in control group; C and D. Parallel β fraction over time for both groups; E and F. α -helical fraction over time for both groups; G and H. anti-parallel fraction over time for both groups. For each dimer, the error bars are the standard deviations of this quantity considering all the simulated structures before this time point.

One of the challenges, in my opinion, comes from the contradictory requirements arising from the structure prediction for natively folded proteins and the structural ensemble modeling of intrinsically disordered proteins. The observation that proteins are not stable enough in simulations compared with experimental thermal profiles motivated the incorporation of nonpolar term in solvation model (**Chapter 2**) and the evaluation of secondary structure specificity (**Chapter 3**). Our findings that the current computational model still destabilizes helical structures in CASP11 refinement trials are consistent with our previous understanding of the inaccuracy in the model (**Chapter 4**). The modifications proposed in **Chapter 2 and 3** are promising for alleviating the instability issue. However, a less compact and coil-rich structural ensemble is exactly what disordered protein requires, which seems to be against the necessity of stabilizing folded structures in simulations. For example, the non-polar term was added to particularly stabilize the more compact (smaller solvent accessible surface area) conformations, which does just the opposite effect to the goal of reproducing physical IDP ensembles. Therefore, the diverged requirements for the folded and disordered protein regimes are unlikely to be settled in the near future, if a universal force field and solvent model combination is needed to accurately describe both types of protein structures.

5.5 Conclusions

To validate the important role of α -helical intermediate structure in IAPP amyloid fibril formation, we applied MD simulations on the monomeric structures of IAPP and its variants, and built dimeric models of wild type IAPP. In rat IAPP and I26P where amyloid fibril formation are abolished, we also observed reduced α -helical fractions in the mutated regions. The simulated structural ensembles of WT, S20G and S19G are consistent with the experimental kinetic rates, although the time scales of the two processes have large gap. The two methods of building dimer models enhance our understanding in the α -helix association and propagation of helix-to-strand transition processes. Not observing the expected results indicate dimer might not the smallest energetically stable oligomer unit. Meanwhile, more experimental data is in need to validate the reliability of computational modeling results.

5.6 Supporting Information

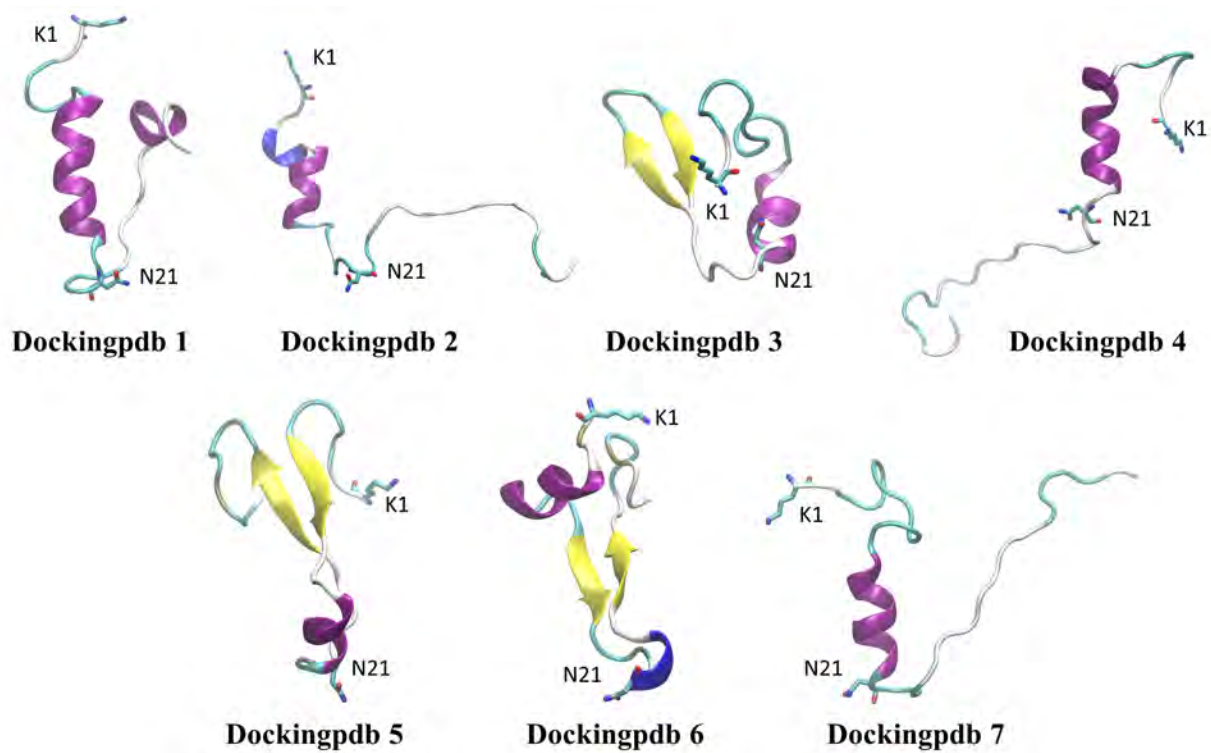


Figure S5.1: The structures of 7 monomers selected as representative α -helical N-terminus with extended C-terminus monomers ready for docking.

The precursor group

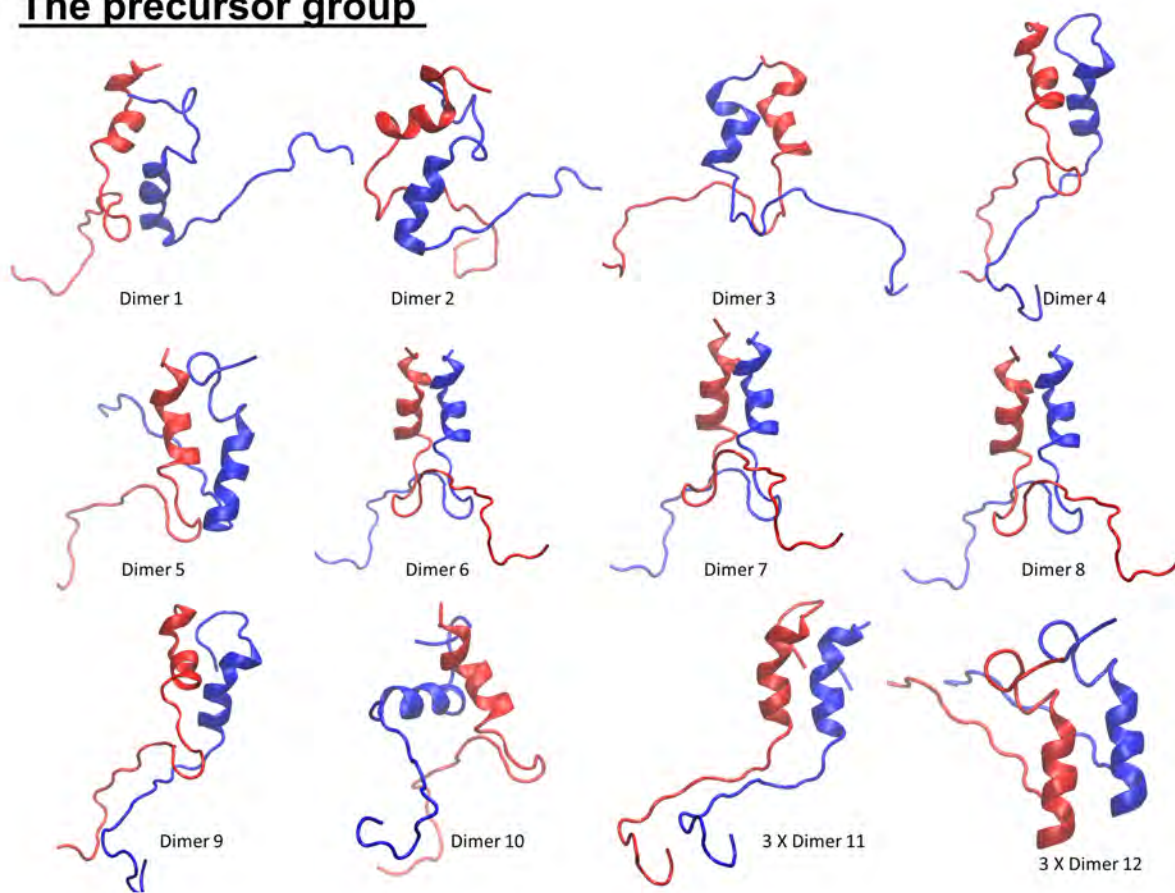


Figure S5.2: The precursor group: 16 dimer structures docked from representative α -helical N-terminus with extended C-terminus monomers. Dimer 11 and Dimer 12 were triplicated due to their large similarity with the hypothesized dimer in **Figure 5.4D**)

The control group

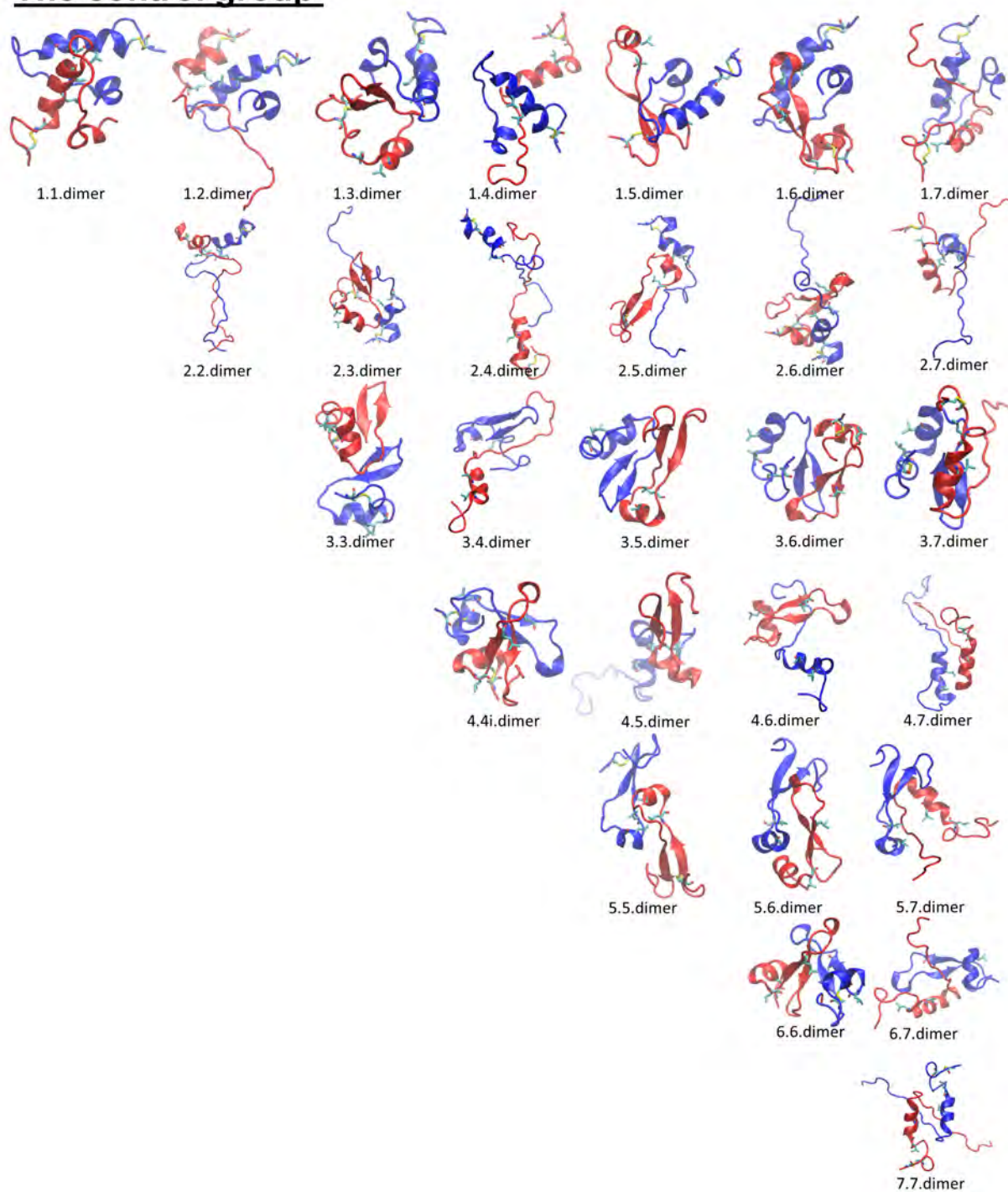


Figure S5.3: The control group: 28 dimer structures docked from representative α -helical N-terminus with extended C-terminus monomers.

Chapter 6

Conclusion and Prospective

Molecular modeling of proteins using Molecular Dynamics has remained an important branch of protein modeling studies, for the understanding of physical principles in protein folding into 3D structures dynamically and carrying out all sorts of functionality. The resulting more accurate structural predictions are beneficial for the disease-associated interference such as rational drug design. In this thesis, four questions have been investigated for more insight.

In Chapter 2, we were interested in how important the nonpolar term of solvation is to protein structure and stability, and developed a GPU-friendly SASA calculation algorithm to accelerate GB/SA solvation in MD simulations. We hypothesized that previously, nonpolar term has been underestimated as a small or even negligible term. Due to the large computational cost of SASA calculation and the incompatibility with GPU of current LCPO algorithm, the applications and optimization of nonpolar term solvation is limited by speed and accuracy. In this chapter, we demonstrated the critical role of nonpolar solvation and quantified its effectiveness in increasing the simulated stability of native structures. The model system HC16 set up a practical situation in which interactions of nonpolar residues and their contributions to the solvation free energy were analyzed under the spotlight; this was achieved by restraining the backbone helices and allowing three Phenylalanine residues to sufficiently sample different conformations relevant to hydrophobic core formation. With respect to the structural equilibria fully sampled in explicit solvent TIP3P, we attributed the discrepancy observed between GB and TIP3P solvent results to nonpolar solvation by reproducing the TIP3P solvent energy profile using GB/SA with a surface tension γ at $7 \text{ cal}/(\text{mol}\cdot\text{\AA}^2)$.

A big contribution of this work is a novel algorithm of SASA estimation, which is done in a pairwise fashion and implemented on GPU in Amber to accelerate GB/SA simulations by up to 30 times compared with the current algorithm LCPO. The cost of computing SASA is minimal as it makes use of the pairwise distances that have already been calculated for non-bonded interactions and GB solvation. It is a two-body algorithm in which only the neighbor atoms within a certain cutoff range of a central atom are iterated for once, therefore, there is nearly no computation overhead and is ideal for GPU parallelization. The accuracy of this two-body method is facilitated by a pre-treatment of 30 SASA atom types, pre-considering 1-2 bonded geometries of protein atoms. The 60 parameters were trained against a novel scrambled peptide data set to acquire the full spectrum of atom geometries

in protein environment. Another 30 parameters were numerically computed accordingly for nonpolar energy calculations. The validity of this algorithm is demonstrated in small protein GB/SA simulations with respect to LCPO algorithm.

The possible next step is to extend the algorithm for nucleic acids SASA calculations, so that DNA, RNA and protein-nucleic acid complexes could be simulated using GB/SA on GPUs. More long-timescale ($> \mu s$ as opposed to the current ns simulations achievable if GPU is not used) GB/SA simulations applied in protein folding, structure prediction etc. could be beneficial to the understanding of interplay of nonpolar solvation free energy with other energy terms, as well as the development of a next-generation nonpolar term which is more accurate than the model developed here.

In Chapter 3, we asked whether the current all-atom force field and implicit solvent can reproduce the amino acid backbone specificity, given all (except Glycine and Proline) the amino acids share the backbone parameters trained on Alanine peptides. To quantify the amino acid specificity of protein backbone in current computational models, we developed a toolbox comparing the simulated backbone dihedral angles with original preferences in 30 high quality crystal structures. A second toolbox previously used by Best *et al*[125] and Perez *et al*[71] was also employed for amino acid specific backbone helical propensity studies. The simulated values were compared against experimental measurements. Two tests agree on the low α stability for residues Serine, Aspartate acid, Tyrosine, Lysine and Arginine.

The discrepancy displayed in Alanine points out new directions for understanding whether training backbone parameters for Alanine is particularly problematic. More investigations are needed especially designed to uncover the sequence-dependence of amino acid backbone stability. For example, in the case of HP36 where 3 Alanine residues are thought to be important for the native conformation stability, more in-depth understanding could be gained from carefully controlled tests in which Alanine backbone parameter is carefully controlled as a single variable varying from ff99SB parameter set, ff14SB parameter set and to correction-map (CMAP) trying to reproduce QM backbone behaviors. More thoroughly designed and validated tools are always in demand, which should provide robust evaluation benchmark for next-generation computational model development and validations.

In Chapter 4, We evaluated how well current Amber force field and implicit solvent perform in the CASP refinement, if unrestrained MD simulations are applied. CASP11 refinement data set served as a good set of proteins of diverse secondary structure compositions. The stability of native structures in MD simulations were evaluated first. By comparing the RMSD deviation of simulations starting from native structures, helix bundle structures were identified to be the least stable in ff14SBonlysc and GBNeck2. It also brought in difficulties in refining the helical secondary structure rich template structures. Cluster analysis of all the refinement trajectories indicates a strong correlation of cluster size and refinement confidence; the larger is the most populated a cluster, the more likely this cluster belongs to a refined or close-to-refined structural ensemble, thus it is more confident that our computational model could get this template structure refined without prior knowledge of experimental structure.

As pointed out in some of the cases where the overall tertiary structures were refined while local secondary structure preferences were in bad shape, for example in the case of TR829, a reasonable future direction is to test the two hypothesis: the lack of nonpolar

term and/or instability of dihedral angles in α . After all, these observations motivated the development of the two projects elaborated in the previous two chapters (**Chapter 2** and **Chapter 3**).

In **Chapter 5**, we tackled another very challenging problem: protein aggregation. **We asked if MD simulations of low-order oligomers of amyloid-forming protein IAPP could shed light on its fibril initialization mechanism.** In wild type IAPP monomer simulations, similar secondary structure preferences were observed when two computational models were employed, which are ff14SBonlysc with GBNeck2 and ff14SB with TIP3P. The α -helical characteristics of monomeric IAPP and its variants were compared and further analyzed, in which two major conclusions could be drawn. Firstly, the predicted α -helical fractions are consistent with the macroscopic fibril formation rate, although that do not explain the mechanism of fibril initialization. Secondly, the majority of α -helices adopted in simulations are short helices, meaning they are transient and short-lived, which is consistent with the experimental findings about IAPP as it is intrinsically disordered. The dimer models were designed under the hypothesis that the smallest unit of IAPP initialization is two molecules and they aggregate through the N-terminal helical regions with propagation to form C-terminal parallel sheet. Although the hypotheses were not validated after preliminary tests, these findings will provide insight and experience for further studies.

In this thesis, more than 100 peptides or proteins were investigated *in silico* at the atomic level and a good number of them were solved from naturally found proteins which perform functional roles in different species. For example HP36 is a subdomain of chicken villin headpiece that functions together with muscle-related protein actin. However, it is mainly from specially designed "model systems" that we answered particular questions and learned the most of how to improve protein modeling accuracy. For example in **Chapter 2**, we derived HC16 model system as a representative structure of the hydrophobic core of HP36 and carefully controlled nonpolar term as a single variable. To get a full spectrum of all possible types of atomic geometries in proteins, we designed scrambled peptides each containing all 20 amino acids. Another example is in **Chapter 3**, where a particular amino acid substituting in $(AAXAA)_3$ peptides were especially designed and characterized in experiments for amino acid specificity studies. For the future studies of issue diagnosis in similar scenarios, more toy models should be designed for answering challenging questions when the interplay of different terms are complicated and the outcomes are hard to separate.

Bibliography

- [1] “Molecular machinery: A tour of the protein data bank.” <https://cdn.rcsb.org/pdb101/learn/resources/2014-mol-mach-poster.pdf>. Online; accessed: 2018-1-19.
- [2] S. Sagar, F. Sharp, and T. Curran, “Expression of c-fos protein in brain: metabolic mapping at the cellular level,” *Science*, vol. 240, no. 4857, p. 1328, 1988.
- [3] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, *et al.*, “Mass-spectrometry-based draft of the human proteome,” *Nature*, vol. 509, no. 7502, p. 582, 2014.
- [4] J. G. Sørensen, T. N. Kristensen, and V. Loeschcke, “The evolutionary and ecological role of heat shock proteins,” *Ecology Letters*, vol. 6, no. 11, pp. 1025–1037, 2003.
- [5] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” *Journal of Molecular Biology*, vol. 7, p. 95499, 1963.
- [6] Y. Kim, G. Chhor, M. Endres, G. Babnigg, and A. Joachimiak, “Crystal structure of KTSC domain protein YPO2434 from *Yersinia pestis*.” <http://www.rcsb.org/structure/4rgi>. To be published.
- [7] C. A. Orengo, A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton, “CATH—a hierarchic classification of protein domain structures,” *Structure*, vol. 5, no. 8, pp. 1093–1109, 1997.
- [8] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, “The Amber biomolecular simulation programs,” *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.
- [9] W. Humphrey, A. Dalke, and K. Schulten, “VMD – Visual Molecular Dynamics,” *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.
- [10] M. Karplus and G. A. Petsko, “Molecular dynamics simulations in biology,” *Nature*, vol. 347, no. 6294, p. 631, 1990.
- [11] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollmann, “Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation,” *Journal of the American Chemical Society*, vol. 115, no. 21, pp. 9620–9631, 1993.

- [12] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *Journal of the American Chemical Society*, vol. 110, no. 6, pp. 1657–1666, 1988.
- [13] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins," *Journal of the American Chemical Society*, vol. 106, no. 3, pp. 765–784, 1984.
- [14] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, "An all atom force field for simulations of proteins and nucleic acids," *Journal of Computational Chemistry*, vol. 7, no. 2, pp. 230–252, 1986.
- [15] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 3, pp. 712–725, 2006.
- [16] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *Journal of Computational Chemistry*, vol. 21, no. 12, pp. 1049–1074, 2000.
- [17] D. S. Cerutti, P. L. Freddolino, R. E. Duke Jr, and D. A. Case, "Simulations of a protein crystal with a high resolution X-ray structure: evaluation of force fields and water models," *The Journal of Physical Chemistry B*, vol. 114, no. 40, pp. 12811–12824, 2010.
- [18] O. F. Lange, D. Van der Spoel, and B. L. De Groot, "Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data," *Biophysical Journal*, vol. 99, no. 2, pp. 647–655, 2010.
- [19] L. Wickstrom, A. Okur, and C. Simmerling, "Evaluating the performance of the ff99SB force field based on NMR scalar coupling data," *Biophysical Journal*, vol. 97, no. 3, pp. 853–856, 2009.
- [20] R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides," *The Journal of Physical Chemistry B*, vol. 113, no. 26, pp. 9004–9015, 2009.
- [21] D.-W. Li and R. Bruschweiler, "Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins," *Journal of Chemical Theory and Computation*, vol. 7, no. 6, pp. 1773–1782, 2011.
- [22] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99SB protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010.

- [23] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, “ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB,” *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3696–3713, 2015.
- [24] H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling, “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*, vol. 136, no. 40, pp. 13959–13962, 2014.
- [25] C. J. Fennell and K. A. Dill, “Physical modeling of aqueous solvation,” *Journal of Statistical Physics*, vol. 145, no. 2, pp. 209–226, 2011.
- [26] R. Anandakrishnan, A. Drozdetski, R. C. Walker, and A. V. Onufriev, “Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations,” *Biophysical Journal*, vol. 108, no. 5, pp. 1153–1164, 2015.
- [27] H. Lei and Y. Duan, “Two-stage folding of HP-35 from ab initio simulations,” *Journal of Molecular Biology*, vol. 370, no. 1, pp. 196–206, 2007.
- [28] A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum, and K. A. Dill, “Blind protein structure prediction using accelerated free-energy simulations,” *Science Advances*, vol. 2, no. 11, p. e1601274, 2016.
- [29] G. Archontis and T. Simonson, “A residue-pairwise generalized Born scheme suitable for protein design calculations,” *The Journal of Physical Chemistry B*, vol. 109, no. 47, pp. 22667–22673, 2005.
- [30] G. Terashi and D. Kihara, “Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent,” *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 189–201, 2018.
- [31] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, no. 5612, p. 585, 1977.
- [32] P. K. Weiner and P. A. Kollman, “AMBER: Assisted model building with energy refinement. a general program for modeling molecules and their interactions,” *Journal of Computational Chemistry*, vol. 2, no. 3, pp. 287–303, 1981.
- [33] R. H. Swendsen and J.-S. Wang, “Replica Monte Carlo simulation of spin-glasses,” *Physical Review Letters*, vol. 57, no. 21, p. 2607, 1986.
- [34] K. Hukushima and K. Nemoto, “Exchange Monte Carlo method and application to spin glass simulations,” *Journal of the Physical Society of Japan*, vol. 65, no. 6, pp. 1604–1608, 1996.
- [35] M. Tesi, E. J. Van Rensburg, E. Orlandini, and S. Whittington, “Monte Carlo study of the interacting self-avoiding walk model in three dimensions,” *Journal of Statistical Physics*, vol. 82, no. 1-2, pp. 155–181, 1996.

- [36] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.
- [37] H. Nymeyer, "How efficient is replica exchange molecular dynamics? An analytic approach," *Journal of Chemical Theory and Computation*, vol. 4, no. 4, pp. 626–636, 2008.
- [38] A. Okur, D. R. Roe, G. Cui, V. Hornak, and C. Simmerling, "Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir," *Journal of Chemical Theory and Computation*, vol. 3, no. 2, pp. 557–568, 2007.
- [39] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327–341, 1977.
- [40] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," *Journal of Chemical Theory and Computation*, vol. 11, no. 4, pp. 1864–1874, 2015.
- [41] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.
- [42] D. R. Roe and T. E. Cheatham III, "PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data," *Journal of Chemical Theory and Computation*, vol. 9, no. 7, pp. 3084–3095, 2013.
- [43] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, "Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms," *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2312–2334, 2007.
- [44] D. Sitkoff, K. A. Sharp, and B. Honig, "Accurate calculation of hydration free energies using macroscopic solvent models," *The Journal of Physical Chemistry*, vol. 98, no. 7, pp. 1978–1988, 1994.
- [45] B. Honig and A. Nicholls, "Classical electrostatics in biology and chemistry," *Science*, vol. 268, no. 5214, pp. 1144–1149, 1995.
- [46] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, no. 16, pp. 6127–6129, 1990.
- [47] L. R. Pratt and D. Chandler, "Theory of the hydrophobic effect," *The Journal of Chemical Physics*, vol. 67, no. 8, pp. 3683–3704, 1977.
- [48] J. Chen and C. L. Brooks III, "Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions," *Physical Chemistry Chemical Physics*, vol. 10, no. 4, pp. 471–481, 2008.

- [49] E. Gallicchio, M. Kubo, and R. M. Levy, "Enthalpy- entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation," *The Journal of Physical Chemistry B*, vol. 104, no. 26, pp. 6271–6285, 2000.
- [50] J. A. Wagoner and N. A. Baker, "Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms," *Proceedings of the National Academy of Sciences*, vol. 103, no. 22, pp. 8331–8336, 2006.
- [51] M. S. Lee and M. A. Olson, "Comparison of volume and surface area nonpolar solvation free energy terms for implicit solvent simulations," *The Journal of Chemical Physics*, vol. 139, no. 4, p. 07B622.1, 2013.
- [52] J. Chen and C. L. Brooks, "Critical importance of length-scale dependence in implicit modeling of hydrophobic interactions," *Journal of the American Chemical Society*, vol. 129, no. 9, pp. 2444–2445, 2007.
- [53] Y. A. Arnautova, Y. N. Vorobjev, J. A. Vila, and H. A. Scheraga, "Identifying native-like protein structures with scoring functions based on all-atom ECEPP force fields, implicit solvent models and structure relaxation," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. 1, pp. 38–51, 2009.
- [54] J. Michel, M. L. Verdonk, and J. W. Essex, "Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization?," *Journal of Medicinal Chemistry*, vol. 49, no. 25, pp. 7427–7439, 2006.
- [55] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities," *Expert Opinion on Drug Discovery*, vol. 10, no. 5, pp. 449–461, 2015.
- [56] R. M. Levy, L. Y. Zhang, E. Gallicchio, and A. K. Felts, "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*, vol. 125, no. 31, pp. 9523–9530, 2003.
- [57] Q. Du, D. Beglov, and B. Roux, "Solvation free energy of polar and nonpolar molecules in water: an extended interaction site integral equation theory in three dimensions," *The Journal of Physical Chemistry B*, vol. 104, no. 4, pp. 796–805, 2000.
- [58] H. Liu, F. Chen, H. Sun, D. Li, and T. Hou, "Improving the efficiency of non-equilibrium sampling in the aqueous environment via implicit-solvent simulations," *Journal of Chemical Theory and Computation*, vol. 13, no. 4, pp. 1827–1836, 2017.
- [59] C. Tan, Y.-H. Tan, and R. Luo, "Implicit nonpolar solvent models," *The Journal of Physical Chemistry B*, vol. 111, no. 42, pp. 12263–12274, 2007.
- [60] E. Gallicchio, K. Paris, and R. M. Levy, "The AGBNP2 implicit solvation model," *Journal of Chemical Theory and Computation*, vol. 5, no. 9, pp. 2544–2564, 2009.

- [61] A. W. Gotz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with AMBER on gpus. 1. Generalized born," *Journal of Chemical Theory and Computation*, vol. 8, no. 5, pp. 1542–1555, 2012.
- [62] H. Nguyen, D. R. Roe, and C. Simmerling, "Improved generalized Born solvent model parameters for protein simulations," *Journal of Chemical Theory and Computation*, vol. 9, no. 4, pp. 2020–2034, 2013.
- [63] A. Onufriev, 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*, vol. 55, no. 2, pp. 383–394, 2004.
- [64] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, and A. Onufriev, "Generalized Born model with a simple, robust molecular volume correction," *Journal of Chemical Theory and Computation*, vol. 3, no. 1, pp. 156–169, 2007.
- [65] A. Onufriev, D. A. Case, and D. Bashford, "Effective Born radii in the generalized Born approximation: the importance of being perfect," *Journal of Computational Chemistry*, vol. 23, no. 14, pp. 1297–1304, 2002.
- [66] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, "Pairwise solute descreening of solute charges from a dielectric medium," *Chemical Physics Letters*, vol. 246, no. 1-2, pp. 122–129, 1995.
- [67] S. Le Grand, A. W. Götz, and R. C. Walker, "SPFP: Speed without compromise—a mixed precision model for GPU accelerated molecular dynamics simulations," *Computer Physics Communications*, vol. 184, no. 2, pp. 374–380, 2013.
- [68] D. Song, 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*, vol. 57, no. 5, pp. 1166–1178, 2017.
- [69] B. Barua, J. C. Lin, V. D. Williams, P. Kummier, J. W. Neidigh, and N. H. Andersen, "The Trp-cage: optimizing the stability of a globular miniprotein," *Protein Engineering, Design & Selection*, vol. 21, no. 3, pp. 171–185, 2008.
- [70] J. C. Lin, B. Barua, and N. H. Andersen, "The helical alanine controversy: An (Ala) 6 insertion dramatically increases helicity," *Journal of the American Chemical Society*, vol. 126, no. 42, pp. 13679–13684, 2004.
- [71] A. Perez, J. L. MacCallum, E. Brini, C. Simmerling, and K. A. Dill, "Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations," *Journal of Chemical Theory and Computation*, vol. 11, no. 10, pp. 4770–4779, 2015.
- [72] C.-Y. Zhou, F. Jiang, and Y.-D. Wu, "Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 1035–1047, 2014.

- [73] A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.
- [74] A. D. Mackerell, “Empirical force fields for biological macromolecules: overview and issues,” *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1584–1604, 2004.
- [75] F. Jiang, W. Han, and Y.-D. Wu, “The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development,” *Physical Chemistry Chemical Physics*, vol. 15, no. 10, pp. 3413–3428, 2013.
- [76] M. S. Shell, R. Ritterson, and K. A. Dill, “A test on peptide stability of AMBER force fields with implicit solvation,” *The Journal of Physical Chemistry B*, vol. 112, no. 22, pp. 6878–6886, 2008.
- [77] B. Lee and F. M. Richards, “The interpretation of protein structures: estimation of static accessibility,” *Journal of Molecular Biology*, vol. 55, no. 3, pp. 379–IN4, 1971.
- [78] E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler, and J. Meiler, “Solvent accessible surface area approximations for rapid and accurate protein structure prediction,” *Journal of Molecular Modeling*, vol. 15, no. 9, pp. 1093–1108, 2009.
- [79] S. J. Wodak and J. Janin, “Analytical approximation to the accessible surface area of proteins,” *Proceedings of the National Academy of Sciences*, vol. 77, no. 4, pp. 1736–1740, 1980.
- [80] W. Hasel, T. F. Hendrickson, and W. C. Still, “A rapid approximation to the solvent accessible surface areas of atoms,” *Tetrahedron Computer Methodology*, vol. 1, no. 2, pp. 103–116, 1988.
- [81] D. Dynerman, E. Butzlaff, and J. C. Mitchell, “Cusa and cude: Gpu-accelerated methods for estimating solvent accessible surface area and desolvation,” *Journal of Computational Biology*, vol. 16, no. 4, pp. 523–537, 2009.
- [82] J. Weiser, P. S. Shenkin, and W. C. Still, “Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO),” *Journal of Computational Chemistry*, vol. 20, no. 2, pp. 217–230, 1999.
- [83] V. Vasilyev and E. O. Purisima, “A fast pairwise evaluation of molecular surface area,” *Journal of Computational Chemistry*, vol. 23, no. 7, pp. 737–745, 2002.
- [84] T. J. Richmond, “Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect,” *Journal of Molecular Biology*, vol. 178, no. 1, pp. 63–89, 1984.
- [85] L. Wesson and D. Eisenberg, “Atomic solvation parameters applied to molecular dynamics of proteins in solution,” *Protein Science*, vol. 1, no. 2, pp. 227–235, 1992.
- [86] M. Schaefer, C. Bartels, and M. Karplus, “Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model,” *Journal of Molecular Biology*, vol. 284, no. 3, pp. 835–848, 1998.

- [87] D. Eisenberg and A. D. McLachlan, "Solvation energy in protein folding and binding," *Nature*, vol. 319, no. 6050, pp. 199–203, 1986.
- [88] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, "The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate Born radii," *The Journal of Physical Chemistry A*, vol. 101, no. 16, pp. 3005–3014, 1997.
- [89] T. Simonson and A. T. Bruenger, "Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment," *The Journal of Physical Chemistry*, vol. 98, no. 17, pp. 4683–4694, 1994.
- [90] D. Sitkoff, K. A. Sharp, and B. Honig, "Correlating solvation free energies and surface tensions of hydrocarbon solutes," *Biophysical Chemistry*, vol. 51, no. 2-3, pp. 397–409, 1994.
- [91] A. Shrake and J. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin," *Journal of Molecular Biology*, vol. 79, no. 2, pp. 351–371, 1973.
- [92] J. Weiser, A. A. Weiser, P. S. Shenkin, and W. C. Still, "Neighbor-list reduction: optimization for computation of molecular van der waals and solvent-accessible surface areas," *Journal of Computational Chemistry*, vol. 19, no. 9, p. 1110, 1998.
- [93] J. L. Morales and J. Nocedal, "Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization",," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, p. 7, 2011.
- [94] E. Jones, T. Oliphant, and P. Peterson, "{SciPy}: open source scientific tools for {Python}." <https://www.scipy.org/>, 2014.
- [95] W. Meng, B. Shan, Y. Tang, and D. P. Raleigh, "Native like structure in the unfolded state of the villin headpiece helical subdomain, an ultrafast folding protein," *Protein Science*, vol. 18, no. 8, pp. 1692–1701, 2009.
- [96] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, "NMR structure of the 35-residue villin headpiece subdomain.," *Nature Structural Biology*, vol. 4, no. 3, pp. 180–184, 1997.
- [97] S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata, "Crystal structure of a ten-amino acid protein," *Journal of the American Chemical Society*, vol. 130, no. 46, pp. 15327–15331, 2008.
- [98] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nature Structural and Molecular Biology*, vol. 9, no. 6, p. 425, 2002.
- [99] P. S. Shah, G. K. Hom, S. A. Ross, J. K. Lassila, K. A. Crowhurst, and S. L. Mayo, "Full-sequence computational design and solution structure of a thermostable protein variant," *Journal of Molecular Biology*, vol. 372, no. 1, pp. 1–6, 2007.

- [100] S. Xiao, V. Patsalo, B. Shan, Y. Bi, D. F. Green, and D. P. Raleigh, “Rational modification of protein stability by targeting surface sites leads to complicated results,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 28, pp. 11337–11342, 2013.
- [101] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.
- [102] D. J. Sindhikara, “Modular reweighting software for statistical mechanical analysis of biased equilibrium data,” *Computer Physics Communications*, vol. 182, no. 10, pp. 2227–2231, 2011.
- [103] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
- [104] D. E. Tanner, 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*, vol. 8, no. 7, pp. 2521–2530, 2012.
- [105] D. R. Roe, A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling, “Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation,” *The Journal of Physical Chemistry B*, vol. 111, no. 7, pp. 1846–1857, 2007.
- [106] N. Homeyer and H. Gohlke, “FEW: a workflow tool for free energy calculations of ligand binding,” *Journal of Computational Chemistry*, vol. 34, no. 11, pp. 965–973, 2013.
- [107] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, “How fast-folding proteins fold,” *Science*, vol. 334, no. 6055, pp. 517–520, 2011.
- [108] J. W. Pitera and W. Swope, “Understanding folding and design: Replica-exchange simulations of “Trp-cage” miniproteins,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 13, pp. 7587–7592, 2003.
- [109] R. Zhou, “Trp-cage: folding free energy landscape in explicit water,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 23, pp. 13280–13285, 2003.
- [110] C. J. Cramer and D. G. Truhlar, “Implicit solvation models: equilibria, structure, spectra, and dynamics,” *Chemical Reviews*, vol. 99, no. 8, pp. 2161–2200, 1999.
- [111] A. Ben-Naim and R. M. Mazo, “Size dependence of the solvation free energies of large solutes,” *The Journal of Physical Chemistry*, vol. 97, no. 41, pp. 10829–10834, 1993.
- [112] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig, “Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects,” *Science*, vol. 252, no. 5002, pp. 106–109, 1991.

- [113] D. M. Huang and D. Chandler, “Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 15, pp. 8324–8327, 2000.
- [114] H. Nguyen, A. Perez, S. Bermeo, and C. Simmerling, “Refinement of generalized Born implicit solvation parameters for nucleic acids and their complexes with proteins,” *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3714–3728, 2015.
- [115] A. Onufriev, D. Bashford, and D. A. Case, “Modification of the generalized Born model suitable for macromolecules,” *The Journal of Physical Chemistry B*, vol. 104, no. 15, pp. 3712–3720, 2000.
- [116] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, “Systematic validation of protein force fields against experimental data,” *PLoS One*, vol. 7, no. 2, p. e32131, 2012.
- [117] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. De Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, “Structure validation by $C\alpha$ geometry: ϕ , ψ and $C\beta$ deviation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 50, no. 3, pp. 437–450, 2003.
- [118] M. B. Swindells, M. W. MacArthur, and J. M. Thornton, “Intrinsic φ , ψ propensities of amino acids, derived from the coil regions of known structures,” *Nature Structural and Molecular Biology*, vol. 2, no. 7, p. 596, 1995.
- [119] S. Xun, F. Jiang, and Y.-D. Wu, “Significant refinement of protein structure models using a residue-specific force field,” *Journal of Chemical Theory and Computation*, vol. 11, no. 4, pp. 1949–1956, 2015.
- [120] A. Leaver-Fay, M. J. O’meara, M. Tyka, R. Jacak, Y. Song, E. H. Kellogg, J. Thompson, I. W. Davis, R. A. Pache, S. Lyskov, *et al.*, “Scientific benchmarks for guiding macromolecular energy function improvement,” in *Methods in enzymology*, vol. 523, pp. 109–143, Elsevier, 2013.
- [121] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe, “Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study,” *Journal of the American Chemical Society*, vol. 129, no. 5, pp. 1179–1189, 2007.
- [122] M. Karplus, “Contact electron-spin coupling of nuclear magnetic moments,” *The Journal of Chemical Physics*, vol. 30, no. 1, pp. 11–15, 1959.
- [123] D. S. Wishart and B. D. Sykes, “Chemical shifts as a tool for structure determination,” in *Methods in enzymology*, vol. 239, pp. 363–392, Elsevier, 1994.
- [124] R. J. Moreau, C. R. Schubert, K. A. Nasr, M. Török, J. S. Miller, R. J. Kennedy, and D. S. Kemp, “Context-independent, temperature-dependent helical propensities for amino acid residues,” *Journal of the American Chemical Society*, vol. 131, no. 36, pp. 13107–13116, 2009.

- [125] R. B. Best, D. de Sancho, and J. Mittal, “Residue-specific α -helix propensities from molecular simulation,” *Biophysical Journal*, vol. 102, no. 6, pp. 1462–1467, 2012.
- [126] Y. Gu, D.-W. Li, and R. Bruschweiler, “NMR order parameter determination from long molecular dynamics trajectories for objective comparison with experiment,” *Journal of Chemical Theory and Computation*, vol. 10, no. 6, pp. 2599–2607, 2014.
- [127] J. Grdadolnik, V. Mohacek-Grosev, R. L. Baldwin, and F. Avbelj, “Populations of the three major backbone conformations in 19 amino acid dipeptides,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 5, pp. 1794–1798, 2011.
- [128] Schrödinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*, November 2015.
- [129] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, “Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1,” *Journal of Molecular Biology*, vol. 285, no. 4, pp. 1735–1747, 1999.
- [130] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev, “H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations,” *Nucleic Acids Research*, vol. 40, no. W1, pp. W537–W541, 2012.
- [131] S. Butterworth, V. Lamzin, D. Wigley, J. Derrick, and K. Wilson, “Anisotropic refinement of a protein g domain at 1.1 Angstrom resolution.” <https://www.rcsb.org/structure/2IGD>. To be published.
- [132] T. Moulaei, I. Botos, N. E. Ziólkowska, H. R. Bokesch, L. R. Krumpke, T. C. McKee, B. R. O’keefe, Z. Dauter, and A. Wlodawer, “Atomic-resolution crystal structure of the antiviral lectin scytovirin,” *Protein Science*, vol. 16, no. 12, pp. 2756–2760, 2007.
- [133] A. Roos, K. Guo, N. Burgess-Brown, W. Yue, J. Elkins, A. Pike, P. Filippakopoulos, C. Arrowsmith, M. Wikstrom, F. Edwards, A. and von Delft, C. Bountra, D. Doyle, and U. Oppermann, “Crystal structure of the second pdz domain of the human numb-binding protein 2.” <http://www.rcsb.org/structure/2VWR>. To be published.
- [134] R. W. Alston, L. Urbanikova, J. Sevcik, M. Lasagna, G. D. Reinhart, J. M. Scholtz, and C. N. Pace, “Contribution of single tryptophan residues to the fluorescence and stability of ribonuclease Sa,” *Biophysical Journal*, vol. 87, no. 6, pp. 4036–4047, 2004.
- [135] J.-Y. Tung, M. D.-T. Chang, W.-I. Chou, Y.-Y. Liu, Y.-H. Yeh, F.-Y. Chang, S.-C. Lin, Z.-L. Qiu, and Y.-J. Sun, “Crystal structures of the starch-binding domain from *Rhizopus oryzae* glucoamylase reveal a polysaccharide-binding path,” *Biochemical Journal*, vol. 416, no. 1, pp. 27–36, 2008.
- [136] C. M. Starks, J. A. Francois, K. M. MacArthur, B. Z. Heard, and T. J. Kappock, “Atomic-resolution crystal structure of thioredoxin from the acidophilic bacterium *Acetobacter aceti*,” *Protein Science*, vol. 16, no. 1, pp. 92–98, 2007.

- [137] S. Szep, S. Park, E. T. Boder, G. D. Van Duyne, and J. G. Saven, “Structural coupling between FKBP12 and buried water,” *Proteins: Structure, Function, and Bioinformatics*, vol. 74, no. 3, pp. 603–611, 2009.
- [138] S. Fisher, J. Helliwell, S. Khurshid, L. Govada, C. Redwood, J. Squire, and N. Chayen, “An investigation into the protonation states of the C1 domain of cardiac myosin-binding protein C,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 64, no. 6, pp. 658–664, 2008.
- [139] L. Roos, A.K. and Tresaugues, C. Arrowsmith, H. Berglund, C. Bountra, and S. G. C. S. et al., “Crystal structure of the phox homology domain of human phosphoinositide-3-kinase-C2-gamma.” <http://www.rcsb.org/structure/2wwe>. To be published.
- [140] R. L. Stanfield, H. Dooley, P. Verdino, M. F. Flajnik, and I. A. Wilson, “Maturation of shark single-domain (IgNAR) antibodies: evidence for induced-fit binding,” *Journal of Molecular Biology*, vol. 367, no. 2, pp. 358–372, 2007.
- [141] M. S. Weiss, G. Mander, R. Hedderich, K. Diederichs, U. Ermler, and E. Warkentin, “Determination of a novel structure by a combination of long-wavelength sulfur phasing and radiation-damage-induced phasing,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, no. 4, pp. 686–695, 2004.
- [142] M. Maestre-Martínez, K. Haupt, F. Edlich, P. Neumann, C. Parthier, M. T. Stubbs, G. Fischer, and C. Lücke, “A charge-sensitive loop in the FKBP38 catalytic domain modulates Bcl-2 binding,” *Journal of Molecular Recognition*, vol. 24, no. 1, pp. 23–34, 2011.
- [143] K. V. Dunlop, R. T. Irvin, and B. Hazes, “Pros and cons of cryocrystallography: should we also collect a room-temperature data set?,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 61, no. 1, pp. 80–87, 2005.
- [144] H. Song, “High resolution structure of barley Bowman-Birk inhibitor.” <http://www.rcsb.org/structure/2fj8>. To be published.
- [145] Q. Liu, Q. Huang, M. Teng, C. M. Weeks, C. Jelsch, R. Zhang, and L. Niu, “The crystal structure of a novel, inactive, lysine 49 PLA2 from *Agkistrodon acutus* venom an ultrahigh resolution, ab initio structure determination,” *Journal of Biological Chemistry*, vol. 278, no. 42, pp. 41400–41408, 2003.
- [146] R. Berisio, F. Sica, V. Lamzin, K. Wilson, A. Zagari, and L. Mazzarella, “Atomic resolution structures of ribonuclease A at six pH values,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 58, no. 3, pp. 441–450, 2002.
- [147] A. Higashiura, T. Kurakane, M. Matsuda, M. Suzuki, K. Inaka, M. Sato, T. Kobayashi, T. Tanaka, H. Tanaka, K. Fujiwara, *et al.*, “High-resolution X-ray crystal structure of bovine H-protein at 0.88 Angstrom resolution,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 6, pp. 698–708, 2010.

- [148] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J.-P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson, *et al.*, “Histone recognition and large-scale structural analysis of the human bromodomain family,” *Cell*, vol. 149, no. 1, pp. 214–231, 2012.
- [149] T. Durek, V. Y. Torbeev, and S. B. Kent, “Convergent chemical synthesis and high-resolution x-ray structure of human lysozyme,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 12, pp. 4846–4851, 2007.
- [150] Y. Nishimiya, H. Kondo, M. Takamichi, H. Sugimoto, M. Suzuki, A. Miura, and S. Tsuda, “Crystal structure and mutational analysis of Ca^{2+} -independent type II antifreeze protein from longsnout poacher, *Brachyopsis rostratus*,” *Journal of Molecular Biology*, vol. 382, no. 3, pp. 734–746, 2008.
- [151] L. Liu, Z. Wei, Z. Chen, and Y. Wang, “Three crystal structures of human coactosin-like protein.” <http://www.rcsb.org/structure/1t3y>. To be published.
- [152] Y. Komarova, C. O. De Groot, I. Grigoriev, S. M. Gouveia, E. L. Munteanu, J. M. Schober, S. Honnappa, R. M. Buey, C. C. Hoogenraad, M. Dogterom, *et al.*, “Mammalian end binding proteins control persistent microtubule growth,” *The Journal of cell biology*, vol. 184, no. 5, pp. 691–706, 2009.
- [153] G. Prasad, D. Carney, B. Small, V. and Sankaran, and P. Zwart, “Crystal structure of human adipocyte fatty acid binding protein (FABP4) at 1.4 Angstrom resolution.” <http://www.rcsb.org/structure/3q6l>. To be published.
- [154] C. R. Garen, M. M. Cherney, E. M. Bergmann, and M. N. James, “The molecular structure of Rv1873, a conserved hypothetical protein from *Mycobacterium tuberculosis*, at 1.38 Angstrom resolution,” *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, vol. 62, no. 12, pp. 1201–1205, 2006.
- [155] M. S. Alpey, M. Gabrielsen, E. Micossi, G. A. Leonard, S. M. McSweeney, R. B. Ravelli, E. Tetaud, A. H. Fairlamb, C. S. Bond, and W. N. Hunter, “Tryparedoxins from *Crithidia fasciculata* and *Trypanosoma brucei*: photoreduction of the redox disulfide using synchrotron radiation and evidence for a conformational switch implicated in function,” *Journal of Biological Chemistry*, vol. 278, no. 28, pp. 25919–25925, 2003.
- [156] E. Jakobsson, B. Xu, S. Mahdi, J. Jansson, G. Kleywegt, and J. Stahlberg, “The crystal structure of the Beta-1,4-D-endoglucanase Cel45A from blue mussel *Mytilus edulis* at 1.2 Angstrom.” <http://www.rcsb.org/structure/1wc2>. To be published.
- [157] G. Kozlov, S. Bastos-Aristizabal, P. Määttänen, A. Rosenauer, F. Zheng, A. Killikelly, J.-F. Trempe, D. Y. Thomas, and K. Gehring, “Structural basis of cyclophilin B binding by the calnexin/calreticulin P-domain,” *Journal of Biological Chemistry*, vol. 285, no. 46, pp. 35551–35557, 2010.
- [158] S. Lifson and A. Roig, “On the theory of helix—coil transition in polypeptides,” *The Journal of Chemical Physics*, vol. 34, no. 6, pp. 1963–1974, 1961.

- [159] V. Munoz and L. Serrano, "Development of the multiple sequence approximation within the AGADIR model of α -helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms," *Biopolymers*, vol. 41, no. 5, pp. 495–509, 1997.
- [160] R. L. McFeeters, C. Xiong, B. R. O’Keefe, H. R. Bokesch, J. B. McMahon, D. M. Ratner, R. Castelli, P. H. Seeberger, and R. A. Byrd, "The novel fold of scytovirin reveals a new twist for antiviral entry inhibitors," *Journal of Molecular Biology*, vol. 369, no. 2, pp. 451–461, 2007.
- [161] F. Jiang, C.-Y. Zhou, and Y.-D. Wu, "Residue-specific force field based on the protein coil library. RSFF1: modification of OPLS-AA/L," *The Journal of Physical Chemistry B*, vol. 118, no. 25, pp. 6983–6998, 2014.
- [162] "UniProtKB: Homepage." <http://www.uniprot.org/uniprot/>. Online; accessed: 2018-03-06.
- [163] "RCSB Protein Data Bank: Homepage." <https://www.rcsb.org/>. Online; accessed: 2018-03-06.
- [164] A. Fiser, "Template-based protein structure modeling," in *Computational Biology*, pp. 73–94, Springer, 2010.
- [165] "Home: Prediction Center." <http://predictioncenter.org/index.cgi>. Online; accessed: 2018-03-06.
- [166] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 4–14, 2016.
- [167] A. Perez, J. A. Morrone, C. Simmerling, and K. A. Dill, "Advances in free-energy-based simulations of protein folding and ligand binding," *Current Opinion in Structural Biology*, vol. 36, pp. 25–31, 2016.
- [168] T. Nugent, D. Cozzetto, and D. T. Jones, "Evaluation of predictions in the CASP10 model refinement category," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. S2, pp. 98–111, 2014.
- [169] H. Fan, X. Periole, and A. E. Mark, "Mimicking the action of folding chaperones by Hamiltonian replica-exchange molecular dynamics simulations: Application in the refinement of de novo models," *Proteins: Structure, Function, and Bioinformatics*, vol. 80, no. 7, pp. 1744–1754, 2012.
- [170] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, p. 725, 2010.
- [171] W. Zhang, J. Yang, B. He, S. E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H.-B. Shen, and Y. Zhang, "Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11," *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 76–86, 2016.

- [172] J. Yang, W. Zhang, B. He, S. E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H.-B. Shen, and Y. Zhang, “Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade,” *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. S1, pp. 233–246, 2016.
- [173] J. P. Rodrigues, M. Levitt, and G. Chopra, “KoBaMIN: a knowledge-based minimization web server for protein structure refinement,” *Nucleic Acids Research*, vol. 40, no. W1, pp. W323–W328, 2012.
- [174] T. Nugent and D. T. Jones, “Membrane protein orientation and refinement using a knowledge-based statistical potential,” *BMC Bioinformatics*, vol. 14, no. 1, p. 276, 2013.
- [175] D. Bhattacharya, J. Nowotny, R. Cao, and J. Cheng, “3Drefine: an interactive web server for efficient protein structure refinement,” *Nucleic Acids Research*, vol. 44, no. W1, pp. W406–W409, 2016.
- [176] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, “Protein structure determination using metagenome sequence data,” *Science*, vol. 355, no. 6322, pp. 294–298, 2017.
- [177] J. Chen and C. L. Brooks, “Can molecular dynamics simulations provide high-resolution refinement of protein structure?,” *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 922–930, 2007.
- [178] A. Raval, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, “Refinement of protein structure homology models via long, all-atom molecular dynamics simulations,” *Proteins: Structure, Function, and Bioinformatics*, vol. 80, no. 8, pp. 2071–2079, 2012.
- [179] V. Mirjalili and M. Feig, “Protein structure refinement through structure selection and averaging from molecular dynamics ensembles,” *Journal of Chemical Theory and Computation*, vol. 9, no. 2, pp. 1294–1303, 2013.
- [180] V. Mirjalili, K. Noyes, and M. Feig, “Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging,” *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. S2, pp. 196–207, 2014.
- [181] J. Zhang, Y. Liang, and Y. Zhang, “Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling,” *Structure*, vol. 19, no. 12, pp. 1784–1795, 2011.
- [182] S. Lindert, J. Meiler, and J. A. McCammon, “Iterative molecular dynamics-Rosetta protein structure refinement protocol to improve model quality,” *Journal of Chemical Theory and Computation*, vol. 9, no. 8, pp. 3843–3847, 2013.
- [183] R. Y.-R. Wang, Y. Song, B. A. Barad, Y. Cheng, J. S. Fraser, and F. DiMaio, “Automated structure refinement of macromolecular assemblies from cryo-EM maps using rosetta,” *Elife*, vol. 5, 2016.

- [184] S. Leelananda and S. Lindert, “Iterative molecular dynamics- Rosetta membrane protein structure refinement guided by cryo-EM densities,” *Biophysical Journal*, vol. 114, no. 3, p. 575a, 2018.
- [185] Q. Cheng, I. Joung, and J. Lee, “A simple and efficient protein structure refinement method,” *Journal of Chemical Theory and Computation*, vol. 13, no. 10, pp. 5146–5162, 2017.
- [186] R. N. Rambaran and L. C. Serpell, “Amyloid fibrils,” *Prion*, vol. 2, no. 3, pp. 112–117, 2008.
- [187] D. Eisenberg and M. Jucker, “The amyloid state of proteins in human diseases,” *Cell*, vol. 148, no. 6, pp. 1188–1203, 2012.
- [188] P. Cao, P. Marek, H. Noor, V. Patsalo, L.-H. Tu, H. Wang, A. Abedini, and D. P. Raleigh, “Islet amyloid: from fundamental biophysics to mechanisms of cytotoxicity,” *FEBS Letters*, vol. 587, no. 8, pp. 1106–1118, 2013.
- [189] C. Esapa, J. H. Moffitt, A. Novials, C. M. McNamara, J. C. Levy, M. Laakso, R. Gomis, and A. Clark, “Islet amyloid polypeptide gene promoter polymorphisms are not associated with type 2 diabetes or with the severity of islet amyloidosis,” *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1740, no. 1, pp. 74–78, 2005.
- [190] W. T. Astbury, S. Dickinson, and K. Bailey, “The X-ray interpretation of denaturation and the structure of the seed globulins,” *Biochemical Journal*, vol. 29, no. 10, p. 2351, 1935.
- [191] L. Haataja, T. Gurlo, C. J. Huang, and P. C. Butler, “Islet amyloid in type 2 diabetes, and the toxic oligomer hypothesis,” *Endocrine Reviews*, vol. 29, no. 3, pp. 303–316, 2008.
- [192] R. P. R. Nanga, J. R. Brender, S. Vivekanandan, and A. Ramamoorthy, “Structure and membrane orientation of IAPP in its natively amidated form at physiological pH in a membrane environment,” *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1808, no. 10, pp. 2337–2342, 2011.
- [193] J. J. Wiltzius, S. A. Sievers, M. R. Sawaya, and D. Eisenberg, “Atomic structures of IAPP (amylin) fusions suggest a mechanism for fibrillation and the role of insulin in the process,” *Protein Science*, vol. 18, no. 7, pp. 1521–1530, 2009.
- [194] T. R. Jahn, M. J. Parker, S. W. Homans, and S. E. Radford, “Amyloid formation under physiological conditions proceeds via a native-like folding intermediate,” *Nature Structural and Molecular Biology*, vol. 13, no. 3, p. 195, 2006.
- [195] J. A. Williamson and A. D. Miranker, “Direct detection of transient α -helical states in islet amyloid polypeptide,” *Protein Science*, vol. 16, no. 1, pp. 110–117, 2007.

- [196] J. A. Williamson, J. P. Loria, and A. D. Miranker, “Helix stabilization precedes aqueous and bilayer-catalyzed fiber formation in islet amyloid polypeptide,” *Journal of Molecular Biology*, vol. 393, no. 2, pp. 383–396, 2009.
- [197] S. A. Jayasinghe and R. Langen, “Lipid membranes modulate the structure of islet amyloid polypeptide,” *Biochemistry*, vol. 44, no. 36, pp. 12113–12119, 2005.
- [198] J. D. Knight, J. A. Hebda, and A. D. Miranker, “Conserved and cooperative assembly of membrane-bound α -helical states of islet amyloid polypeptide,” *Biochemistry*, vol. 45, no. 31, pp. 9496–9508, 2006.
- [199] A. Abedini and D. P. Raleigh, “A critical assessment of the role of helical intermediates in amyloid formation by natively unfolded proteins and polypeptides,” *Protein Engineering, Design & Selection*, vol. 22, no. 8, pp. 453–459, 2009.
- [200] C. Betsholtz, L. Christmansson, U. Engström, F. Rorsman, V. Svensson, K. H. Johnson, and P. Westermark, “Sequence divergence in a specific region of islet amyloid polypeptide (IAPP) explains differences in islet amyloid formation between species,” *FEBS Letters*, vol. 251, no. 1-2, pp. 261–264, 1989.
- [201] P. Westermark, U. Engström, K. H. Johnson, G. T. Westermark, and C. Betsholtz, “Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation,” *Proceedings of the National Academy of Sciences*, vol. 87, no. 13, pp. 5036–5040, 1990.
- [202] P. Westermark, A. Andersson, and G. T. Westermark, “Islet amyloid polypeptide, islet amyloid, and diabetes mellitus,” *Physiological Reviews*, vol. 91, no. 3, pp. 795–826, 2011.
- [203] P. Cao, L.-H. Tu, A. Abedini, O. Levsh, R. Akter, V. Patsalo, A. M. Schmidt, and D. P. Raleigh, “Sensitivity of amyloid formation by human islet amyloid polypeptide to mutations at residue 20,” *Journal of Molecular Biology*, vol. 421, no. 2-3, pp. 282–295, 2012.
- [204] R. Akter, “Accelerating amyloid formation by islet amyloid polypeptide,” 2013. The Graduate School, Stony Brook University: Stony Brook, NY. Master Thesis.
- [205] C. Wu and J.-E. Shea, “Structural similarities and differences between amyloidogenic and non-amyloidogenic islet amyloid polypeptide (IAPP) sequences and implications for the dual physiological and pathological activities of these peptides,” *PLoS Computational Biology*, vol. 9, no. 8, p. e1003211, 2013.
- [206] T. M. Doran, E. A. Anderson, S. E. Latchney, L. A. Opanashuk, and B. L. Nilsson, “Turn nucleation perturbs amyloid β self-assembly and cytotoxicity,” *Journal of Molecular Biology*, vol. 421, no. 2-3, pp. 315–328, 2012.
- [207] R. Akter, P. Cao, H. Noor, Z. Ridgway, L.-H. Tu, H. Wang, A. G. Wong, X. Zhang, A. Abedini, A. M. Schmidt, *et al.*, “Islet amyloid polypeptide: structure, function, and pathophysiology,” *Journal of Diabetes Research*, vol. 2016, 2016.

- [208] P. Cao, F. Meng, A. Abedini, and D. P. Raleigh, “The ability of rodent islet amyloid polypeptide to inhibit amyloid formation by human islet amyloid polypeptide has important implications for the mechanism of amyloid formation and the design of inhibitors,” *Biochemistry*, vol. 49, no. 5, pp. 872–881, 2010.
- [209] N. F. Dupuis, C. Wu, J.-E. Shea, and M. T. Bowers, “Human islet amyloid polypeptide monomers form ordered β -hairpins: a possible direct amyloidogenic precursor,” *Journal of the American Chemical Society*, vol. 131, no. 51, pp. 18283–18292, 2009.
- [210] R. Laghaei, N. Mousseau, and G. Wei, “Effect of the disulfide bond on the monomeric structure of human amylin studied by combined Hamiltonian and temperature replica exchange molecular dynamics simulations,” *The Journal of Physical Chemistry B*, vol. 114, no. 20, pp. 7071–7077, 2010.
- [211] A. S. Reddy, L. Wang, S. Singh, Y. L. Ling, L. Buchanan, M. T. Zanni, J. L. Skinner, and J. J. De Pablo, “Stable and metastable states of human amylin in solution,” *Biophysical Journal*, vol. 99, no. 7, pp. 2208–2216, 2010.
- [212] W. Xu, H. Su, J. Z. Zhang, and Y. Mu, “Molecular dynamics simulation study on the molecular structures of the amylin fibril models,” *The Journal of Physical Chemistry B*, vol. 116, no. 48, pp. 13991–13999, 2012.
- [213] N. F. Dupuis, C. Wu, J.-E. Shea, and M. T. Bowers, “The amyloid formation mechanism in human IAPP: dimers have β -strand monomer-monomer interfaces,” *Journal of the American Chemical Society*, vol. 133, no. 19, pp. 7240–7243, 2011.
- [214] R. Laghaei, N. Mousseau, and G. Wei, “Structure and thermodynamics of amylin dimer studied by Hamiltonian-temperature replica exchange molecular dynamics simulations,” *The Journal of Physical Chemistry B*, vol. 115, no. 12, pp. 3146–3154, 2011.
- [215] J. Guo, Y. Zhang, L. Ning, P. Jiao, H. Liu, and X. Yao, “Stabilities and structures of islet amyloid polypeptide (IAPP22–28) oligomers: From dimer to 16-mer,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1840, no. 1, pp. 357–366, 2014.
- [216] M. Duan, J. Fan, and S. Huo, “Conformations of islet amyloid polypeptide monomers in a membrane environment: implications for fibril formation,” *PloS One*, vol. 7, no. 11, p. e47150, 2012.
- [217] Y. Zhang, Y. Luo, Y. Deng, Y. Mu, and G. Wei, “Lipid interaction and membrane perturbation of human islet amyloid polypeptide monomer and dimer by molecular dynamics simulations,” *PLoS One*, vol. 7, no. 5, p. e38191, 2012.
- [218] R. D. Murphy, J. Conlon, T. Mansoor, S. Luca, S. M. Vaiana, and N.-V. Buchete, “Conformational dynamics of human IAPP monomers,” *Biophysical Chemistry*, vol. 167, pp. 1–7, 2012.
- [219] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, “Water dispersion interactions strongly influence simulated structural properties of disordered protein states,” *The Journal of Physical Chemistry B*, vol. 119, no. 16, pp. 5113–5123, 2015.

- [220] F. Palazzesi, M. K. Prakash, M. Bonomi, and A. Barducci, “Accuracy of current all-atom force-fields in modeling protein disordered states,” *Journal of Chemical Theory and Computation*, vol. 11, no. 1, pp. 2–7, 2014.
- [221] J. Henriques, C. Cragnell, and M. Skepo, “Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment,” *Journal of Chemical Theory and Computation*, vol. 11, no. 7, pp. 3420–3431, 2015.
- [222] I. T. Yonemoto, G. J. Kroon, H. J. Dyson, W. E. Balch, and J. W. Kelly, “Amylin pro-peptide processing generates progressively more amyloidogenic peptides that initially sample the helical state,” *Biochemistry*, vol. 47, no. 37, pp. 9900–9910, 2008.
- [223] L. E. Buchanan, E. B. Dunkelberger, H. Q. Tran, P.-N. Cheng, C.-C. Chiu, P. Cao, D. P. Raleigh, J. J. De Pablo, J. S. Nowick, and M. T. Zanni, “Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient β -sheet,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 48, pp. 19285–19290, 2013.