Improving the accuracy of Amber force field for biomolecular simulation

A Dissertation Presented

by

Chuan Tian

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

December 2019

Stony Brook University

The Graduate School

Chuan Tian

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, Department of Chemistry

Isaac Carrico - Chairperson of Defense
Associate Professor, Department of Chemistry

Jin Wang
Professor, Department of Chemistry

Qin Wu
Staff Scientist
Brookhaven National Laboratory, Center for Functional Nanomaterials

This dissertation is accepted by the Graduate School

Eric Wertheimer
Dean of the Graduate School

Abstract of the Dissertation

Improving the accuracy of Amber force field for biomolecular simulation

by

Chuan Tian

Doctor of Philosophy

in

Chemistry

Stony Brook University

2019

Molecular dynamics (MD) simulations have become increasingly popular in studying the motions and functions of biomolecules. The accuracy of the simulation, however, is highly determined by the classical force field (FF), a set of functions with adjustable parameters to compute the potential energies from atomic positions. The relatively simple terms in most of the current force fields are computationally advantageous which enable the simulation of biologically important macromolecules at biologically relevant timescale. However, the overall quality of the FF, including our previously published ff99SB and ff14SB, is limited by the assumptions that were made years ago. (1) An overly symmetric φ/ψ dihedral energy map arises from the uncoupled cosine functions used to model these two degrees of freedom in the protein backbone. (2) The model does not show sufficient dependence of the backbone energetics on the amino acids, probably because the parameters developed for the simple amino acid Ala were applied to all other amino acids without checking the quality of the transferability. (3) The fixed partial charges were trained for aqueous solution, but the dihedral parameters were all fit to gas-phase quantum mechanics (QM), thus the resulting dihedral parameters actually counteract the intended polarization effect and introduce significant internal inconsistency in the model.

This dissertation seeks to overcome these limitations in FF. In ff19SB model, we have significantly improved the backbone profiles for all 20 amino acids. We fit coupled φ/ψ parameters

using 2D φ/ψ conformational scans for multiple amino acids, using as reference data the entire 2D QM energy surface. We address the polarization inconsistency during dihedral parameter fitting by using both QM and MM in solution. Finally, we examine possible dependency of the backbone fitting on side chain rotamer. To extensively validate ff19SB parameters, and to compare to results using other Amber models, we have performed a total of ~6 milliseconds MD simulations in explicit solvent with several different explicit water models. Our results show that after amino-acid specific training against QM data with solvent polarization, ff19SB not only reproduces the differences in amino acid specific Protein Data Bank (PDB) Ramachandran maps better, but also shows significantly improved capability to differentiate amino acid dependent properties such as helical propensities. We also conclude that an inherent underestimation of helicity is present in ff14SB, which is (inexactly) compensated by an increase in helical content driven by the TIP3P bias toward overly compact structures. In summary, ff19SB, when combined with a more accurate water model such as OPC, should have better predictive power for modeling sequence-specific behavior, protein mutations, and also rational protein design.

In addition, we have further investigated the potential source of errors in non-bonded parameters that can be improved in the long term. We quantified the magnitude of errors for the non-bonded terms including hydrogen bond, 1-4 scaling factor and partial charges. Our preliminary results show that there is still significant room for improvement on classical mechanical force field model. However, systematic refitting will be required to fundamentally improve force field.

Dedicated to my eternal wife and closest friend, Yu Zhu

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CD	Circular Dichroism
CMAP	Correction Map
CSD	Chemical Shift Deviation
ESP	Electrostatic Potential
ff99SB	force field 2006 Stony Brook
ff14SB	force field 2014 Stony Brook
ff19SB	force field 2019 Stony Brook
GB	Generalized Born
GB3	Third Immunoglobin-Binding Domain Of Protein G
HEWL	Lyozyme
HF	Hartree−Fock
IDP	Intrinsic Disordered Protein
ILDN	Isoleucine, Leucine, Aspartate, And Asparagine
iRED	Isotropic Reorientational Eigenmode Dynamics
MD	Molecular Dynamics
MM	Molecular Mechanics
MP2	Møller−Plesset Perturbation Theory of the Second Order
NMR	Nuclear Magnetic Resonance
Orig	Original Karplus Parameters
PDB	Protein Data Bank
ppII	Polyproline Helix, Type II
QM	Quantum Mechanics
REMD	Replica Exchange MD
RMSD	Root Mean Squared Deviation
Ubq	Ubiquitin
vdW	van der Waals

Acknowledgments

The five years since 2014 is the most valuable and unforgettable time in my life. I got acquaintance with many, many extraordinary and bright people and had fascinating experiences in this country. A good balance between work and life is the most thing I would appreciate for.

Before everything, I thank my advisor **Professor Carlos Simmerling**. I wouldn't be me without him! My parents gave birth to me but Carlos made me proud of myself. His knowledge, wisdom and all his advising throughout years have endowed me power to tackle obstacles in the rest of my career. I thank to his understanding and encouraging especially at the early stage of my PhD when I made little progress and was embarrassed to show little and poor data. I thank to him for being supportive all the time. I'm so grateful for having him as my advisor.

I thank my committee **Professor Isaac Carrico** and **Professor Jin Wang**. I thank them for their guidance, suggestions and helping me keep on track. I still remember how important the raised questions turned out to be from my 1st committee meeting in Nov, 2015.

I thank to my co-PI **Dr. Qin Wu** for his great contribution to the ff19SB project. His broad knowledge on quantum chemistry and organic chemistry has proved extremely helpful to my PhD work. I learnt a lot from him through our bi-weekly force field group meeting since 2016.

I thank to **Professor Daniel Raleigh** for his useful suggestions on experimental validation. One of the most important part of my PhD work is to validate model through comparing to experiments. His extensive knowledge really helps a lot.

I thank to **Professor Ken Dill**, the director of Laufer Center for his leading and bringing us the cutting-edge seminars and speakers, and the staff in Laufer Center **Nancy Rohring**, **Dr. Feng Zhang** for maintaining the research environment from different aspects.

I thank to **Kellon A.A Belfon** for sharing bright ideas with me. We had a lot of great conversations in the past five years and he has also contributed to the ff19SB work.

I thank to **Dr. Junjie Zou** for sharing insights on force field development. His knowledge and experience on free energy calculations are proved to be crucially important for force field modifications and improvements.

I thank to Dr. Angela N. Migues for advising in the force field development project and all her encouraging words and support.

I thank to all members in **Simmerling's lab**: Koushik Kasavajhala, Kenneth Lam, Zachary Fallon, Lauren Raguette, Yuzhang Wang for every week's discussions and suggestions in the lab meeting.

And finally, I thank my parents, **Xiujun Wang** and **Doctor Hongxiao Tian** for always supporting and encouraging me. They are the most kind and nice people in the world!

I am most grateful to my beloved wife, **Yu Zhu**. She is the only one in the world who knows deeply about me. Her pleasant characteristics and firm and persistent will have given me every calorie of energy to accomplish my PhD. With her company, I am confident to face difficulties and pursue excellence in the rest of my life.

Chapter 1

Introduction

"By 'life', we mean a thing that can nourish itself and grow and decay"

--- Aristotle

Understanding the underlying mechanisms that govern the biological process requires scrutiny at spatial and temporal resolutions that can be challenging for current experimental techniques. Over the past three decades, molecular dynamics (MD) simulation has evolved to become an indispensable tool for studying biological phenomena with atomic precision and at relevant timescales. Such simulations have served as a "computational microscope", providing information previously unattainable by experiments. Today, MD simulations are having a profound impact on biology. For example, researchers have utilized MD simulations to capture the assembly of a whole virus inside host cells and tried to explain how virus cause diseases[1]. Insight provided by these findings could be particularly informative for biological studies. The MD results, however are strongly dependent on the accuracy of the computed energies and forces. More broadly, whether it is proteins in bio-systems or conjugated polymers in organic electronic materials, understanding of the functions of macromolecules requires detailed information about their structures and conformations in condensed phase. Molecular modelling using classical force

fields (FF) is an indispensable tool in obtaining information for large biomolecules and long timescales. However, significant limitations remain, and making the next step in FF accuracy requires unique combination of skills, including extensive experience in developing force fields, their applications and limitations, along with expertise on quantum mechanics calculations that can be used to improve FFs.

The subject of this dissertation is to significantly improve the classical force field in Amber. In this chapter, first, however, the general concepts of proteins will be described. Then the quantum mechanics and classical mechanics that provide microscopic energetic information of molecules will be presented. The introduction will then be switched to force fields, recent developments of force fields and molecular dynamics. Finally, an outline of the dissertation will be included.

1.1 Proteins

Almost any property that characterizes a living organism can be influenced by proteins. Proteins store and transport various particles ranging from electrons to macromolecules. Some proteins control the electron flow in photosynthesis; some proteins control the passage of molecules across membranes; antibodies play important roles in immune system to defend against intruders; proteins also regulate gene expression via binding to nucleic acids; proteins are responsible for converting chemical energy into mechanical energy and control muscles; proteins are crucial for sight, hearing and other senses as well. Even though proteins have diverse biological function, their structures are relatively homogeneous. All proteins belong to the same type of linear polymers, made up from the same 20 building blocks, amino acids (**Figure 1.1**), but they differ in the polymeric sequence of amino acids. A single amino acid molecule may also be named a residue indicating a repeating unit of a polymer. Proteins can undergo condensation reactions, in which the amino acids lose one water molecule per reaction in order to attach to one another with a peptide bond. The chemical diversity of the amino acid and sequence is important to the functional diversity of proteins, but the highly flexible three-dimensional geometries are more determinant of the functions.

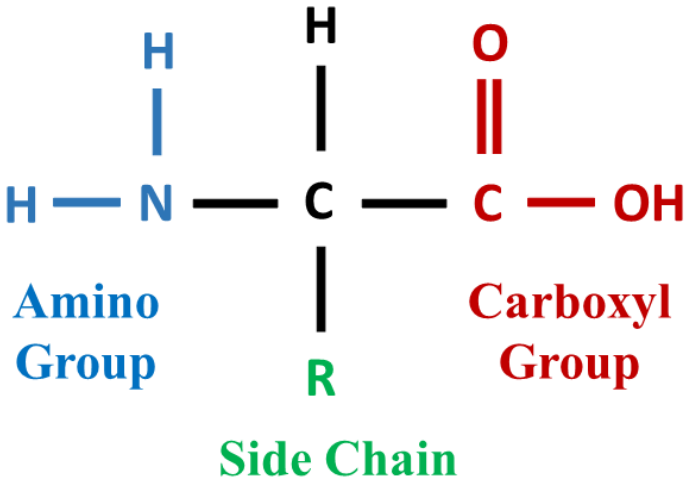


Figure 1.1 Chemical composition of single amino acid molecule with amino and carboxyl group.

The allowable geometric space of proteins is highly determined by the chemical composition of polypeptide. There are three repeating torsion angles along the peptide backbone chain called φ, ψ and ω. Because of the delocalization of carbonyl π electrons and the nitrogen lone pair, the peptide bond ω has partial double-bond character and tends to be planar (0° or 180°). Thus the carbonyl oxygen, carbonyl carbon, and amide nitrogen/hydrogen that make up the peptide bond are coplanar, and thus the free rotation around this ω bond is limited. The φ (C-N-Cα-C) and ψ (N-Cα-C-N) in the basic repeating of backbone are σ bonds and free rotation is permitted provided there is no steric clash from the side chains. Interactions between side chain rotation (determined by χ dihedral angle) and backbone atoms such as steric conflicts, hydrogen bond formation will limit the free rotation of φ and ψ angles. Thus a protein is a polymer with rotatable covalent bonds (φ and ψ) alternating with rigid planar ones (ω). This combination greatly restricts the number of possible conformations that a polypeptide chain can adopt. The energy barriers to the free rotation of dihedral angles in molecules are fundamental in structural properties and can be either measured by experimental techniques such as NMR spectroscopy or computed by theoretical methods such as quantum mechanics or molecular mechanics.

1.2 Quantum mechanics

In the early twentieth century, physicists found that the motions of small particles such as electrons and nuclei of atoms can be rigorously described by a set of rules/equations called quantum mechanics (QM). QM can be applied to a variety of problems in chemistry, for instance, to calculate thermodynamic properties such as heat capacity; to interpret molecular spectra and determine molecular properties such as molecular geometries, dipole moments, barrier to internal rotation, conformational energies, etc. Even though it is still difficult to apply QM calculations onto large biological molecules of interest, more and more researchers have begun to take advantage of QM to understand structures of biological molecules, enzyme-substrate binding, drug-protein binding and solvation of biomolecules.

The QM rigorously applies to microscopic "particles" such as electrons. To describe the state of a QM system for instance a hydrogen atom, the existence of a function Ψ is postulated as state function or wave function. The wave function contains all possible information about the system. Since the state will change with time, Ψ is a function of time. For one-particle, one-dimensional system, the future state of a system form its present state is defined as:

$$-\frac{\hbar}{i}\frac{d\Psi(x,t)}{dt} = -\frac{\hbar^2}{2m}\frac{d^2\Psi(x,t)}{dx^2} + V(x,t)\Psi(x,t) \qquad (1.1),$$

where ħ is Planck's constant divided by 2π, m is mass of particle, t is time, x is the coordinate of particle, V is the potential energy function of the system. **Equation 1.1** is also called time-dependent Schrödinger equation. The time-independent Schrödinger equation is often used to identify the energy of a system, following:

$$-\frac{\hbar^2}{2m}\frac{d^2\varphi(x)}{dx^2} + V(x)\varphi(x) = E\varphi(x) \qquad (1.2),$$

For the one-particle system like hydrogen atom, the exact wave function is known. For helium and lithium that contain interacting particles, very accurate wave functions can be calculated by approximation methods such as variation theorem. For atoms of higher atomic number, one way to find an accurate wave function is by first applying Hartree-Fork method, and approximate electron correlation with Møller-Plesset (MP) perturbation theory, coupled cluster theory, etc. For polyatomic molecules (for example, biomolecules), even more crude calculations become necessary including density-functional method, semi-empirical method, etc.

1.3 Molecular mechanics

Most of the biochemical problems to be tackled are unfortunately too large to be solved by quantum mechanics. As mentioned in Section 1.2, QM methods rigorously resolve electronic structures in a system, but the calculations become too expensive since a large number of particles need to be considered. Molecular mechanics (MM) is quite different from QM since it does not deal with an electronic Hamiltonian or wave function or an electron density. MM only applies to macroscopic particles, and only nuclear motions are considered. The energy of a system is merely computed as a set of functions of the nuclear positions. Therefore, MM is advantageously applied to perform calculations on systems containing significant number of atoms, for instance biopolymers. The cost of doing MM calculations scales as $O(N^2)$ where N is the number of atoms in the system. On the other hand, the simplest QM calculations formally scale as $O(N^3)$ or worse. The QM calculations in CCSD level scale as $O(N^7)$. The N is the number of particles including both nucleus and electron. More specifically, MM provides a reasonable trade-off between the wanted accuracy and the computer time. However, considering its missing description of physics, MM cannot be applied to questions that are highly related to molecule's electronic distribution and quantum effect, such as enzyme reaction, charge transfer, etc. The practical use of MM relies on several assumptions. One of the most important is, the motion of macro biomolecules can be dominantly determined by nuclear motions that can be described by relatively simple and computationally efficient functions. In this regard, MM work reasonably well when functions such as Hooke's law are used to describe certain contribution to the total energy.

1.4 Force fields

The MM method is sometimes called force field method because it relies almost entirely on classical mechanic force field (1.4) whose equation together with a set of empirical parameters can be used to calculate the energies and forces of molecules.

$$E_{total} = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} 4\varepsilon\left[\left(\frac{\sigma}{R_{ij}}\right)^{12} - \left(\frac{\sigma}{R_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 R_{ij}} \quad (1.3)$$

Force fields such as AMBER[2], CHARMM[3], OPLS[4] have very similar functional forms, while the differences mostly come from specific parameters. In a canonical force field equation (1.3), the vibrational mode for a typical bond is represented by a bond stretching term. That is based on assumptions that bond cannot break in simulations. Thus this model cannot be applied to chemical reaction study where covalent bonds can break and form. A true bond-stretching potential is better expressed as Morse curve[5]. But the Morse potential is not usually used in force field because of its inefficiency in computation. Simpler model like Hooke's law formula is often used. This functional form is reasonable because it well approximates the Morse curve at the bottom of the potential. The reference bond length r_{eq} is parameterized to fit values obtained by electron diffraction[6] or X-ray data[7]. The forces between valent bonds are very strong and significant energies are required to cause a bond to deviate largely from its equilibrium value. This is reflected by the force constant (K_r) in a bond stretching term. Many of the K_r come from normal mode calculations, in which the K_r values vary to give the best fit to experimental frequencies[7].

The deviation of bond angle from its reference value is also described in harmonic potential. The development of bond angle parameters followed a similar route. Values of θ_{eq} come from experimental data, while normal mode calculations also play a large role in the choice of K_θ values[7]. These two bonded terms are often regarded as 'hard' degrees of freedom, since quite substantial energies are required to break them. Higher order terms like cubic or quartic terms are also used to model the Morse curve more accurately or used to treat certain molecules.

Atoms can interact through space, usually described as non-bonded interactions. In force field, they are usually modeled as electrostatic and van der Waals (vdW) interactions. Electrostatic interactions result from electron distributions around atoms. In a simpler model however, the charges are restricted to nuclear centers, and are referred to partial charges. The polarization effect

is often neglected and the charge distribution is approximated with these discrete atom-centered partial charges. These fixed point charges can be derived from fitting to the QM electrostatic potential such as in RESP fitting[8], or fitting to reproduce thermodynamics properties of experiments using Monte Carlo methods. The electrostatic interactions are calculated between pairs of point charges using Coulomb's law (1.3) where R_{ij} is the distance between two atoms and ε_0 is the dielectric constant. The limitation of fixed-charge model is neglect of polarization. The polarization arises from changes in atomic charge distribution by an external field coming from local environment. Polarizable force fields such as AMOEBA[9] and DRUDE[10] explicitly consider the polarizability of molecules. Advances in polarization algorithms and computing hardware have significantly reduced the computational overhead of polarizable force fields with the accuracy and coverage improving in recent years[11]. One promising example is, the binding free energy between Mg^{2+} and $H2PO4^-$, determined by AMOEBA, CHARMM fixed-charge force field, and QM with a mixed explicit/continuum solvent model, was −2.23, −41.0, and −3.3 kcal/mol, respectively, compared with the experimental value of −1.7 kcal/mol[12]. But there is need to further calibrate the underlying physics of polarizable models to improve modeling of for instance charge penetration and transfer, and the timescale and length scale of polarizable simulations are still highly dependent on advances of efficient algorithm and hardware.

The vdW interaction is explicitly considered in a force field. It arises from a balance between attractive and repulsive forces. The attractive forces are long-range and dominant by London dispersive forces, arising from interactions between induced dipoles. In spite of its simplicity of only considering dipole-dipole interaction, this model gives quite reasonable results. The repulsive interactions arise from the Pauli principle, which prohibits any two electrons in a system from having the same set of quantum numbers. The best known vdW potential function (Lennard-Jones potential) contains an attractive part that varies as r^6 and a repulsive part that varies as r^{12}, the collision diameter σ is the separation for which energy is zero while the well depth ε determines the lowest energy. Since there is no strong theoretical reason of using the twelfth power term (just for the sake of computational efficiency), different powers such as values of 9, 10, 14 or even exponential function are also adopted in various models.

Due to the simplicity, Lennard-Jones potential and Coulomb's law are not accurate enough in modelling short-range interaction[13]. In practice, 1-4 (separated by three covalent bonds) scaling factor[14] is used to scale the short-range interaction by certain amount to reduce the errors. In

addition, 1-4 scaling factor is also meant to compensate for dihedral potential where outer atoms are separated by three bonds, and the insufficiency of dihedral potential in modeling QM orbital effects.

Most of variations in structure and relative energies come from the complex interplay between the torsional and non-bonded contributions. The energy barriers to the free rotation of dihedral angles is fundamental in structural properties. QM calculations suggested that these barriers resulted from the bonding/antibonding orbital effects, as rotation of the molecular orbitals results in phase changes of wave functions and change the rotational energies[15]. Since MM doesn't have explicit orbitals, this rotational profile is usually described in a dihedral term of force field using truncated Fourier series expansion (1.4). In the dihedral term, φ is any 4-atom torsion angle (both backbone and side chain), V_n relates to the 'barrier' height, n is the multiplicity which gives the number of minimum or maximum in one period, γ (initial phase) determines where the torsion angle passes through the minimum. For instance, the rotation of partial double bond ω that results from pure quantum orbital effects is calibrated by summation of one-fold and two-fold cosine functions. In practice, considering the simplicity of force field model, these dihedral parameters are not only simply accounting for the missing orbital effects in a classical model, but also making up for all differences between QM and MM model. These differences include the missing rotation-dependent polarization effects resulting from fixed-charge model, and the rotation-dependent errors in bond stretching, angle bending, and non-bonded interactions[2c]. This is also why dihedral parameter fitting is usually the last step in parametrization. As a result, the missing effects and errors from bonded and non-bonded terms are implicitly considered in the dihedral parameters. But since dihedral fitting was only done on a limited number of molecules, the parameters might not be transferable to dihedrals not explicitly included in training, or applied into dihedrals where the local environment nearby was changed.

Since FFs were developed when parameters were trained using smaller molecules than those typically simulated, the concept of "atom types" was developed[16], leading to many of the problems addressed below. These atom types typically reflect the element, hybridization and the nature of the functional group (e.g. hydroxyl vs. carbonyl oxygen). Atom types are used to assign vdW parameters, along with bonds, angles and dihedrals. While this assumption of broad transferability can work well, it becomes a serious weakness when atom types are used to select dihedral corrections. Dihedral terms are intended to be corrections on top of the energy profile for

the rest of the force field terms, in order to improve the match to training data. Because of the relatively few atom types, the same dihedral terms (defined using 4 consecutive atom types) occur frequently in a peptide sequence (for example, the amide in the protein backbone and in Gln/Asn side chains). In each case, the same dihedral parameters are often applied to segments that have different partial charges, since charges are assigned with much higher specificity than the atom types. A single dihedral term is unlikely to be an equally accurate correction in situations where the charge distribution is different, or when the neighboring functional groups vary. This overly broad application of atom types is a significant inconsistency and weakness in current models.

The force field terms are implicitly coupled to some extent. For example when the bond angle decreases, the bonds stretches in order to reduce the interaction between 1-3 atoms. Quantum calculations suggested that stretch-stretch, stretch-bend, bend-bend, stretch-torsion and bend-bend-torsion were important[17]. For instance, in MM2/MM3 force field[18], a cross-term is used to model stretching of two bonds adjoining an angle. Also in CHARMM force field[19], a 'Urey-Bradley' cross-term is there to compensate for angle bending by a harmonic function of the distance between the 1-3 atoms.

1.5 Recent development of force fields

The three major families of biomolecular force fields are AMBER, CHARMM and OPLS (briefly summarized in **Figure 1.2**). The history of AMBER force field for biomolecules can be traced back to over 30 years ago[4a, 7, 14]. The development of a molecular mechanics force field is timely, given the recent advances in the understanding of biomolecular interactions. Weiner et al.[7] have developed a force field that reasonably reproduce structures, energies, and vibrational frequencies of model systems. For reasons of computational efficiency, that force field used a united atom (spherical) representation of CH, CH_2 and CH_3 groups. Because of this approximation, compromises have to be made which lead in some cases to less than optimum fits with experiment. The following calculations have also suggested that when one is examining small energy differences, a spherical representation of CH groups leads to poorer agreement with experiment

than an all atom representation[14]. Simulations of NMR relaxation of methyl group rotations should benefit from an explicit treatment of all hydrogen atoms, and such a representation also makes comparisons to observed vibrational spectra much more straightforward.

The first generation all-atom force field associated with AMBER software[20] was developed by Weiner et al.[7, 14]. That was developed in the era before one could study big molecules in explicit solvent. As computer power grew, Cornell et al.[21] developed the second generation force field (denoted as ff94) for simulations in explicit solvent. In ff94, infrared spectroscopy and crystal structure geometry data were used for bond stretching and angle bending parameters. Different from OPLS charges which were derived empirically using liquid properties as a guide, QM calculations were used in ff94 to derive electrostatic potential (ESP). Specifically, ESP is calculated at a large number of grid points around the molecule at the QM level. At each grid point, the electrostatic interaction energy, between a molecular electronic and nuclear charge distribution and an external charge placed at grid point in the neighboring space around the molecule will be computed and is equivalent to electrostatic potential when the external charge is chosen to be positive unit charge[22].

Although the ab initio derived charges fluctuate significantly with small basis sets, after one reaches a basis set of 6-31G* quality, the ESP is close to convergent with respect to improvements in the basis set[21]. A 6-31G* based ESP-fit charge model is capable of giving an excellent reproduction of condensed-phase properties such as liquid enthalpies and densities[23]. But the 6-31G* standard ESP charges are less than ideal. First, the variation is considerable when charges are generated using different conformations of a molecule[8]. Second, the charges on buried atoms (such as sp3 carbons for butane) are underdetermined and often assume large values for nonpolar atoms. Thus, a restrained ESP-fit (RESP) was proposed[8]. In RESP model, hyperbolic restraints were added on charges of non-hydrogen atoms (such as buried carbons) to reduce their resulted charges without impacting the fit. New vdW parameters such as the ones for sp3 carbon were optimized to reproduce liquid properties and the remaining vdW parameters such as sp2 and sp3 N and sp2 O were retained from OPLS model[4a]. Monte Carlo simulations with adjustable σ and ε parameters were carried out to reproduce the densities and enthalpies of vaporization. Dihedral parameters such as backbone dihedral parameters φ (C-N-C_α-C) and ψ (N-C_α-C-N) were fit against accessible gas-phase minimum QM energies[24] of a set of glycine and alanine dipeptide conformations at that time, and side chain parameters were trained based on side chain analogues.

The overall hypothesis of FF training include (1) fitting RESP charges against QM, (2) fitting vdW parameters against liquid properties and (3) adjusting conformational energies with dihedral terms against small peptide QM data, and seem justified to reproduce experimental conformational energies for limited set of molecules better than models prior to 1990s.

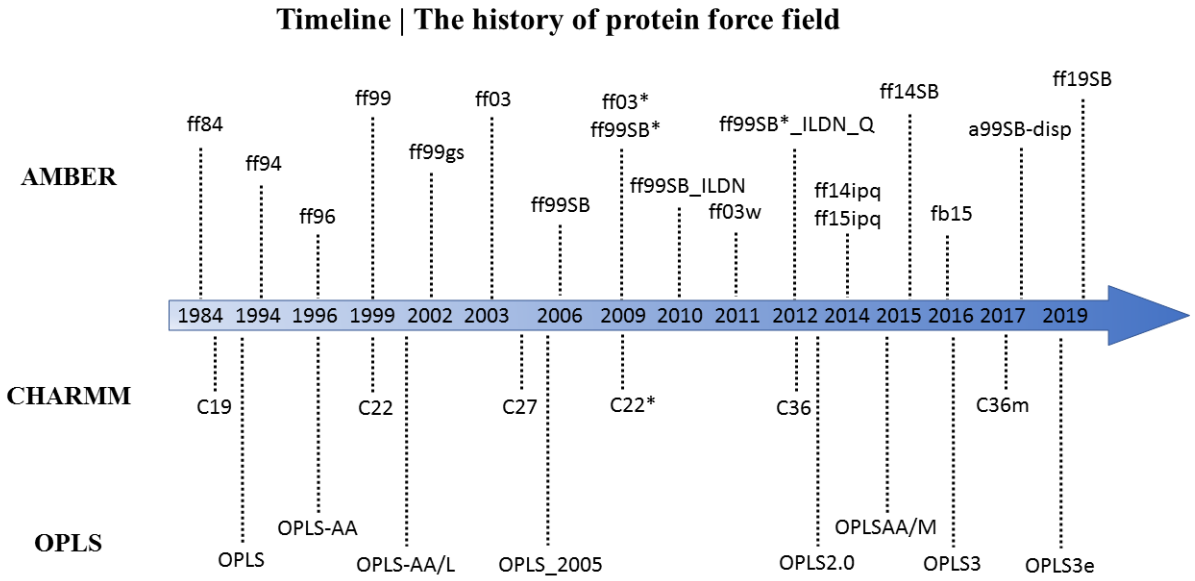


Figure 1.2 Thirty years of development of protein force fields. A selected number of protein force fields from AMBER, CHARMM and OPLS families are listed.

Because of limited computational resources at the time, dihedral parameters were fit to a small number of low energy conformations of glycine and alanine dipeptides. A possible limitation of using dipeptides is that their gas-phase energy surfaces do not have a local minimum in the α-helical region, which occurs with high frequency in protein structures[2b]. In ff96[25] and ff99[2d], φ and ψ backbone parameters were fit to better reproduce QM energies of alanine dipeptides and tetrapeptides. The use of tetrapeptides is advantageous over dipeptides because tetrapeptides can form an intermolecular hydrogen bond and have a local helical minimum in gas-phase. Because gas-phase QM is the reference energy in fitting, the use of gas-phase minima of tetrapeptide provides training data in helical region and in turn makes the model more accurate in secondary structure prediction. However, limitations in ff99 are that φ and ψ parameters were fit to reproduce relative energies for alanine with the existence of φ' (C-N-C_α-C_β) and ψ' (C_β-C_α-C-N) parameters retained from ff94 (**Figure 1.3**). And the new fitted φ and ψ parameters were applied to glycine

and proline afterward. Thus, it makes no physical meaning when these new φ and ψ parameters were applied to glycine where there were no φ' and ψ' dihedrals (missing C_β atom). This problem was carefully re-considered in ff99SB[2b]. While using glycine tetrapeptides in fitting φ and ψ parameters, a set of alanine tetrapeptides were used to fit φ' and ψ'. These φ' and ψ' parameters were then applied to the rest of amino acids containing C_β. The pair-wise relative energies were used in fitting instead of absolute energies with arbitrarily zeroing the energy of one conformation as in ff99[2d]. Both ff99 and ff99SB uses QM data (LMP2/cc-pVTZ) at high level of accuracy as reference for dihedral fitting. The modification of ff99SB achieved a good balance between secondary structures and became widely used[26].

A different approach was taken by Duan et al.[27] who introduced a more extensive modification of ff94/ff99 (called ff03), in which a fundamentally different concept to derivation of partial atomic charges was used. Instead of relying on the HF/6-31G* approach to provide aqueous-phase charges, a low-dielectric continuum model corresponding to an organic solvent environment was included directly in the QM calculation of the dihedral parameters and electrostatic potential (from which the charges are obtained). Because of these differences, ff03 should be considered a distinct force field model rather than extension of previous Amber force fields. Ff03 behaves very similarly to ff99SB in short peptides such as Ala_3 and Gly_3, but it may still be slightly over stabilizing α-helices comparing to ff99SB (see Table II and Figure 3 in ff99SB paper[2b]). Considering that in Ala dihedral fitting, φ' and ψ' dihedral parameters were trained against Ala peptides with the presence of φ and ψ dihedral parameters previously trained against Gly peptides. The dependence in training might prevent from an ideal match to the target data.

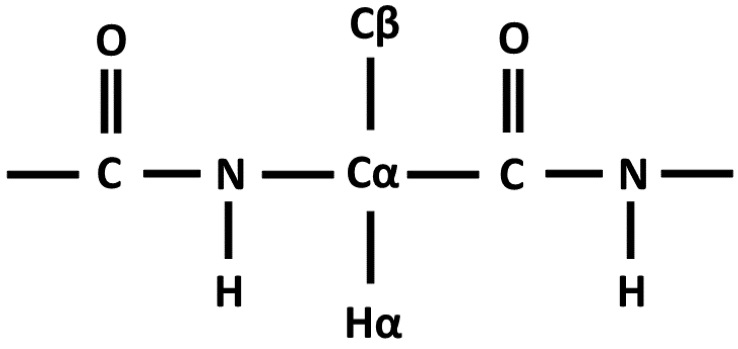


Figure 1.3 The definition of φ(C-N-C_α-C), ψ(N-C_α-C-N), φ'(C-N-C_α-C_β) and ψ'(C_β-C_α-C-N) in Amber residue.

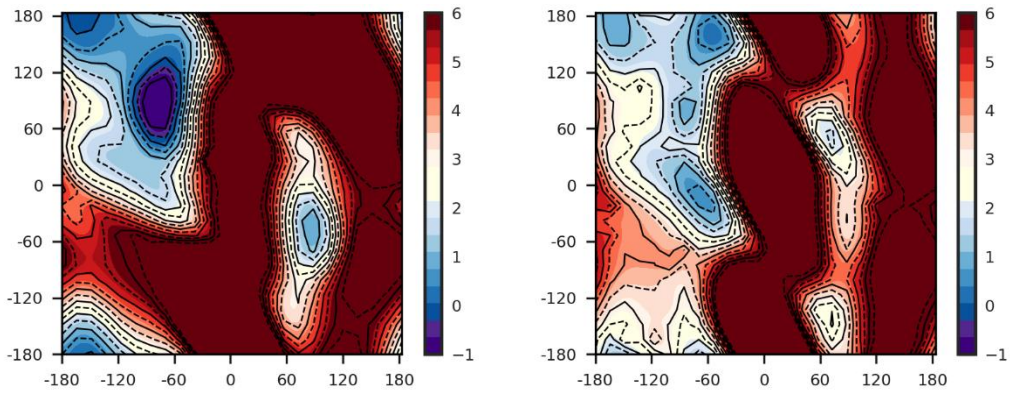


Figure 1.4 Ala dipeptide Ramachandran energy (kcal/mol) surfaces calculated in (left) QM in gas-phase and (B) QM in solution. The values beyond color bar range are depicted in dark red. Solid contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies.

Soon after ff99SB, an extensive data set on in-solution scalar coupling of short peptides provided good opportunity to validate the force field parameters[28]. Several studies noted room for improvements in ff99SB and have suggested that helical structures are not stable enough in ff99SB[29]. Since ff99SB dihedral parameters were trained against gas-phase QM energies (at the time when condensed-phase QM data was not widely available), the problems of reproducing condensed-phase properties (such as helical propensity in solution) might exist due to the lack of polarization effect in fixed-charge force field. Some force fields including ff99SB were compared against these NMR data and the inaccuracy in reproducing these condensed-phase data were noted[26a, 29a]. Taking advantage of these experimental data, a single dihedral term was adjusted to reproduce the fraction of helix measured in solution resulting with later variation of ff99SB (ff99SB*)[29b]. The φ/ψ dihedral parameters were calibrated to reproduce protein chemical shifts resulting with ff99SBnmr[30]. Recently, an array of small empirical perturbations to φ' and ψ' (ff14SB) were designed[2c] to better reproduce scalar coupling data in order to overcome the problem of using gas-phase minima as training set as in ff99SB.

In addition to helical bias in ff99SB, the rotamer preference was also found to be less accurate when comparing to data from rotamer library[31]. The four residues (I, L, D, N) whose rotamer distributions of simulations differed most from PDB were re-parameterized. The χ1 and/or χ2 side chain parameters were refit against high-level gas-phase QM data (MP2/aug-cc-pVTZ), and improvements in reproducing NMR data were noted. In a revised force field[32], charges were refitted on charged residues (K, R, D, E) against QM data. These new charges together with the

ILDN side chain parameters resulted in even better improvements in reproducing experimental data such as helical propensity[32]. A more systematic side chain parameter reworking was performed in ff14SB[2c]. Multi-dimensional conformational scan was performed following with high-level QM calculations to develop specific side chain parameters for each amino acid separately. Ff14SB achieves significant improvement on QM reproduction most of the residues comparing to ff99SB. Moreover, better agreements with side chain scalar coupling data were also noted comparing to both ff99SB and ff99SB_ILDN[2c].

CHARMM is another widely used force field for simulating biomolecules. The first generation is the united-atom CHARMM19[33]. To further improve the accuracy, an all-atom model for proteins was developed as CHARMM22[19]. In contrast to CHARMM19, all atoms were treated explicitly in CHARMM22 force field. The functional form in CHARMM is different from AMBER as it includes a Urey-Bradley 1-3 non-bonded cross term to compensate for the bond stretching and angle bending potential. The CHARMM22 force field is more transferable than CHARMM19 since more atom types (55 in CHARMM22 vs. 29 in CHARMM19) were introduced for proteins and parametrization was done with broader conformational space and better ab initio QM data available[19].

However, the weakness in reproducing QM energy surface and poor sampling in secondary structure such as π helices were noted[34], and a novel strategy of introducing a grid-based correction map was presented resulting with CHARMM22/CMAP force field[35]. The grid-based correction, CMAP, allowed accurate reproduction of φ/ψ QM energy surface. But it is desirable for a force field to reproduce condensed-phase properties, and rigorous reproduction of gas-phase QM data wouldn't always ensure satisfactory reproduction of condensed-phase properties (lack of polarization in force field). Thus it is important to overcome this gap. Best et al.[3c] tried to solve this problem by an iterative empirical perturbation on CMAP values to reproduce condensed-phase measurements such as residual dipolar coupling, scalar coupling and chemical shifts (CHARMM36/CMAP). The propensity to over stabilize helices in C22/CMAP is corrected and more reasonable results for the fraction helix are obtained. The agreement with side chain NMR data for χ1 is improved comparing to C22/CMAP, and fine tuning of χ2, χ3 and χ4 torsion potential left room for future development[3c]. It also highlighted the importance of considering a range of experimental data measured in condensed-phase.

The OPLS force field followed a similar philosophy with AMBER and CHARMM throughout the course of its development. As in united-atom model OPLS_UA[36] and all-atom model OPLS_AA[4b] force field, bonded parameters (bond stretching, angle bending) were retained from AMBER[7, 14], and non-bonded parameters were derived in conjunction with Monte Carlo simulations by computing thermodynamic and structural properties of 34 organic liquids, and torsion parameters were derived from ab initio calculations for more than 50 organic molecules and ions[4]. As computer became more powerful, higher level QM calculations were performed in parametrization, larger systems like tetrapeptides and broader conformational space were employed. In OPLS_AA/L[37], a modification on OPLS_AA, local MP2 theory with basis set cc-pVTZ were used to calculate QM energies of dipeptides. To overcome the limitations of using gas-phase local minima as training, a weighting scheme were introduced to prioritize reproduction of the most important parts of QM energy surface. But it has been argued that the reproduction of condensed-phase properties is still under-explain[37]. As the computer get advanced, the resources permit higher level investigations and further improvements. In OPLS-AA/M[38], higher accuracy QM methods (B2PLYP-D3BJ/aug-cc-pVTZ) were used to optimize both backbone and side chain dihedral parameters with peptide models, and NMR scalar coupling data were used to assess the quality of the new parameters. An overall improvements on both MD sampling and NMR data reproduction over the previous version of OPLS force fields were noted[38]. But since the canonical uncoupled cosine terms were still adopted in their training (unlike the 2D grid-based correction in CHARMM[35a]), a rigorous reproduction of QM surface is not guaranteed[35], and this is an inherent limitation of uncoupled cosine terms. The $\chi 1$ and $\chi 2$ dihedral parameters in OPLS-AA/M was not equally treated during fitting[39]. The $\chi 2$ parameter was only fit if the QM/MM improvement is observed. For Glu and Gln, $\chi 2$ parameters were empirically adjusted to improve the fitting.

A substantial number of studies have tried to compare different force fields and evaluate their accuracy against experimental data over years[29a, 40]. Best and Hummer[29a] performed extensive simulations on Ala_5 in explicit solvent with ff03*, ff99SB, CHARMM27, OPLS-AA/L and GROMOS force fields and compared to NMR J coupling data. They found that most force fields over stabilize α helices with quantitative results depending on the choice of Karplus relation and termini. Piana et al[40a] found that among ff99SB*_ILDN, ff03, CHARMM22* and CHARMM27, the free-energy surface and mechanism of folding of villin headpiece vary substantially even though the agreement with experiments are similarly good. Extensive validation of force fields

against experimental data were performed afterward by Lindarff-Larsen et al[40c] and Pande et al[40b]. Hundreds of NMR measurements on folded proteins and multiple force fields (AMBER, CHARMM and OPLS) were employed in both comparison. The results suggest that force fields have improved over time, while not perfect, provide an accurate description of many structural and dynamical properties of proteins. Robustelli et al[40d] tested force fields against both folded and unfolded proteins and found that none of the tested force fields simultaneously provided accurate descriptions of folded proteins and of the secondary structure propensities of disordered proteins. The force fields tested are state-of-the-art models in AMBER and CHARMM communities.

1.6 Molecular dynamics

The force field method is used not only for single energy calculations, geometries and vibrational frequencies, but also for molecular dynamics (MD) simulations. In MD, the motion of a particle is governed by Newton's second law:

$$F = ma = m\frac{d^2x}{dt^2} \quad (1.4),$$

where F is the force acting on the particle, m is its mass, and dt is the time step for integration; a is the acceleration, given by $a = dv/dt = d^2x/dt^2$, where v is the velocity. In MD, successive configurations of the system are generated by integrating Newton's laws of motions. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time. There is no explicit rule of choosing the most appropriate time step. Too small time step will make the trajectory cover limited phase space and too big will cause instabilities in the integration algorithm. Such instabilities would lead to a violation of energy and could result in a simulation failure due to numerical overflow. The Verlet algorithm[41] is typically used for integrating motion in MD simulation. The Verlet algorithm uses the positions and accelerations at time t, and the positions from the previous step r(t-dt) to calculate positions at t+dt, r(t+dt). Implementation of the Verlet algorithm is straightforward and the data storage is modest. A few variations on Verlet

algorithm have been developed such as leap-frog algorithm[42] in which the verlet velocity is explicitly included.

1.7 Outline

This dissertation is mainly comprised of four chapters. Chapter 2 discusses the training and testing of amino-acid specific backbone dihedral parameters in ff19SB. The details of training including QM and MM calculations, CMAP fitting are elaborated. The extensive tests performed to systematically validate ff19SB were discussed in detail. The results were shown afterward with highlighting the improvements of ff19SB over previous force fields. In chapter 3, the physical cause of errors that were corrected by dihedral parameters in ff19SB are investigated. The investigation is on four aspects: model system, 1-4 empirical scaling factor, partial charges and MM solvation. Both methods and preliminary results are included. The last chapter 4 discusses the future directions of force field development and validation.

Chapter 2

Develop amino-acid specific protein backbone dihedral parameters using quantum mechanics energy in solution

2.1 Acknowledgements

This chapter is adapted with permission from Chuan Tian, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. Journal of Chemical Theory and Computation. DOI: 10.1021/acs.jctc.9b00591.

2.2 Introduction

State-of-the-art computational methods have been able to complement experimental structural biology with information that is both interesting and difficult to obtain without computers. Recent simulation highlights are the time-resolved, atomic-detail folding of ubiquitin during a 1-millisecond MD simulation[43], or the accurate reproduction of a large set of protein-ligand binding affinities[44]. Moreover, simulations are typically used during the refinement of high resolution structures obtained using experimental data such as crystallography, NMR or cryo-electron microscopy. However, two significant caveats apply to the hypothetical power of simulations: (1) the energy function must provide an accurate model of the underlying physics of the system, and (2) the simulation must adequately sample the important regions of the resulting energy landscape. These problems are coupled, and improving the physics model typically gains accuracy at the expense of greater computational cost, reducing the conformational diversity that can be sampled. One of the main challenges in successfully employing simulations is the need to optimize this precision/accuracy compromise based on the requirements of each research project. All-atom molecular dynamics (MD) is likely the most widely used biomolecular simulation sampling method. These often employ simple classical energy functions (force fields, FFs) which usually have many adjustable parameters, most often obtained by fitting to data from experiments or QM. Most modern FFs have very similar functional forms, but differ significantly in choice of model systems and source of the training data. Although using even more complex models than those discussed here (such as including explicit polarizability[45]) may improve accuracy, these gains come at the cost of computational complexity and corresponding reduction in the sampling that is usually the limiting factor in the application of force fields.

Many approximations are made in fitting FF parameters. The FFs used for simulation of biomolecules in water tend to be relatively simple, due to the large number of atom pair interactions that contribute to the overall forces. In this article we focus on the FFs associated with the Amber simulation package[46], though others tend to be very similar. Amber FFs include harmonic terms for covalent structure, such as bond stretching and angle bending. The intramolecular and intermolecular nonbonded interactions are modeled as a Lennard-Jones 12-6

potential for vdW interactions, and a simple Coulomb term for electrostatics typically using fixed partial atomic charges obtained using QM-based electrostatic potentials on intact peptides. The final and crucially important component is the dihedral (torsion) correction terms, which modify the energy of the system as a function of rotation around bonds. These bond rotations control the flexibility of the biopolymer, and different corrections can alter barrier heights as well as the relative energies of various stable rotamers, directly influencing the sampled ensembles[2b].

The physical motivation for the dihedral corrections is that the rest of the FF is purely classical, and therefore lacks quantum orbital effects such as the increased energy barrier for rotation around a double bond. In practice, these corrections are used broadly to empirically optimize force fields during training, accounting for quantum effects as well as other weaknesses in the simple model, such as lack of conformation-dependent polarization that could impact electrostatic interaction profiles, or even to remedy lack of agreement with experiments. In Amber and most other atomistic FFs, the dihedral correction is modeled as a simple truncated Fourier series with amplitudes and phases that are parameters in the FF. These parameters are optimized at the last stage in order to improve the agreement between training data and MM properties calculated without the dihedral terms. Some FFs add one or more additional empirical adjustment steps to improve agreement with experiments.

Importantly, these force fields rely on an implicit assumption that each term is independent, with no coupling between parameters for bonds, angles and dihedrals. This additivity assumption extends to the non-bonded pairs as well, and is a major source of efficiency in force field calculations. In reality, coupling exists to varying extents, and parameters for one component may depend on the conformations of other nearby functional groups. This is neglected in most current biomolecular force fields. Another important key assumption is transferability: that a FF trained on one set of molecules (typically small) will perform as well on different, perhaps much larger molecules. Transferability also applies to neglecting the coupling between parameters, since it is usually assumed that one set of parameters (for example, for rotation around a bond) will perform well for multiple conformations of neighboring groups. Since transferability is imperfect, one way to improve FF accuracy is to ensure that the training data more closely reflect the situations in which the parameters will be applied, and by implicitly accounting for any coupling with neighboring groups at least in a mean-field way. Choice of model systems is therefore crucial.

Enabled by greater computer power, this has led to a trend away from fitting against QM data for small organic compounds[2d, 21] to that for larger peptides.

An important example is the protein backbone φ and ψ dihedral parameters that can alter the energy profiles for these rotations, and thus influence secondary structure preferences and loop conformations. These have been frequently revised over the years based on observations of secondary structure biases in prior models[47]. While early FFs used capped single amino acids (dipeptides) to train the backbone, our ff99SB[2b] FF used tetrapeptides[48], allowing φ and ψ parameters to be trained in a context of conformational diversity of neighboring amino acids in a longer peptide. The improvement was significant, and ff99SB has been widely adopted.

Since that time, widespread use of ff99SB exposed weakness in some amino acid side chain dihedral parameters[32], probably because they were carried over from ff99 which trained them against a limited set of energy minima for simple organic compounds[2d]. In ff14SB[2c], we performed complete refitting of all side chain parameters using QM data for capped amino acids. An important update was the use of multidimensional QM conformational grid scans for every side chain, rather than fitting each rotatable bond separately. Likewise, fitting was done using both α and β peptide backbone contexts. Though it stopped short of explicit dihedral parameter coupling, this approach allowed implicit inclusion of coupling of rotational profiles to neighboring groups in a mean-field way, by fitting parameters for each bond rotational energy profile in the context of multiple conformations of neighboring groups, as was done for the backbone in ff99SB. ff14SB was a notable improvement; for example, a recent study[49] of the ability of protein MD to reproduce high resolution experimental crystal data concluded that ff14SB performed best among all force fields tested, including several older Amber variants and even the empirically tuned CHARMM C36[50].

In addition to the weaknesses in side chain dihedral parameters, some studies also noted weaknesses in ff99SB backbone preferences. Several groups focused on empirically adjusting the ff99SB backbone parameters via comparison to experimental data such as NMR scalar couplings for very short peptides[3c, 29b], or amino acid helical propensities[29b, 32]. Similar to these other groups, we also included in ff14SB a small empirical adjustment to ff99SB (using TIP3P water[2c]) to improve agreement with NMR data for short alanine peptides. Empirical corrections can improve performance on training data but also can be problematic when extrapolated too far. The relative scarceness of experimental data compared to the number of parameters in the FF leaves the

empirical fitting problem severely under constrained. Also, the common target of NMR J coupling data is sensitive to the choice of Karplus parameters[51], they are not equally sensitive to variations in φ and ψ, and the χ^2 values typically used to score performance[29a] can be highly sensitive to small details in the energy landscape yet relatively insensitive to the large differences that are observed between force fields[52]. Fitting backbone parameters to helical propensities is also challenging; it was shown that updating side chain dihedral parameters had a substantial impact on the backbone helical tendencies of some amino acids[32], perhaps because side chain positioning details may play a role in helicity by shielding backbone hydrogen bonds[47b], or due to side chain parameter changes modulating side chain entropy changes, which may influence helix formation[53]. Thus it is possible to erroneously adjust one part of the model (such as the backbone) to improve agreement with experiment, instead of fixing the more fundamental source of the error (e.g., the side chain rotamer energies). Designing or implicitly accepting cancellation of error can lead to models with unphysical and unwanted dependence between components, where one part cannot easily be improved (or even used) without exposing the compensating weakness in another.

Another major challenge to empirical fitting against experiment is deconvolution of the solvent model from the solute FF, each of which may contribute inaccuracies that lead to deviation from experiment. This is complicated further when empirical adjustment creates dependence between solute and solvent models. Shell et al. showed that the best results for predicting small protein structures were obtained using the ff96 force field with the GB-OBC implicit solvent model, despite each having well-established deficiencies[54]. The CHARMM C22[3a] FF was trained using TIP3P water, and backbone refitting was needed to use a different water model[55]. ff14ipq[56] was developed with extensive training to TIP4P-Ew[57] in the initial stages[56], requiring refitting in ff15ipq[58] to enable use with SPC/E_b water[59]. Moreover, weaknesses are apparent in studies of systems that sample diverse ensembles such as the unfolded states of proteins, or simulations of intrinsically disordered proteins (IDPs). This may well arise because of the vast number of nearly degenerate states, and the need for much higher accuracy than what is sufficient for simulating proteins in stable native basins. Currently, the challenge to FFs seems too great; for example, a recent study of IDPs found that the simulated ensembles depended dramatically on FF, but much less so on peptide sequence[60]. Piana et al. showed that the unfolded ensembles in their successful protein folding simulations were much more compact than expected from experiments[61]. Palazzesi compared simulations to NMR data, again finding generally poor agreement regardless of FF

used[62]. In these and other cases, simulated ensembles are generally too compact. Several groups attempted to address the problem empirically by re-training backbone parameters against PDB coil libraries, and flattening energy landscapes[63]. Robustelli et al. carried out extensive refitting to improve the ability of ff99SB to model IDPs while retaining the ability to simulate folded proteins.[40d]

More recent IDP work has implicated overly weak water-protein interactions[51, 64], consistent with other studies showing that protein-protein association in water is too favorable regardless of force field tested[65]. Best et al developed the ff03w model, empirically increasing the water-protein dispersion interaction.[51] Piana et al. developed the TIP4P-D water model, with 50% larger dispersion energies[64b], further adjusted later[40d]. Both adjustments resulted in improved match to IDP experimental data such as Rg values inferred from SAXS and FRET. Recently, the Amber team's new OPC 4-point explicit water model was shown to better reproduce liquid water properties than most other models.[66] It also results in much less compact ensembles for IDPs.[67] Such studies demonstrating that newer water models improve IDP behavior again highlight the dangers in empirically adjusting specific protein FF parameters to fix what may just be a symptom of a different problem. This weakens transferability, and emphasizes the value of independent development and validation of solute and solvent models.

Despite the issues described above, current force fields clearly are good enough to have enabled many excellent biophysical simulation studies. In terms of simulating global structure of proteins of various sequences, protein force fields have improved with time[40b]. Current force fields typically result in stable simulations of folded proteins, with many reports of good match to experimental solution NMR observables such as NOEs, RDCs and S^2 order parameters. More challenging are studies that attempt to predict structure from sequence[68]. A particularly impressive achievement was the successful brute-force folding of ubiquitin in MD simulations[43, 69]. We reported accurate folding via MD for 16 out of 17 diverse proteins up to 100 amino acids long[70]. Despite these successes, a growing number of studies have suggested that even after the recent updates to backbone and side chain parameters, as well as water models, the models still have significant limitations in protein simulations. There is a mounting consensus that current force fields do not accurately reproduce differences between backbone preferences of different amino acids. This is especially apparent in studies where the quantitative relative energies of basins are important, such as analysis of the effect of point mutations, or studies of flexible systems with

many nearly isoenergetic minima. Pande et al. suggested 6 of 19 amino acids were outliers vs. NMR and should be re-optimized[40b]. Best[32] et al. and later, we reported[71] that Amber does not accurately reproduce experimental[72] amino acid specific behavior such as helical propensities, shown in **Figure 2.1** for ff14SB used with TIP3P. Correlation is generally poor, with most amino acids having similar helicity in simulation. In principle, nonbonded interactions should account for the impact of the side chain on backbone energetics (hereafter denoted "sequence dependence"), but weaknesses in the nonbonded function may limit the accuracy in modeling the short-range interactions that are responsible for backbone-side chain coupling.

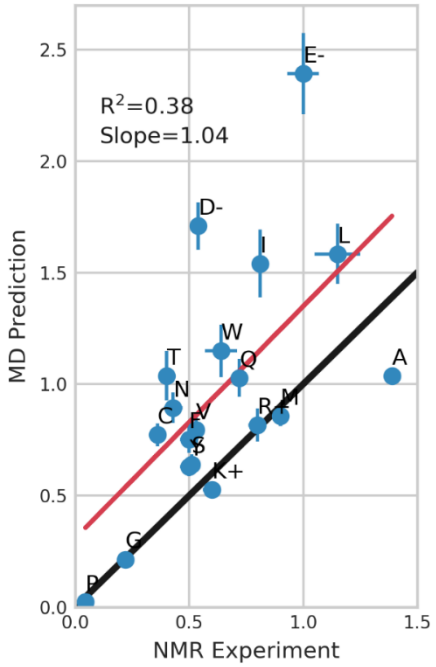


Figure 2.1 Helical propensities in ff14SB+TIP3P (Y) vs experiment[72] (X) for amino acids (1 letter codes). Values on the X-axis represent the data based on NMR and the reported standard deviations.[72] Values on Y-axis represent the helical propensities fit against the combined trajectory (3.2 μs * 12), with error bars calculated via bootstrapping analysis (see **Methods: Bootstrapping analysis on helical propensity**). Black lines represent perfect agreement. Linear regression (red line) was performed against the data points, with R^2 and slope quantifying the goodness of fit.

Importantly, alanine is an outlier in having its helicity significantly under predicted (below the diagonal line). This is concerning since alanine is used as the model system in all recent Amber protein force fields for fitting of backbone dihedral parameters that also are applied to the other amino acids (except Gly). CHARMM also uses the same alanine-based backbone map for all

amino acids except Pro and Gly. The data clearly show that empirical correction of all amino acids using alanine helicity as a target would introduce a significant overall positive helical bias for the remaining amino acids.

β-branched amino acids are not modeled correctly. Experimentally, steric clash between β-branched side chains and the backbone carbonyl reduces helical propensity[72-73]. Troublingly, simulations in ff14SB show that β-branched Ile, Val and Thr all have *higher* or similar helical propensity than Ala, the reverse of the experimental trend (**Figure 2.1**). In high resolution structures of folded proteins, the same trend of backbone-side chain coupling is apparent[74], where the helical basin is narrower in valine than alanine, along with a broader, flatter region at high ψ values corresponding to polyproline 2 (ppII) and β conformations as compared to alanine (**Figure 2.2**). It is challenging for force fields to reproduce these differences, and the alanine and valine MD Ramachandran landscapes are similar using ff14SB (see **Results**). These observations are further corroborated by solution NMR data; higher H^N-H^α scalar couplings for Val dipeptide than Ala dipeptide suggest more structures along the β-ppII transition for valine than for alanine[28], again not reproduced in the MD data (see **Results**). The situation is similar for CHARMM C36, where errors vs. NMR remained large for valine even though the force field was empirically adjusted to obtain a good fit for alanine[45]. Taken together, the results suggest that alanine may not be an ideal model for training other amino acids, in contrast to the central assumption in >20 years of Amber and CHARMM FF development.

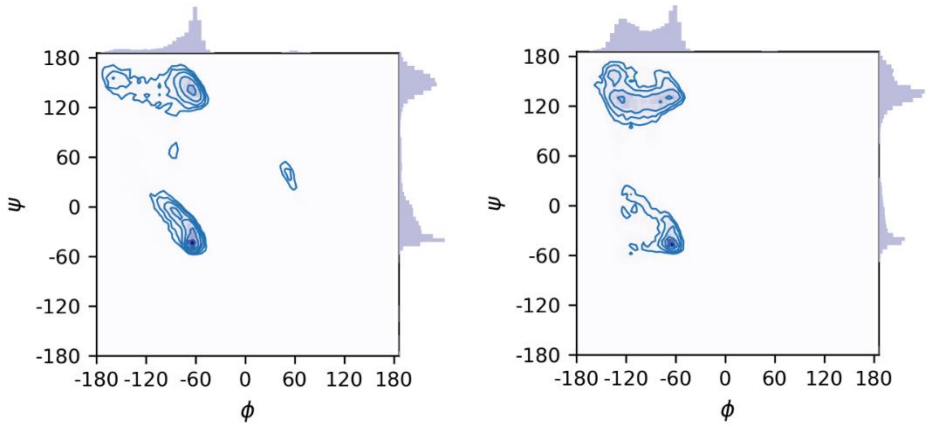


Figure 2.2 Ramachandran sampling in PDB shown for Ala (left) and Val (right) (using data from Lovell et al.[74]) Each contour line represents a doubling in population. Density is also shown as grids filled with light (no density) to dark (maximum density). Side histograms on each subplot represent independent distributions on φ and ψ.

We previously developed empirical backbone corrections for some amino acids in order to improve residue-specific helical propensities[71]. Alternatively, Best et al. found that empirically enforcing the alanine backbone partial charges on all amino acids also resulted in improvement for charged amino acids[32], but this also may have been successful because it eliminated an inconsistency between using atom-specific partial charges and atom-type based dihedral parameters. Other recent work (for example, RSFF[63a, 75] and ff99IDPs/ff14IDPs[63b, 76]) used PDB φ/ψ distributions to develop amino-acid specific empirical backbone parameters. However, in addition to the general problems with empirical fitting discussed above, these crystal data have significant limitations that prevent them from being used as an accurate source of thermodynamic training data (such as inconsistent and cryogenic temperatures, crystal packing effects, limited or noisy data outside low-energy basins, etc.). As a specific example, although the achiral glycine should have a fully symmetric φ/ψ energy profile, PDB-based distributions show significantly enhanced incidence of glycine in the positive φ region[74], which would be reflected erroneously in force fields fit to these statistical distributions.

Going beyond empirical adjustment requires insight into the physical weaknesses in the model. What is the source of this unsatisfactory sequence dependence, despite good reproduction of QM side chain rotational energy profile data[2c] in ff14SB? Speculation leads to several reasonable possibilities, including, but not limited to, lack of charge polarization of the backbone from the side chain (or weaknesses in the charge model overall), the inability of the current functional form to reproduce strong interactions between backbone and bulky side chains, or inaccurate empirical nonbonded scaling factors. Certainly using uncoupled cosine terms for backbone dihedrals limits the accuracy attainable even with ideal QM training data or extensive empirical adjustment. The relative orientation of the two adjacent amides depends on both ϕ and ψ of the intervening amino acid, thus independent cosine terms may be insufficient at correcting the interaction energy or lack of polarization between these groups.

In this work, we revisit the ff14SB protein backbone description with an aim to improve the performance for amino-acid specific behavior discussed above. We hypothesize that several specific weaknesses in the ff99SB strategy may be dominant factors limiting accuracy. (1) Fitting only *alanine* data, and only at the gas-phase minima, poorly constrained the resulting energy landscape for many biologically relevant conformations[77], or at locations of the slightly shifted

φ/ψ minima sampled by other amino acids[74]. (2) The φ/ψ landscape is overly symmetric, arising from neglect of coupling in the simple cosine functional form. (3) Dihedral parameters are shared too broadly due to assignment by simple atom typing that does not discriminate amino acids. (4) Polarization was treated inconsistently in ff99SB and ff14SB, dating back to the original ff94 model. "Pre-polarized" Amber MM partial charges[7, 14] intended for aqueous solution simulations[2a, 8] are used while fitting dihedral parameters against gas-phase QM data, thus forcing the rotational energy profiles back towards the gas phase profiles and thereby counteracting the intended effect of better modeling charge polarization.

We describe here modifications to the protein backbone parameters that at least partially address these issues. We continue our previous philosophy for the Amber "SB" (Stony Brook) force fields, assuming that physics-based force field development can provide excellent models with good transferability beyond their training data. Different approaches also have merit, such as in CHARMM, where physics-based training is followed by iterative rounds of empirical adjustments that improve match to experimental data[50, 78]. The a99SB-disp model[40d] derives from our ff99SB, followed by extensive empirical refitting of torsion parameters, nonbonded pair interactions, atomic partial charges and water dispersion energetics in order to improve agreement with experiments. Likewise, the recent "Force Balance" approach is a promising method to automate iterative improvement through iterative cycles of fitting and comparison to experiment[79]. These adjustments can significantly enhance agreement with experiment, but the complex mapping of experimental observables to individual force field terms can also lead to the introduction of fortuitous (and non-transferable) cancellation of error between the various force field components. We attempt here to overcome the ff14SB weaknesses discussed above by a more self-consistent reconsideration of the physics-based training of protein backbone energetics, developing improved backbone parameters based on fitting to a wider variety of high-level QM, and eliminating a series of inconsistencies in past fitting that are likely to have negatively impacted the resulting models.

The first departure from ff14SB is that we fit coupled φ/ψ parameters using 2D φ/ψ conformational scans, followed by fitting the entire 2D QM energy surface. This will eliminate the problem of unconstrained energies outside the energy minima used to train ff99SB/ff14SB backbone parameters. This also explicitly accounts for coupling between these correction terms. As shown in **Results**, the correction profile needed to match the ff14SB MM to QM for the ψ

rotation differs depending on the value of φ. In other words, in ff99SB/ff14SB, it is not possible to use a 1D correction profile to accurately reproduce QM energy profiles for ψ at all values of φ. This 2D "CMAP" approach was pioneered in the CHARMM force field[80], and extended here. The CMAP approach was also used for backbone fitting in RSFF2+CMAP[75b], but in that case the free energy surface derived from PDB statistics was used as the fitting target, rather than QM data as we use here. Previously, the "CMAP" approach was employed by other Amber force fields as well. In ff99IDPs/ff14IDPs[63b, 76], the 2D energy profile was fitted against statistical data from PDB coil library. In ff12SB-cMAP[71], only the minimum region in CMAP such as α basin and β basin were corrected by fitting to helical propensities and β strand population in MD.

The second difference from ff14SB is that we address the polarization inconsistency during dihedral parameter fitting. While fitting the entire gas phase surface using CMAPs would ensure sampling of energies for regions populated in solution, a significant problem arises during dihedral fitting when comparing *in vacuo* energies between QM and MM. The MM partial charges in most non-polarizable Amber models are traditionally fit to HF-level QM, which results in partial charges larger than expected in the gas phase, intending to mimic the higher dipoles induced in aqueous solution and avoid the need to explicitly include polarization in the FF calculation[2a, 8]. However, using these "pre-polarized" charges to compare to higher level QM providing *gas phase* conformation energies during dihedral fitting introduces error, and enforcing a match results in dihedral parameters that (at least partially) cancel out the effect of charge polarization. The ff03 Amber model addressed this by fitting new charges to QM calculations in low-dielectric organic solvent[81], but the subsequent protocol for backbone dihedral fitting (also in organic solvent) resulted in erroneous double-counting of solvation effects[82]. The recent "ipq" force fields[56, 58] addressed polarization inconsistency by using two independent charge sets, one for MD, fit to QM calculations that included a specific explicit water model[83] that was used in MD simulations, while a second set of gas-phase partial charges was used during fitting dihedrals corrections to gas-phase QM rotational energy profiles. Our approach differs; we train backbone dihedrals using the same pre-polarized MM charges as used in MD, but using continuum aqueous solvation rather than gas-phase energies, and with reference QM data also in aqueous implicit solvent to resolve the gas/aqueous phase inconsistency (following precedent in RNA parameter fitting[82, 84]). An additional benefit is that the resulting dihedral parameters also can absorb conformation dependent

changes in solute polarization that are not reproduced in a fixed-charge model[85] (also absent in the "ipq" models since dihedral fitting is done in the gas phase[56, 58]).

More accurate reproduction of the QM training surfaces and resolving polarization inconsistencies allow us to undertake **the third difference** from ff14SB, that of exploring amino-acid specific correction maps. Amber already used separate parameters for proline and glycine, and finer differentiation is a reasonable next step. In our experience, optimizing amino-acid specific backbone parameters using simple uncoupled cosine terms (as done by other groups[58, 63a]) is unlikely to result in significant improvement for ff14SB since these are not able to accurately reproduce the QM training data even for a single amino acid (see **Results**). For example, despite fitting sets of uncoupled cosine parameters for several groups of amino acids, simulations using the ff15ipq[58] force field show reduced accuracy for β-branched amino acids[58].

Alanine and valine (together with other β-branched isoleucine and threonine) are conformational outliers, justifying separate CMAP treatment. Alanine is very helical, whereas valine has a very flat φ distribution according to PDB φ/ψ distributions (**Figure 2.2**). Many residues exhibit conformational preferences between those of alanine and valine. Leucine is likely a better model for most amino acids (since all but Ala and Gly include a γ-carbon). We therefore used the CMAP fit to Leu for several other amino acids, including those with aromatic rings (Phe, Trp, Tyr) and nonpolar but non-β branched side chains (Met) and the three protonation states of His (His^+, His^δ, His^ε). Polar or charged side chains (Ser, Cys, Thr, Asp^-, Asp, Asn, Glu^-, Glu, Gln, Arg^+, Lys^+) all received individual CMAPs, Pro received its own CMAP and the β-branched Ile used the CMAP fit to the similar Val. Other force fields also fit different parameters for different amino acids. For example in Amber fb15[86], full scanning over φ/ψ and χ1/χ2 dihedrals were performed for each amino acid, then the 4D φ/ψ/χ1/χ2 grid was mapped onto 2D φ/ψ grid by searching for lowest energy side chain conformation at each φ/ψ. Then, uncoupled (1D) cosine functions were used for each dihedral φ, ψ, χ1 and χ2, with all phases and amplitudes fit simultaneously. Here, we fit 2D CMAPs to φ/ψ energy maps using a single rotamer for each amino acid, in order to avoid transferring errors in the χ energy profiles into the φ/ψ correction, as could happen if the φ/ψ grid points also vary in χ values.

Finally, we examine possible dependence of the backbone CMAP on side chain rotamer. In ff99SB and ff14SB backbone training (also CHARMM[50]), the coupling between backbone and rotamer was avoided by using the ff94 approach of Ala as a model for all other amino acids, thus

ignoring any possible backbone-sidechain coupling correction. To account for rotamer dependency in RSFF2+CMAP[75b], the 2-dimensional ϕ/ψ CMAP was supplemented by the use of additional two-dimensional free energy surfaces including $\phi/\chi1$ and $\psi/\chi1$. Here, we find that the 2D CMAPs that we fit to QM data in solution, in combination with the high-quality side chain energy profiles from ff14SB, result in a model that is reasonably transferable to side chain rotamers not included in the training data.

Extensive MD simulations (a total of ~6 milliseconds in explicit water) were performed to validate the performance of the ff19SB model. We show below that ff19SB, using amino-acid specific training against QM data with solvent polarization, reproduces the amino-acid differences in Ramachandran maps much better than ff14SB or other older Amber models. For example, the reproduction of amino-acid specific helical propensity is significantly improved with ff19SB. We also show that the QM-based ff19SB is in reasonable agreement with experiments when combined with an accurate solvent model, while ff14SB performs poorly with the same solvent model and relies on cancellation of error with the less accurate TIP3P model in order to reproduce properties such as the helical content a Baldwin-type peptide. We conclude that an inherent underestimation of helicity is present in ff14SB, which is (inexactly) compensated by an increase in helical content driven that is likely driven by the TIP3P bias[40d, 64a, 87] toward overly compact structures. The improvements in modeling helicity with ff19SB do not appear to result in less accurate performance on β systems. With ff19SB, the overall excellent performance of ff14SB and ff99SB in NMR order parameter reproduction is also generally maintained with even smaller RMSD values relative to experimental structures. Future work will examine the performance of ff19SB on IDP model systems.

2.3 Methods

2.3.1 Structure preparation & simulations

Unless noted otherwise, all crystal and NMR structures were downloaded from the PDB[88] at www.rcsb.org. Alternate structures including fully extended and fully helical used to initiate independent were built via the LEaP module of AmberTools in the Amber v16 software[46]. Helical and extended conformations are defined as (φ, ψ) = (-60°, -45°) and (φ, ψ) = (-180°, -180°) In explicit solvent MD simulations, TIP3P[89], OPC[66], OPC3[90], TIP4P-Ew[57], SPC/E_b[59] and fb3[91] solvent models were used to solvate systems as noted. A truncated octahedron periodic box was used for all simulations. Implicit solvent MD simulations with GBneck2 parameter set[92] of the GBneck solvent model[93] and ff14SB[2c] were performed to generate additional initial structures. ff14SB[2c], ff15ipq[58], fb15[86] and ff19SB were used for explicit solvent MD simulations as noted. System-specific details are discussed below with additional details in **Table 2.1**.

Table 2.1 Systems used for validation of the ff19SB force field. Independent runs represent MD runs starting from random initial velocity and different initial conformation. Force field + solvent model combinations included ff14SB+GBneck2, ff14SB+TIP3P, ff14SB+TIP4P-Ew, ff14SB+OPC, ff19SB+GBneck2, ff19SB+TIP3P, ff14SB+TIP4P-Ew, ff19SB+OPC, ff19SB+OPC3, ff15ipq+SPC/Eb and fb15+fb3.

Peptide PDBID/Sequence	Octahedron box size (Å)	Number of water molecules	Simulation details (MD length * independent runs * force fields * sequences)
Ace-X-Nme	36.0±0.1	997	800 ns * 2 * 4 * 20 (ff14SB/ff19SB+TIP3P/OPC)
$^{+}H_3N$-A5-COOH	36.2±0.1	995	800 ns * 2 * 4 * 1 (ff14SB/ff19SB+TIP3P/OPC)
Ace-A_4XA_4-NH_2	56.2±0.1	3989	3.2 μs * 12 * 4 * 21 (ff14SB/ff19SB+TIP3P/OPC)
Ace-A_4XA_4-NH_2	56.2±0.1	3989	3.2 μs * 12 * 5 * 12 (ff14SB/ff19SB+TIP4p-Ew, ff19SB+OPC3, ff15ipq+SPC/E_b, fb15+fb3)
Ace-A_4XA_4-NH_2	--	--	2 μs * 12 * 1 * 21

			(ff14SB+GBneck2)
Ace-A_9XA_9-NH_2	--	--	2 μs * 12 * 1 * 21 (ff14SB+GBneck2)
Ace-GGG$(KAAAA)_3$K-NH_2	60.6±0.1	4978	3.2 μs * 10 * 3 * 1 (ff14SB+TIP3P/OPC, ff19SB+OPC)
CLN025, 2RVD, $^{+}H_3N$-YDPETGTWY-COO^{-}	64.0±0.1	5989	7.2 μs * 8 * 3 * 1 (ff14SB+TIP3P/OPC, ff19SB+OPC)
GB3, 1P7E, 56-residue	57.6±0.0	3944	200 ns * 4 * 3 * 1 (ff14SB+TIP3P/OPC, ff19SB+OPC)
Ubiquitin, 1UBQ, 76-residue	62.1±0.0	4923	200 ns * 4 * 3 * 1 (ff14SB+TIP3P/OPC, ff19SB+OPC)
Lysozyme, 6LYT, 129-residue	66.4±0.0	5870	200 ns * 4 * 3 * 1 (ff14SB+TIP3P/OPC, ff19SB+OPC)

Dipeptides

Acetyl and N-methyl capped dipeptides of the natural amino-acids (Ace-X-Nme) were used for force field training and testing. In training, 16 amino acids (including two protonation states of Asp and Glu, but excluding Ile, Trp, Tyr, Phe, Met and His) were fully scanned in backbone dihedral space using implicit solvation (see Structure preparation & simulations and Geometry scanning). In testing, helical and extended conformations for all natural amino acids (including two protonation states each for Glu and Asp side chains, and three protonation states for His side chain) were used as initial structures in 800ns MD simulations. The number of explicit water molecules was equalized across all dipeptide systems and solvent models (**Table 2.1**).This was achieved by adjusting the value of buffer distance until desired number of water molecules was obtained. Four combinations including ff14SB[2c]+TIP3P[89], ff14SB[2c]+OPC[66], ff19SB+TIP3P[89] and ff19SB+OPC[66] were tested for dipeptides.

Ala_5

Ala_5 with a free N- and protonated C-terminus was used in simulation, corresponding to pH=2 used in the NMR studies[28] (see Parameter derivation for protonated C-terminal Ala). Both helical and extended conformations were used as initial structures for 800ns MD simulation. The number of water molecules was equalized across all runs (**Table 2.1**). Four combinations including ff14SB[2c]+TIP3P89, ff14SB[2c]+OPC66, ff19SB+TIP3P[89] and ff19SB+OPC[66] were tested for Ala_5.

$A_4\underline{X}A_4$ and $A_9\underline{X}A_9$ peptides

Acetyl and NH2 capped polypeptides (matching pH=7 in NMR[72]) of the 20 natural amino-acids (A4**X**A4: Ace-$A_4\underline{X}A_4$-NH_2 where **X** denotes the amino acid tested) were used to test amino-acid specific helical propensities. Two independent runs of 800 ns each starting from the helical and extended conformations were initially performed with ff14SB[2c]+GBneck2[92], and cluster analysis (see **Cluster analysis**) was carried out on the combined trajectory. Cluster centroids from the top four clusters, together with helical and extended conformations were then selected as initial structures for MD simulations in explicit solvent. Each of these six initial structures seeded 2 independent runs each with different initial velocity assignment (using ig=-1 in Amber). Therefore, a total of 12 initial states were simulated for 3.2 μs each, in each explicit solvent (~4000 water molecules for both OPC and TIP3P runs, see **Table 2.1**, for each one of the 20 peptide sequences, for a total of 768 μs for each force field + solvent model combination. Helical propensities were calculated using eight FF+water combinations including ff14SB[2c]+TIP3P[89], ff14SB[2c]+TIP4P-Ew[57], ff14SB[2c]+OPC[66], ff19SB+TIP3P[89], ff19SB+TIP4P-Ew[57], ff19SB+OPC3[90], ff19SB+OPC[66], ff15ipq[58]+SPC/E_b[59] and fb15[86]+fb3[94].

Acetyl and NH2 capped polypeptides of the 20 natural amino-acids in a longer peptide (A9**X**A9: Ace-$A_9\underline{X}A_9$-NH2 where **X** denotes the amino acid tested) were used to test the sensitivity of the helical propensities to chain length. Two independent runs, starting from helical and extended conformations, were initially performed for 800 ns with ff14SB+GBneck2, and cluster analysis (see **Cluster analysis**) was carried out on the combined trajectory. Cluster centroids from the top four clusters were then selected as initial structures for additional MD simulations in GBneck2. Each of these six initial structures seeded 2 independent runs with different initial velocity assignment (using ig=-1 in Amber). Therefore, a total number of 12 initial

states were simulated in ff14SB[2c]+GBneck2[92] for each one of the 20 Ace-A_9**X**A_9-NH2 systems, and each simulation was 2 μs long, for a total of 480 μs. These A9XA9 results were compared to data from A4XA4 (also in ff14SB+GBneck2) by extending the 800ns simulations described above to 2 μs.

K19 helical peptide

Consistent with our previous work[2c, 95], the sequence of Ace-GGG(KAAAA)$_3$K-NH2 was chosen to validate parameter quality in folding helices. Since it was unfeasible to run long simulations starting from fully extended conformations that require very large numbers of water molecules to solvate, a fully extended conformation was not selected for explicit solvent simulations. For instance, 12000 water molecules would be needed to solvate a fully extended conformation of K19 with 8 Å buffer. Instead, several semi-extended initial conformations were generated. Two independent runs starting from helical and extended conformations were run for 800 ns with ff14SB[2c]+GBneck2[92], and clustering analysis (see **Cluster analysis**) was performed on the combined trajectory. The cluster centroids (**Figure 2.3**) from the top 1st and 2nd were disregarded because both were partially helical with 2.7 Å and 4.4 Å RMSD (backbone C, N, CA atoms) referenced to a fully helical conformation. Therefore, the centroids from top 3rd (c2), 4th ,(3), 5th (c4) and 6th (c5) clusters were selected as semi-extended. Both semi-extended and helical conformations are immersed in explicit water. The number of water molecules was equalized across all runs (**Table 2.1**). Each initial structure was used for 2 independent runs with random initial velocity assignment (ig=-1 in Amber). Therefore, a total of 10 initial states were simulated with each force field + explicit solvent combination, and each simulation was 3.2 μs. Three combinations including ff14SB[2c]+TIP3P[89], ff14SB[2c]+OPC[66], and ff19SB+OPC[66] were tested for K19.

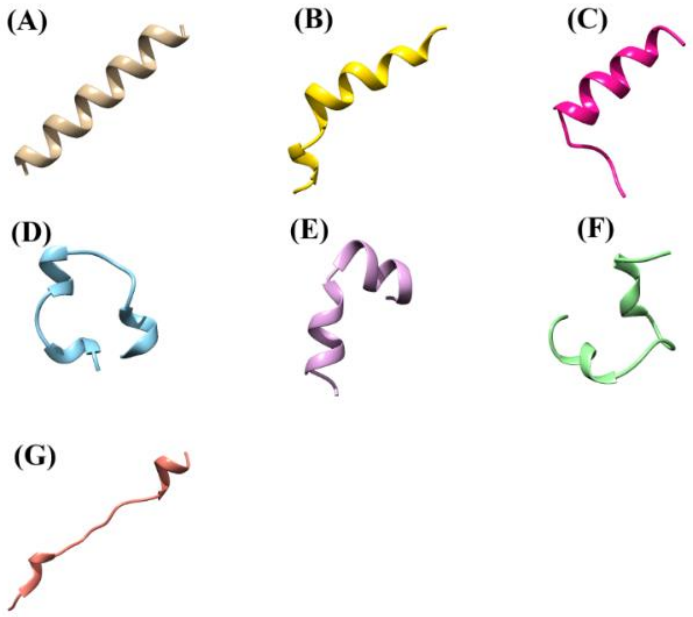


Figure 2.3 Representative conformations (depicted in ribbon) of K19 including (A) fully helical conformation and cluster centroids from (B) top 1st (c0), (C) 2nd (c1), (D) 3rd (c2), (E) 4th (c3), (F) 5th (c4) and (G) 6th (c5) clusters. Only (A), (D), (E), (F) and (G) were selected for K19 MD.

CLN025 hairpin

CLN025 (PDBID: 2RVD[96], $^{+}H_3N$-YDPETGTWY-COO^{-}) is an engineered fast-folding hairpin that is a thermally optimized variant of Chignolin[97]. The native conformation was chosen as the 5th conformation in the NMR ensemble[96] since that conformation was closest to the average of the NMR ensemble. A fully extended conformation of the same sequence was also used, and 4 independent runs (ig=-1 in Amber) were performed with an explicit solvent for both native and extended conformations. Each simulation was 7.2 μs long and the number of water molecules was equalized across all runs (**Table 2.1**). Three combinations including ff14SB[2c]+TIP3P[89], ff14SB[2c]+OPC[66], and ff19SB+OPC[66] were tested for CLN025. A cutoff of 1.5 Å RMSD was chosen to delineate native from non-native structures because the highest population peak at low RMSD across all force field + solvent models ends near 1.5 Å (**Figure 2.4**).

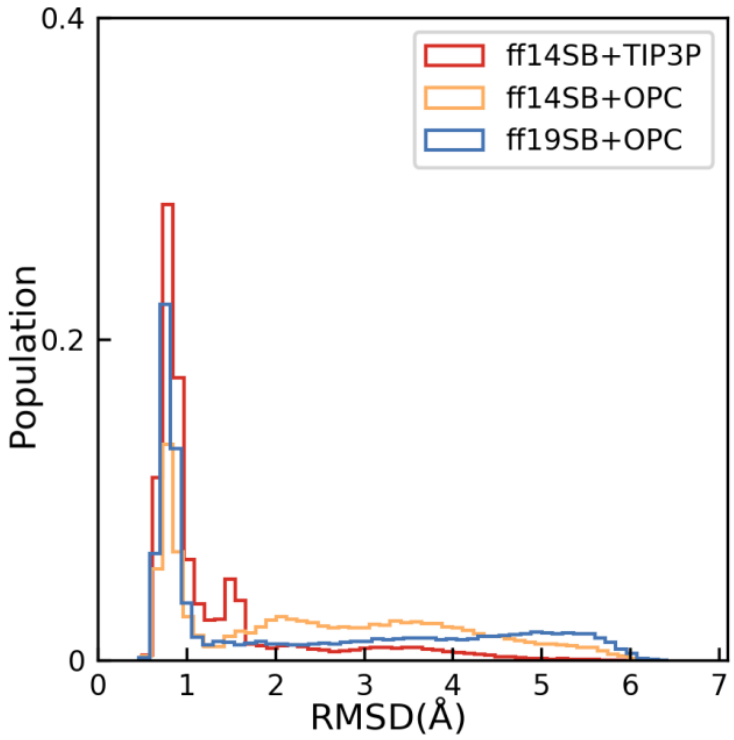


Figure 2.4 Backbone RMSD histograms for the combined four extended (ext) and four native (nat) runs of CLN025 with ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC. Y-axis represents normalized population and X-axis represents the RMSD to the NMR structure (PDBID: 2RVD[96]).

Folded proteins

Three folded proteins were simulated for comparison to NMR-based backbone dynamics measurements. First was the third Igg-binding domain of protein G (GB3). The native structure was defined from a liquid crystal NMR structure (PDBID: 1P7E[98]). Second was Ubiquitin (Ubq), with the native structure defined from a crystal structure (PDBID: 1UBQ[99]). Third was hen egg white Lysozyme (HEWL), with the native structure defined from a crystal structure (PDBID: 6LYT[100]). Four independent runs with random initial velocity assignment (ig=-1 in Amber) were performed for each system in explicit solvent. Each simulation was 200 ns long and the number of water molecules was equalized across runs for each system (**Table 2.1**). These folded proteins were tested using three combinations including ff14SB[2c]+TIP3P[89], ff14SB[2c]+OPC[66], and ff19SB+OPC[66].

2.3.2 Geometry scanning

Backbone geometry scans were performed to generate structures for parameter training. All scans were carried out via the LEaP module of AmberTools in Amber v16 software[46, 101]. All 16 dipeptides (see **Dipeptides**) were 2D scanned on φ and ψ dihedrals over ranges of -180° to 165° with an interval of 15°. For glycine dipeptide, a finer grid scanning was performed in the beta region: -180° to -125° and 120° to 175° on φ and ψ dihedrals with an interval of 5° resulting with an additional 12*12 finer grid. This was done because the QM energy surface in beta region is highly sensitive to the structure/energy of the picked grid point and using 15° interval might unintentionally miss the structure/energy in the "actual" minimum. For proline dipeptide, structures were limited to -180° to 120° on φ in order to exclude structures with excessive ring strain. For dipeptides containing one or more heavy atom χ dihedrals (Val, Leu, Ash, Asp^-, Asn, Glh, Glu^-, Gln, Lys^+, Arg^+, but excepting Ser, Cys and Thr, see below; Ash and Glh are neutral Asp and Glu, respectively), χ dihedral values were initialized to the most populated rotamer for that amino acid, according to Lovell's rotamer library[74].

2.3.3 Molecular mechanics (MM) optimization and energy calculations

For Cys and Met, Lennard-Jones (LJ) parameters were taken from GAFF2 for sulfur and hydrogen (in -SH and -S-), and also incorporated into ff19SB. This was done to keep consistent with the most recent LJ parametrization on these atoms performed by Wang et al[102].

Unless otherwise noted, use of the term "GBSA" in this paper denotes the combination of GBneck2 (igb=8 in Amber) and SASA (gbsa=1 in Amber).

Dipeptide structures were minimized with restraints after geometry scanning. MM optimization and energy calculations were performed with Amber v16[46, 101] using ff14SB[2c] and GBneck2[92] implicit solvent model with the mbondi3 radii set[92] for polar solvation and SASA-based nonpolar solvation[103]. The default 0.005 kcal mol^{-1} $Å^{-1}$ surface tension was adopted.

Dipeptides taken from geometry scanning were minimized using ff14SB[2c] and GBSA including restraints on φ and ψ values with harmonic force constant of 1000 $kcal\ mol^{-1}\ rad^{-2}$. All χ dihedrals were relaxed during minimization without restraints, except Ser, Cys and Thr, for

which the χ_2 dihedral (defined as CA-CB-OG-HG for Ser, CA-CB-SG-HG for Cys and CA-CB-OG1-HG1 for Thr) was restrained (10 *kcal mol^{-1} rad^{-2}*) to 165° to prevent the hydroxyl group (O-H bond) from approaching too closely to the backbone amides during minimization. As we noted for ff14SB[2c], this was done to avoid incorporating into the backbone dihedral parameters any difference between the quantum mechanical (QM) and MM models in the short-range potential between side chain and backbone. Our strategy assumes that the largest contribution to rotamer dependency is errors in the MM short-range nonbonded model, which may be present for backbone conformation using a rotamer with steric clashes or strong electrostatic interactions. If correction to these errors were to be incorporated into the backbone parameter for that φ/ψ grid point, it consequently would be applied for conformations sampling the same φ/ψ values but with different rotamers that lack these inaccurate interaction energies.

We adopted the strategy of initializing all structures on the grid at the same rotamer conformation, then minimizing with backbone restraints to relax the rotamer to a local minimum. The rationale for using a single initial rotamer for the entire φ/ψ grid scan is to reduce the likelihood of transferring any errors in the ff14SB side chain rotamer energy profiles to the CMAP (which can occur if neighboring grid points also differ significantly in χ dihedral values). The same relaxed rotamer was used in the QM calculations (discussed below).

Structures were minimized for a maximum of 10,000 cycles in ff14SB+GBSA with no cutoff on non-bonded interactions. Steepest descent was employed for the first 10 cycles in the minimization and conjugate gradient for the following cycles. Single point energies were calculated for the MM-optimized structures using ff14SB00+GBSA. ff14SB00 is defined as the original ff14SB[2c] force field with the amplitudes of dihedrals sharing the same central two atoms with φ and ψ (C-N-CA-C, C-N-CA-CB, N-CA-C-N, CB-CA-C-N, HA-CA-C=O) set to zero (**Table 2.2**). The convergence criterion for energy gradient is when the root-mean-square of the Cartesian elements of the gradient is less than 10^{-4} *kcal mol^{-1} Å^{-1}*.

Table 2.2 AMBER standard frcmod file for the modified ff14SB00.

ff14SB00				
MASS				
SH	32.06	2.900		
S	32.06	2.900		

HS	1.008	0.135		
DIHE				
H1-CX-C-O	1	0.000	0.0	-1
H1-CX-C-O	1	0.000	0.0	-2
H1-CX-C-O	1	0.000	180.0	3
C-N-CX-C	1	0.000	0.0	-4
C-N-CX-C	1	0.000	0.0	-3
C-N-CX-C	1	0.000	0.0	-2
C-N-CX-C	1	0.000	0.0	1
N-CX-C-N	1	0.000	0.0	-4
N-CX-C-N	1	0.000	180.0	-3
N-CX-C-N	1	0.000	180.0	-2
N-CX-C-N	1	0.000	180.0	1
C8-CX-N-C	1	0.000	0.0	-4
C8-CX-N-C	1	0.000	0.0	-3
C8-CX-N-C	1	0.000	0.0	-2
C8-CX-N-C	1	0.000	0.0	1
CT-CX-N-C	1	0.000	0.0	-4
CT-CX-N-C	1	0.000	0.0	-3
CT-CX-N-C	1	0.000	0.0	-2
CT-CX-N-C	1	0.000	0.0	1
2C-CX-N-C	1	0.000	0.0	-4
2C-CX-N-C	1	0.000	0.0	-3
2C-CX-N-C	1	0.000	0.0	-2
2C-CX-N-C	1	0.000	0.0	1
3C-CX-N-C	1	0.000	0.0	-4
3C-CX-N-C	1	0.000	0.0	-3
3C-CX-N-C	1	0.000	0.0	-2
3C-CX-N-C	1	0.000	0.0	1

N-C-CX-C8	1	0.000	0.0	-4
N-C-CX-C8	1	0.000	0.0	-3
N-C-CX-C8	1	0.000	0.0	-2
N-C-CX-C8	1	0.000	0.0	1
N-C-CX-CT	1	0.000	0.0	-4
N-C-CX-CT	1	0.000	0.0	-3
N-C-CX-CT	1	0.000	0.0	-2
N-C-CX-CT	1	0.000	0.0	1
N-C-CX-2C	1	0.000	0.0	-4
N-C-CX-2C	1	0.000	0.0	-3
N-C-CX-2C	1	0.000	0.0	-2
N-C-CX-2C	1	0.000	0.0	1
N-C-CX-3C	1	0.000	0.0	-4
N-C-CX-3C	1	0.000	0.0	-3
N-C-CX-3C	1	0.000	0.0	-2
N-C-CX-3C	1	0.000	0.0	1
NONB				
SH	1.9825	0.2824		
S	1.9825	0.2824		
HS	0.6112	0.0124		

2.3.4 CMAP fitting groups

A total of 16 CMAPs were fit and then applied to the 20 natural amino acids with several having alternate protonation states (**Table 2.3**). Ala, Gly, Pro were fit separately because the allowable regions in Ramachandran plot according to PDB are notably different from each other.[104] Ser, Cys and Thr were fit separately from others because of the proximity of the polar group to the backbone, and from each other because the polarity of their side chains is different (Ser vs. Cys) or the side chain β-branching structure is different (Ser vs. Thr). Val CMAP was fit and applied to both Val and Ile since Val and Ile are the only two amino acids having β-branched

non-polar side chain. Arg^+, Lys^+, Asp^-, Ash, Asn, Glu^-, Glh and Gln were fit separately because the charge state is different (Arg^+ and Lys^+ vs. Asp^- and Glu^-), the polarity of side chain is different (Arg^+ vs. Lys^+, Asp^- vs. Ash vs. Asn, Glu^- vs. Glh vs. Gln), or the length of side chain is different (Asp^- vs. Glu^-, Ash vs. Glh, Asn vs. Gln). Leu CMAP was fit and applied to long non-polar and non-charged side chains including amino acids with aromatic rings (Phe, Trp and Tyr), Met, Cys in disulfide bonds (Cyx) and Cys interacting with metal (Cym). Leu CMAP was also applied to the three protonation states of His (His^+, His^ε, His^δ).

Table 2.3 Amino acid used to fit CMAP for each of the standard amino acids.

Amino acid	CMAP model
Gly	Gly
Ala	Ala
Val, Ile	Val
Ser	Ser
Cys	Cys
Thr	Thr
Leu, Cyx, Cym, Met, Phe, Trp, Tyr, His^+, His^ε, His^δ	Leu
Asp^-	Asp^-
Ash	Ash
Asn	Asn
Glu^-	Glu^-
Glh	Glh
Gln	Gln
Arg^+	Arg^+
Lys^+, Lyn	Lys^+
Pro	Pro

2.3.5 CMAP fitting

A CMAP is defined by a 24*24 grid that is evenly spaced (15°) in φ/ψ dihedral space, the same spacing as used in C22/CMAP[3b], C36[3c], C36m[3d] and RSFF[63a, 75] force fields. At each grid point, the energy $U_{cmap}(\varphi, \psi)$ corresponds to the following:

$$U_{cmap}(\varphi,\psi) = E_{QM}^{gas} + E_{QM}^{polarization} + E_{QM}^{solvation} - (E_{MM}^{ff14SB00} + E_{MM}^{solvation}) \quad (2.1),$$

where E_{QM}^{gas} represents gas-phase QM energy, $E_{QM}^{polarization}$ represents the contribution from solute-solvent polarization from QM solvation and $E_{QM}^{solvation}$ represents the remaining specific solvation effects in QM. $E_{MM}^{ff14SB00}$ represents MM energy calculated in ff14SB00 (**Table 2.2**) using pre-polarized charges, and $E_{MM}^{solvation}$ represents MM solvation energy calculated in GBSA. In practice, E_{QM}^{gas}, $E_{QM}^{polarization}$ and $E_{QM}^{solvation}$ cannot be separated since the solute electron density is evaluated self-consistently with the solvent polarization represented in a reaction field.

In Amber, the bicubic spline function is fit once against the 24*24 grid values of the CMAP, and is later used to interpolate MM energy at any arbitrary φ/ψ dihedral. The bicubic spline function for each residue is as following:

$$U_{cmap}(\varphi,\psi) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}\phi^i\psi^j \quad (2.2),$$

where φ and ψ are dihedral values in radians, and a_{ij} are the coefficients of the bicubic spline function that are solved from a set of linear equations derived from values at the four corners of the grid cell. The resulting CMAP forces are calculated by the chain rule and added to the total forces[105]. The CMAPs are intended to be used as direct replacement for the old cosine-based dihedral terms in ff14SB.

The CMAP code was originally implemented in AMBER with the support of CHAMBER module[105]. The LEaP module reads the ff19SB frmod file and locates the CMAP section. Then, one type of CMAP will be assigned to each protein residue based on matching residue names listed in the frcmod CMAP_RESLIST with those in the molecule. The information of each residue-based CMAP (index, grid values, etc) together with the atom indices referring to five backbone atoms (C, N, CA, C, N) of the corresponded residue are written into the Amber topology file. The atom indices are linked to residue-based CMAP parameters by CMAP index. The MD engine (pmemd, sander) reads the topology file and fits bicubic spline functions for each CMAP, and calculates CMAP forces for each residue based on the listed five atoms and the bicubic function of the indexed CMAP.

2.3.6 QM energies in solution

To calculate $E_{QM}^{gas} + E_{QM}^{polarization} + E_{QM}^{solvation}$ (**Equation 2.1**), we used the SMD solvent model that includes both polar and nonpolar solvation components[106]. The polar component uses the integral-equation-formalism polarizable continuum model (IEFPCM)[107], where the solute cavity is defined through superposition of atom-centered spheres with reparametrized "intrinsic radii". The non-polar component is a product of the solvent-accessible surface area (SASA) and the surface tension, which is a function of several element-specific parameters. These empirical parameters for effective radii and surface tension were iteratively optimized to reproduce 2346 solvation free energies of both neutral solutes and ions[106]. In the original work[106], the authors concluded that among various QM theories used in their parameter fitting, the DFT method M05-2X[108] yielded the best performance. Taking these results into consideration, particularly performance for amides, we selected the hybrid functional M05-2X with basis set 6-311G** together with SMD to compute the total solvation energy in QM. In the original paper[106], 6-31+G** was shown to have smaller mean unsigned error in aqueous solvation free energy for all tested molecules, including four amides, compared to other basis sets such as MIDI!6D, 6-31G* and cc-pVTZ. The diffuse functions in 6-31+G**, however, cause convergence issues in some of our calculations where the geometries are far from equilibrium. Instead, we use the comparable 6-311G** basis set. We also tested whether M05-2X/6-311G** is accurate in calculating conformational energies of Ala dipeptide in gas-phase.

Nine conformations of Ala dipeptide were chosen with φ dihedral to be -150°, -60°, and 60°, and ψ dihedral to be 60°, -45° and 150°. QM calculations were performed with Gaussian 091. First, geometry optimization was performed on the nine conformations at B3LYP/6-31G* level of theory with φ/ψ restraints. Then, single point energies were calculated at levels of theory including HF/6-31G*, MP2/6-311++G**, B3LYP/6-31G*, wB97XD/6-311G**, M05-2X/6-311G**, B3LYP/Aug-cc-pVTZ, B3LYP/Def2QZVP, M06-2X/cc-pVTZ, MP2/cc-pVTZ, MP2/Aug-cc-pVTZ and MP2/cc-pVQZ. The average REE was calculated for each of the methods against MP2/cc-pVQZ and plotted in **Figure 2.5**.

The calculated average relative energy error (see **2.3.18 Average relative energy error (REE) calculation**) against MP2/cc-pVQZ for nine conformations of Ala dipeptide is ~0.35 kcal/mol, very close to MP2/cc-pVTZ level of accuracy (average REE = 0.2 kcal/mol). Based on

our results, M05-2X/6-311G** is reasonably accurate relative to MP2/cc-pVQZ at reproducing relative energy of Ala dipeptides in gas phase (**Figure 2.5**), and errors are likely comparable to those arising from other sources such as the spacing of the grid scan and fundamental inaccuracies in the MM treatment.

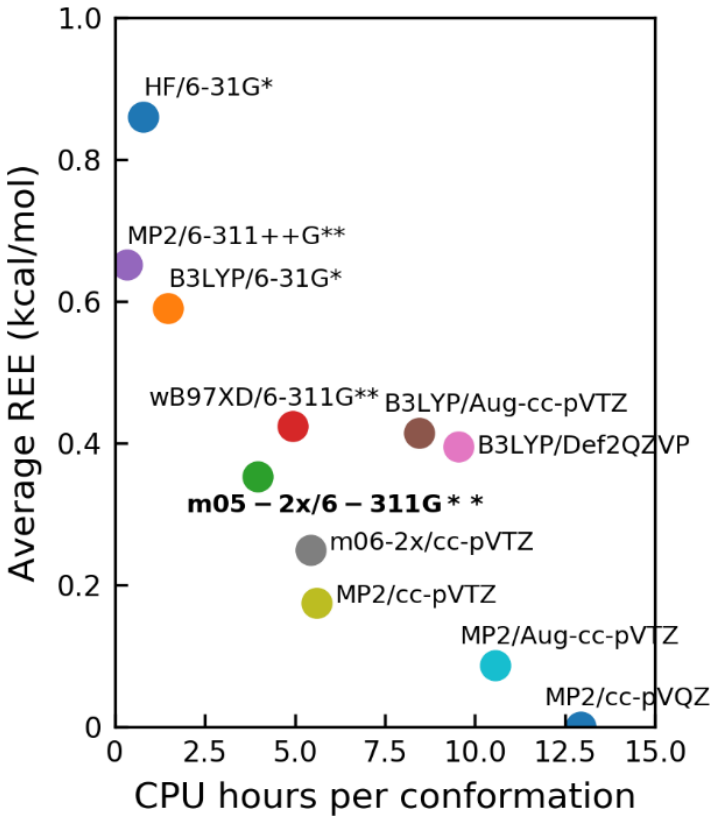


Figure 2.5 The average relative energy error (REE) of nine Ala dipeptide conformations versus CPU hours per conformation for various QM theory and basis set combinations. The MP2/cc-pVQZ energy was used as reference for error calculations.

2.3.7 QM optimization and energy calculations

QM calculations were performed with Gaussian 09[109]. Geometry optimizations and single point energy calculations were performed on the 16 dipeptides at the M05-2X/6-311G**/SMD level of theory[108]. Grimme's dispersion correction with the original D3 damping function[110] was used to correct for long-range dispersion. The solvation environment was represented as a self-

consistent reaction field, with exterior dielectric set to default 78.3553, using SMD[106] with consideration of both polar and nonpolar solvation energy components.

Very tight optimization convergence criterion was used to generate data for fitting. To maintain the structure on the φ/ψ grid, one of the dihedrals sharing the same central two atoms with φ, and one dihedral sharing the central two atoms with ψ were restrained to the values from the structures taken from the last step of MM optimization. In order to avoid inclusion of errors in the χ energy profiles into the QM-MM energy difference used for CMAP fitting, we also restrained one of the dihedrals for each χ dihedral to the value from the last step of MM optimization (see **Molecular mechanics (MM) optimization and energy calculations**) (details on restrained dihedrals provided in **Table 2.4**).

Table 2.4 Dihedrals to be restrained during QM optimization.

Amino acid	Dihedrals in atom name
Gly	C-N-CA-C, N-CA-C-N
Ala	C-N-CA-C, N-CA-C-N
Val	C-N-CA-C, N-CA-C-N, N-CA-CB-CG1
Ser	C-N-CA-C, N-CA-C-N, N-CA-CB-OG, CA-CB-OG-HG
Cys	C-N-CA-C, N-CA-C-N, N-CA-CB-SG, CA-CB-SG-HG
Thr	C-N-CA-C, N-CA-C-N, N-CA-CB-OG1, CA-CB-OG1-HG1
Leu,	C-N-CA-C, N-CA-C-N, N-CA-CB-CG, CA-CB-CG1-CD1
Asp^-	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-OD1
Ash	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-OD2, CB-CG-OD2-HD2
Asn	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-OD1
Glu^-	C-N-CA-C, N-CA-C-N, N-CA-CB-CG.

	CA-CB-CG-CD, CB-CG-CD-OE1
Glh	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-CD, CB-CG-CD-OE1
Gln	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-CD, CB-CG-CD-OE1
Arg^+	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-CD, CB-CG-CD-NE, CG-CD-NE-CZ
Lys^+	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-CD, CB-CG-CD-CE, CG-CD-CE-NZ
Pro	C-N-CA-C, N-CA-C-N, N-CA-CB-CG. CA-CB-CG-CD

For glycine dipeptide, QM optimization and energy calculations were done on both 24*24 coarse grid with 15° interval (same as other amino acids) and 12*12 sub-grid with 5° interval (specific to glycine). In the region with fine grid data, the QM energies of coarse grid points were replaced with lowest energy in the surrounding fine grid points (within 10° from the coarse grid). This was only done for QM energy calculations on Gly. MM calculations on Gly were performed the same as other amino acids, and the 24*24 grids were used to obtain the CMAPs.

2.3.8 Parameter derivation for protonated C-terminal Ala

Following the original RESP method for peptide partial charge assignment[2d, 8], new charges were trained for Ala with acetylated N-terminus and protonated C-terminus. Helical and extended conformations were used for RESP fitting. The partial charges on all atoms except the –COOH group were restrained to the charges from ff94[2a]; –COOH group charges were refit via RESP. QM calculation was performed with Gaussian 09[109]. HF/6-31G* was used for geometry optimization. MK[111] population analysis was performed on the optimized geometry. Antechamber, espgen and residuegen as implemented in Amber v16[101] were used in RESP fitting.

The resulting atomic charges are listed in **Table 2.5**. The –COOH functional group in the protonated C-terminal Ala was assigned the same atom types as –COOH in side chains of Alh or

Glh, thus sharing existing bonds, angles, dihedral and LJ parameters. When simulating a system with a protonated (uncapped) C-terminal Ala in ff19SB, the ff14SB parameters were applied to the C-terminal residue without application of a CMAP due to lack of C-terminal amide.

Table 2.5 Partial charges for protonated C-terminal Ala.

Atom name	Atom type	Partial charge
N	N	-0.4157
H	H	0.2179
CA	CX	0.0337
HA	H1	0.0823
CB	CT	-0.1825
HB1	HC	0.0603
HB2	HC	0.0603
HB3	HC	0.0603
C	C	0.6717
OX1	O2	-0.5303
OX2	O2	-0.6122
HX	HO	0.5001

2.3.9 MD simulations

The following methods were used for all MD simulations unless otherwise noted. Bonds to hydrogen atoms were constrained with the SHAKE algorithm[112] using a geometrical tolerance of 0.000001Å. The direct space non-bonded interaction cutoff was 10.0 Å for explicit solvent simulations and 9999.0 (no cutoff) for implicit solvent simulation. Long-range electrostatic interactions in explicit solvent were calculated via the particle mesh Ewald (PME) approach[113].

There were a total of 9 steps of equilibration in both implicit and explicit simulations. For explicit solvent simulation, initially energy minimization was performed for up to 10000 cycles. All atoms except water and H atoms used positional restraints with force constant 100 $kcal\ mol^{-1}\ rad^{-2}$. Steepest descent was applied for the first 10 cycles and conjugate gradient was applied for the following cycles in the minimization. The same restraints were maintained as the system was heated in NVT starting from a low temperature of 100 K and reaching 298 K over 1 ns of

simulation time. Langevin coupling (ntt=3 in Amber) was applied to maintain constant temperature, with a coupling constant of 1.0 (gamma_ln=1.0 in Amber). Next, Langevin dynamics in NPT was performed for 1ns to equilibrium the box density/pressure. A strong pressure coupling of 0.1 ps to the barostat was used in order to quickly equilibrate the box to the final density and pressure. NPT simulation with lowered restraints 10.0 *kcal mol^{-1} rad^{-2}* was run for 1 ns with weaker pressure coupling (0.5 ps). The system was then minimized again with restraints acting on backbone atoms. All backbone C and N atoms were restrained with force constant 10.0 *kcal mol^{-1} rad^{-2}* so that the side chains were free to relax. Three 1ns NPT simulation with lowered restraint force constants of 10.0 *kcal mol^{-1} rad^{-2}*, 1.0 *kcal mol^{-1} rad^{-2}* and restraints 0.1 *kcal mol^{-1} rad^{-2}* on backbone C and N atoms were consecutively performed. Finally, 1ns NPT unrestrained simulation was performed prior to production runs. For implicit solvent simulation, the protocol was same except that no periodicity and pressure coupling were applied, and PME was off during the entire equilibration (ntb=0, igb =8 in AMBER). The exterior dielectric was 78.5.

For production runs, the time step was increased to 4 fs using the hydrogen-mass repartitioning method implemented as described previously[114], and explicit solvent simulations were changed to the NVT ensemble (ntb=1, ntp=0 in Amber).

2.3.10 Cluster analysis

Unless noted otherwise, cluster analysis was performed on the combined trajectories starting from helical and extended conformations. The hierarchical agglomerative (bottom-up) approach was used with average linkage (defined by RMSD of C, N and CA atoms) to generate a maximum of 10 clusters using default settings in Cpptraj[115]. This was performed to divide the trajectory into 10 clusters without setting a threshold on how similar the structures are to each other within a cluster. The representative structures extracted from these clusters were used as initial conformations in independent MD runs to check convergence.

2.3.11 RMSD calculations

Unless otherwise noted, all RMSD calculations were done on backbone C, N and CA atoms via Cpptraj[115]. In all cases, terminal residues and capping groups on termini were neglected.

2.3.12 Helical propensity

Following the Best et al.[29b] protocol of implementing Lifson-Roig model[116] for computing helical propensities, we explored the helical propensities of each amino acid in the context of the sequence Ace-A_4**X**A_4-NH_2 to compare to experimental data[72]. This model measures the equilibrium properties of coil-to-helix transitions: three states are defined: coil, start/end of the helix, and within a helix. Their relative weights are 1, v_i *and* w_i, respectively. The start/end of the helix is defined when residue *i* is in the helical region but either of its two adjacent residues is not in the helical region. The residue within a helix is defined when residue *i* and its two adjacent residues are all in the helical region. Everything else is considered to be random coil within the model. A residue is considered helical if inside the α region using the basin definition in **Table 2.6**. The sensitivity to this definition was tested by calculating helical propensity with a wider range definition (**Table 2.6** and **Figure 2.6**).

Table 2.6 The definition of φ/ψ range for α, wider-α, β and ppII conformations used in this work.

	φ range	ψ range
α	(-90°: -30°)	(-77°: -17°)
wider-α	(-100°: -20°)	(-90°: 0°)
β	(-175°: -115°)	(105°: 165°)
ppII	(-105°: -30°)	(90°: 150°)

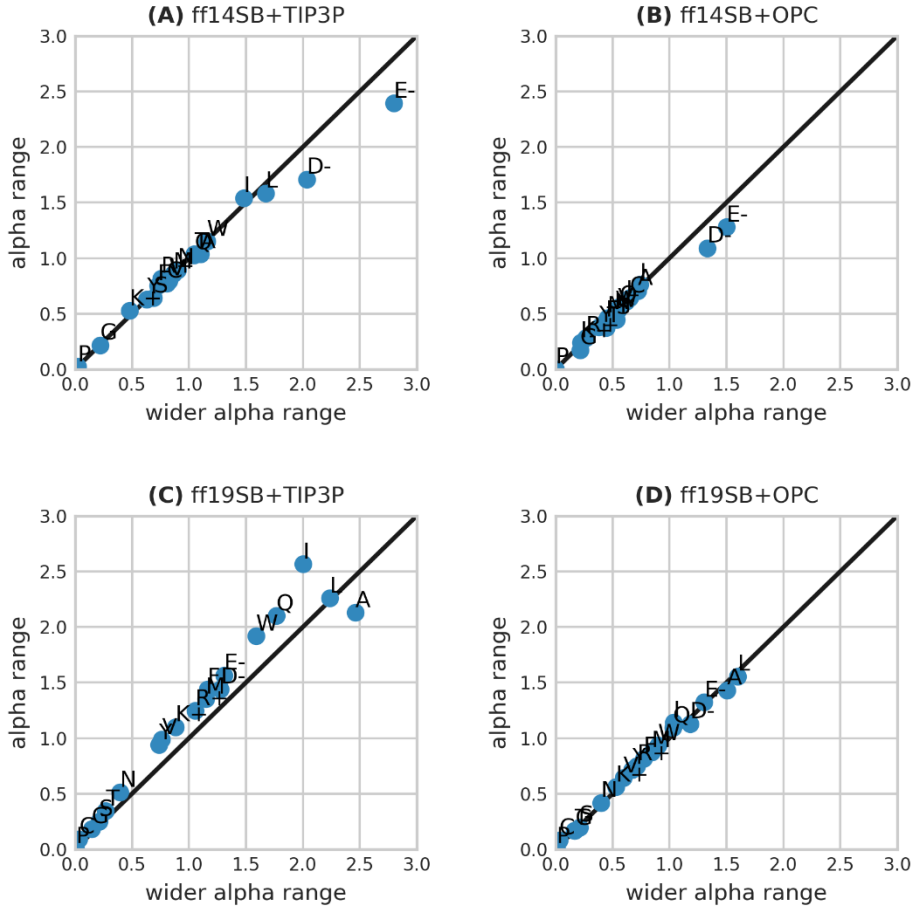


Figure 2.6 Correlation between helical propensity *w* from simulations with wider alpha range and standard alpha range (defined in **Table 2.6**) for (A) ff14SB+TIP3P, (B) ff14SB+OPC, (C) ff19SB+TIP3P and (D) ff19SB+OPC.

Following Best et al.[29b], the partition function for the blocked peptide of length N (N=9) is defined as:

$$Z = (0\ 0\ 1)\prod_{i=1}^{N} M_i\ (0\ 0\ 1)^T, \text{ where } M_i = \begin{vmatrix} w_i & v_i & 0 \\ 0 & 0 & 1 \\ v_i & v_i & 1 \end{vmatrix} \quad (2.3),$$

The log-likelihood that residue *i* will be assigned a helical propensity parameter w_i is given by:

$$ln(L) = \sum_i N_{w,i} ln(w_i) + \sum_i N_{v,i} ln(v_i) - N_k ln(Z) \quad (2.4),$$

where v_i and w_i are the parameters for fitting, N_k is the total number of frames in the simulation, $N_{w,i}$ and $N_{v,i}$ are the total number of times in the simulation that residue *i* is within a helix and start/end of a helix, respectively. More specifically, $N_{w,i}$ is incremented if residue *i* is

within a helix and $N_{v,i}$ is incremented if residue *i* is start/end of a helix. The subscript *i* indicates the amino acid (Ala, Val, Leu, etc). The model parameters (*v* and *w*), and their distributions, were optimized by performing genetic algorithm following the protocol of Perez et al.[71] to maximize the objective function, *ln(L)*, which maximized the likeliness of residue *i* being assigned to specific *v* and *w*. Mutation and crossover moves were performed to change $ln(w_i)$ and $ln(v_i)$, with a rate of 0.3 and 0.7 respectively. A total of 1000 genetic optimization cycles were performed to yield specific *v* and *w* for residue *i*. v_{ala} and w_{ala} were initially evaluated for all Ala in the capped $A_4\underline{\mathbf{A}}A_4$ peptide, then v_i and w_i for **X** were evaluated in capped $A_4\underline{\mathbf{X}}A_4$ peptide with *v* and *w* parameters for Ala being fixed to the values of previously optimized v_{ala} and w_{ala}.

Histidine was excluded because the imidazole protonation state (δ, ε or both) is difficult to assign, and the reported experimental scales for 20 natural amino acids vary the most for His, with it being the least helical from one experimental scale but almost in the middle of the helicity from another [72, 117] For instance, Pace and Scholtz[117] summarized a helical propensity scale based on NMR measurements of helix propensity from 11 systems, including both proteins and short peptides, at different pH values and temperatures. All helical propensities were reported in ΔΔG relative to Ala (0 kcal/mol, the most helical) and normalized by setting Gly=1 kcal/mol, the least helical. In that report, His exhibits a value of 0.61±0.11 (error bar calculated from 13 reported measurements) averaged across systems and protonation states (estimated based on experimental pH). Specifically, for neutral His, the helical propensity is 0.56±0.07 (uncertainty calculated from seven reported measurements), and for the protonated His^+, the helical propensity is 0.66±0.10 (uncertainty calculated from six reported measurements). This value is much lower (closer to Ala, meaning more helical) than several other amino acids including Asn, Thr, Cys and Asp. However, according to the NMR data[72] (reported as helical propensity w instead of ΔΔG), His is the least helical along with Gly (see **Table 2.7**). These NMR data are generally consistent with Pace and Scholtz[117] except for His (**Figure 2.7**). Due to these uncertainties, we decided to remove His from the helical propensity comparisons in **Figure 2.25**. The helical propensity data (from both NMR and MD) of all including His are provided in **Table 2.7** and **Table 2.8**.

Table 2.7 Helical propensities of 20 standard amino acids from NMR experiments[72], ff14SB+TIP3P, ff14SB+OPC, ff19SB+TIP3P and ff19SB+OPC. Error bars for calculated helical propensities were estimated via bootstrapping analysis.

Residue	NMR[72]	ff14SB+TIP3P	ff14SB+OPC	ff19SB+TIP3P	ff19SB+OPC

Ala	1.39±0.01	1.04±0.02	0.71±0.03	2.13±0.03	1.43±0.02
Leu	1.15±0.10	1.58±0.14	0.61±0.07	2.26±0.18	1.56±0.13
Glu^-	1.00±0.07	2.39±0.18	1.28±0.20	1.57±0.16	1.32±0.12
Met	0.90±0.03	0.86±0.05	0.46±0.07	1.35±0.10	0.87±0.05
Ile	0.81±0.03	1.54±0.15	0.76±0.11	2.57±0.25	1.14±0.16
Arg^+	0.80±0.04	0.82±0.07	0.28±0.05	1.25±0.14	0.75±0.07
Gln	0.72±0.03	1.03±0.09	0.56±0.11	2.10±0.16	1.09±0.10
Trp	0.64±0.07	1.15±0.12	0.51±0.09	1.92±0.14	0.93±0.12
Lys^+	0.60±0.04	0.53±0.04	0.24±0.04	1.10±0.10	0.56±0.04
Asp^-	0.54±0.03	1.71±0.11	1.09±0.11	1.44±0.12	1.13±0.08
Val	0.53±0.03	0.79±0.06	0.53±0.07	0.99±0.07	0.64±0.06
Ser	0.51±0.03	0.64±0.05	0.44±0.06	0.25±0.03	0.20±0.02
Phe	0.50±0.04	0.75±0.06	0.42±0.07	1.44±0.14	0.82±0.07
Tyr	0.50±0.04	0.63±0.04	0.38±0.09	0.94±0.10	0.72±0.08
Asn	0.43±0.04	0.89±0.07	0.49±0.08	0.51±0.04	0.42±0.04
Thr	0.40±0.02	1.04±0.11	0.37±0.06	0.34±0.03	0.17±0.02
Cys	0.36±0.01	0.77±0.05	0.64±0.10	0.09±0.01	0.08±0.01
Gly	0.22±0.02	0.21±0.01	0.17±0.02	0.18±0.01	0.17±0.02
Pro	0.05±0.01	0.02±0.02	0.01±0.01	0.01±0.01	0.01±0.01
His^δ	0.20±0.02	0.72±0.05	0.31±0.05	1.20±0.08	0.74±0.07
His^ε	0.20±0.02	1.15±0.08	0.94±0.10	1.01±0.08	1.04±0.12
His^+	0.20±0.02	0.26±0.04	0.11±0.02	0.81±0.09	0.33±0.05

Table 2.8 Helical propensities of 12 amino acids for NMR experiment[72], ff14SB+TIP4P-Ew, ff19SB+TIP4P-Ew, ff19SB+OPC3, ff15ipq+SPC/Eb and fb15+fb3. Error bars for calculated helical propensities were estimated via bootstrapping analysis.

	NMR[72]	ff14SB+TIP4P-Ew	ff19SB+TIP4P-Ew	ff19SB+OPC3	ff15ipq+SPC/E_b	fb15+fb3
Ala	1.39±0.00	0.84±0.01	1.58±0.01	1.94±0.01	1.25±0.01	0.81±0.01
Leu	1.15±0.10	1.15±0.10	1.97±0.10	1.79±0.10	0.70±0.10	1.27±0.10
Glu^-	1.00±0.07	1.53±0.07	1.42±0.06	1.70±0.07	0.20±0.07	1.65±0.07

Ile	0.81±0.03	1.20±0.03	1.82±0.03	2.13±0.03	0.35±0.03	0.63±0.03
Arg^+	0.80±0.04	0.42±0.04	0.73±0.04	0.94±0.04	0.16±0.04	0.76±0.04
Gln	0.72±0.03	0.55±0.03	1.49±0.03	1.85±0.02	0.48±0.03	1.60±0.03
Trp	0.64±0.07	0.83±0.07	1.15±0.07	1.38±0.06	0.66±0.07	0.39±0.07
Lys^+	0.60±0.04	0.12±0.04	0.75±0.04	1.08±0.04	0.10±0.04	0.39±0.04
Val	0.53±0.03	0.45±0.03	0.93±0.03	0.92±0.02	0.16±0.03	0.82±0.03
Phe	0.50±0.04	0.61±0.04	1.20±0.04	1.35±0.04	0.35±0.04	0.64±0.04
Asn	0.43±0.04	0.46±0.04	0.39±0.04	0.57±0.04	0.30±0.04	0.43±0.04
Gly	0.22±0.02	0.23±0.02	0.16±0.03	0.17±0.01	0.64±0.02	0.37±0.02

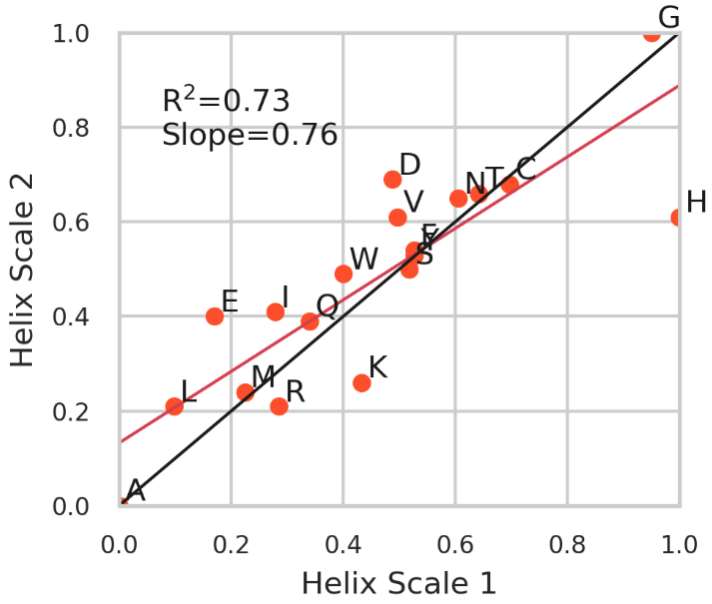


Figure 2.7 Correlation of helical propensity between two experimental data sets. Helix scale 1[72] was reported in helical propensity parameter w and helix scale 2[117] was reported in ΔΔG (kcal/mol) relative to Ala (Ala=0 kcal/mol and Gly=1kcal/mol). The helical propensity parameter data were further converted by applying -RTln(w) and normalized here by forcing Ala to be zero and Gly to be one so that we can have a consistent comparison between two scales. Orange dots represent the normalized values. Amino acids are represented with one letter codes. Linear regression (red line) was performed against the data points, with R^2 and slope quantifying the goodness of fit.

2.3.13 Bootstrapping analysis on helical propensity

In order to quantify the uncertainty of the computed *w*, bootstrapping analysis was performed for each system. When the sample size is insufficient for straightforward statistical

inference, bootstrapping provides a way to account for the distortions caused by a specific sample that may not be fully representative of the population. First, a combined trajectory from 12 independent runs of each $A_4\underline{\mathbf{X}}A_4$ (3.2 μs for each run) was used to fit the *v* and *w* for that **X**. Second, the combined trajectory was split into 10 segments with same length. Third, 50 times of resampling with replacement were done on the 10 sub-trajectories. This step generated 50 trajectories with the same length of the initially combined one (3.2 μs * 12) but with some segments repeated. 50x resampling has been suggested to lead to reasonable standard error estimates[118]. Lastly, we fit the *v* and *w* parameters with each of the 50 trajectories respectively and calculated the standard deviation of the 50 resulting w values for each amino acid. According to the distribution of the sampled *w* parameters (**Figure 2.8**), all amino acids have a high peak and a narrow range on *w* which suggests good quality sampling and precise estimates of helical propensity.

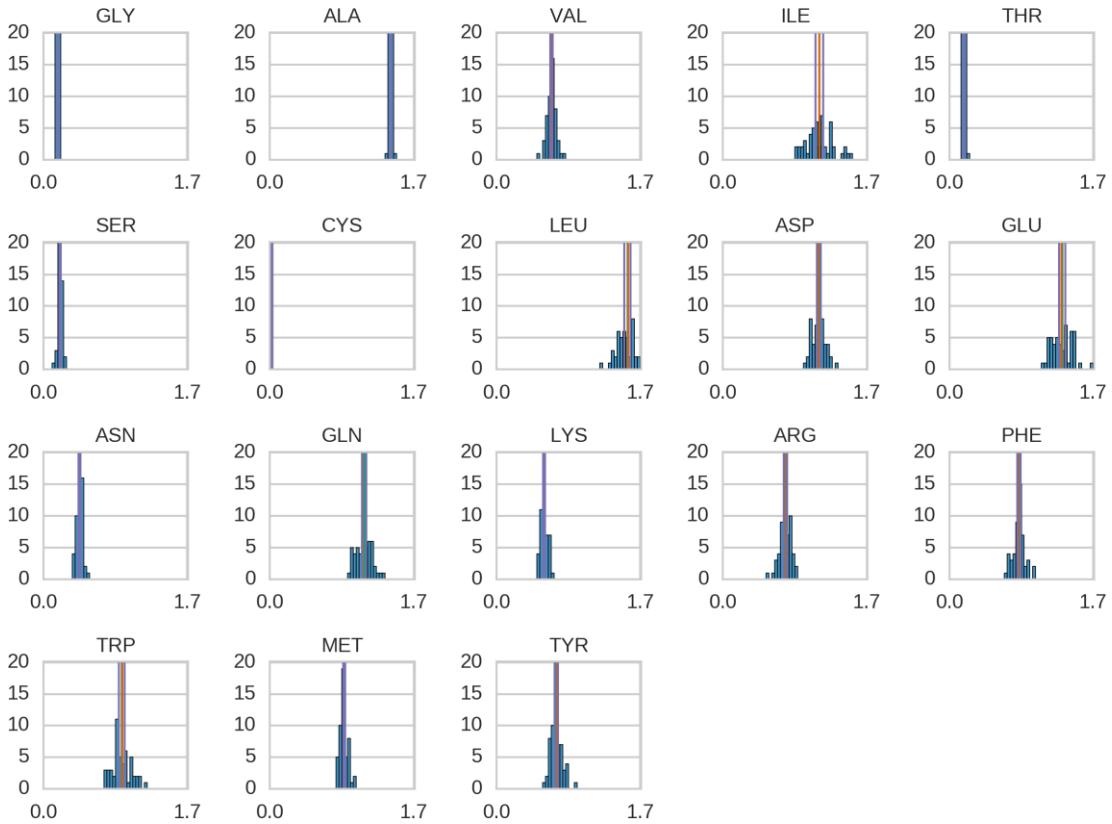


Figure 2.8 Distribution of sampled helical propensity *w* from bootstrapping for all amino acids in ff19SB+OPC.

2.3.14 NMR scalar coupling calculations

Following Best et al.[29a] and our previous work[26a], scalar couplings were calculated from simulations using Karplus relations[119] and the "Orig parameters"[120] also adopted by Graf et al.[28]. To quantify the discrepancy between experimental scalar couplings and those calculated from simulation, χ^2 error was defined in Best et al.'s work[29a] and also here as:

$$\chi^2 = N^{-1} \sum_{j=1}^{N} (\langle J_j \rangle_{sim} - J_{j,exp})^2 / \sigma_j^2 \quad (2.5),$$

where N is the total number of scalar coupling types, $<J_j>_{sim}$ is the averaged scalar coupling from the simulation for scalar coupling type j. $J_{j,exp}$ is the NMR data for type j. σ_j is the estimated systematic error of the Karplus equation for type j adopted by both Best et al.'s work[29a] and our previous work[26a]. Precision of χ^2 is estimated as half the difference of χ^2 calculated from two simulations starting from either helical or extended conformation. For dipeptides, **Table 2.9** lists $^3J_{HNHA}$ data and an estimated systematic error of 0.91 was used in the χ^2 calculation[26a, 29a]. For Ala_5, **Table 2.10** lists all scalar coupling types and the corresponding systematic errors[2c]. Since the NMR data[121] were measured at pH=4.9, side chains for Arg, Lys and His were modeled in protonated state. For Glu and Asp, both deprotonated and protonated states were simulated, and the error was reported as a weighted average value. Constant pH simulations were performed to obtain the appropriate ratio of protonated state versus deprotonated state respectively.

Table 2.9 NMR *$^3J(HNHA)$* values and calculated *$^3J(HNHA)$* values from MD simulation (with error bars calculated from independent runs) for 19 dipeptides.

Residue	*$^3J(HNHA)$* (NMR)	*$^3J(HNHA)$* (ff14SB+TIP3P)	*$^3J(HNHA)$* (ff14SB+OPC)	*$^3J(HNHA)$* (ff19SB+TIP3P)	*$^3J(HNHA)$* (ff19SB+OPC)
Gly	5.85	6.41±0.01	6.35±0.01	6.07±0.01	6.01±0.01
Ala	6.06	6.25±0.03	6.17±0.01	5.91±0.01	5.81±0.01
Val	7.30	6.81±0.01	6.62±0.01	6.92±0.01	6.79±0.02
Thr	7.35	6.69±0.01	6.71±0.03	7.11±0.02	7.15±0.04
Ile	7.33	6.49±0.01	6.35±0.02	6.66±0.01	6.52±0.04
Ser	7.02	6.56±0.02	6.52±0.01	6.29±0.01	6.22±0.02
Cys	7.31	6.30±0.02	6.23±0.02	6.43±0.01	6.44±0.01
Leu	6.88	6.65±0.02	6.59±0.01	6.63±0.01	6.52±0.01
Phe	7.18	6.65±0.01	6.55±0.03	6.46±0.02	6.40±0.02

Trp	6.91	6.71±0.01	6.55±0.01	6.50±0.02	6.54±0.01
Tyr	7.13	6.58±0.03	6.49±0.08	6.42±0.01	6.35±0.02
Met	7.02	6.59±0.01	6.48±0.02	6.50±0.01	6.41±0.01
Asp	6.93	6.68±0.02	6.62±0.01	7.55±0.01	7.43±0.01
Asn	7.45	6.55±0.01	6.56±0.01	6.80±0.01	6.85±0.01
Glu	6.63	6.65±0.01	6.58±0.02	6.10±0.03	5.98±0.02
Gln	7.14	6.60±0.02	6.59±0.02	6.46±0.01	6.44±0.04
Arg^+	6.85	6.62±0.03	6.57±0.02	6.46±0.01	6.48±0.02
Lys^+	6.83	6.44±0.01	6.37±0.01	6.54±0.01	6.55±0.01
His^+	7.89	6.63±0.05	6.71±0.00	6.55±0.03	6.76±0.04

Table 2.10 Scalar coupling type, NMR measurements, the calculated scalar couplings with different force field + solvent model (with error bars), and the systematic error[26a, 29a] of Karplus equation/"Orig parameters" for Ala_5 tetrapeptide.

Scalar coupling	NMR	ff14SB+ TIP3P	ff14SB+ OPC	ff19SB+ TIP3P	ff19SB+ OPC	Systematic error (σ)
$^3J(HNHA)$	5.74	5.97±0.04	5.78±0.01	5.52±0.00	5.33±0.03	0.91
$^3J(HAC)$	1.86	1.95±0.19	1.66±0.08	2.01±0.01	1.93±0.05	0.38
$^3J(HACB)$	2.24	1.91±0.04	1.98±0.01	1.91±0.01	1.94±0.01	0.39
$^1J(NCA)$	11.26	11.11±0.05	11.35±0.02	10.75±0.02	10.86±0.01	0.59
$^2J(NCA)$	8.55	8.30±0.01	8.45±0.03	8.15±0.01	8.33±0.01	0.50
$^3J(HNCA)$	0.68	0.49±0.01	0.46±0.01	0.53±0.00	0.50±0.01	0.10

2.3.15 Constant pH simulation

Constant pH simulations of 800 ns with TIP3P[89] and OPC[66] explicit solvent were performed on the capped dipeptide forms for the titratable residues Glu and Asp. Initially, these titratable residues were assigned to be protonated, and the state change was attempted every 100 MD steps through Monte Carlo approach using a Generalized Born implicit solvent model (igb=2)[122] which was the model used to parameterize the reference compounds in constant pH simulation[123]. Following published protocol, the intrinsic Born radii of carboxylate oxygen atoms were shrunk in order to reduce artifacts arising from including all four alternate hydrogen atom

positions in the GB descreening calculation.[123] 200 steps of solvent relaxation dynamics (in which the solute was held fixed) were performed before resuming simulation if any protonation states were changed[123]. The solvent pH value was set to 4.9 in agreement with the NMR experiment[121]. The rest of the input was retained from the standard MD protocol described above.

These constant pH simulations have limitations, such as using an older GB model[122] (igb=2) for reference compound energy, and neglect of updating dihedral parameters when protonation state switches[123]. Therefore, the constant pH simulations were only used to estimate the percentage of protonation states for titratable residues, and the sampled ensembles were not used directly for χ^2 analysis. The χ^2 analysis was performed on the combination of protonated and deprotonated trajectories in explicit water, weighted by the ratio of protonated state versus deprotonated state.

2.3.16 NMR order parameters

The ability of a force field to model local dynamics accurately in well-folded proteins in solution was examined by comparing to NMR experimental backbone NH S^2 order parameters for GB3[98], ubiquitin[99] and lysozyme[100]. We adopted the model-free approach originally proposed by Lipari and Szabo[124] and used iRED[125] as implemented in Cpptraj. iRED is based on a covariance matrix analysis of inter-nuclear vector orientations, represented by spherical harmonics, extracted from MD simulations. For this analysis, we averaged iRED results calculated for windows of length five times the tumbling correlation time (τ_c), which was suggested to best reproduce the model-free S^2 order parameters[126]. Thus, window sizes of 2 ns, 4 ns and 8 ns were used for GB3, ubiquitin and lysozyme respectively, in agreement with previous work[30, 127]. The uncertainties in the computed S^2 were calculated by taking the standard deviation from four independent MD runs.

2.3.17 Statistical analysis of PDB data

To compare the φ/ψ distributions from simulation against PDB data, we used Lovell's rotamer library[74, 128] of 7957 high-resolution, quality-filtered protein chains to generate the PDB-based φ/ψ distributions. Two filters were applied to select a portion of the original 7957 structures. Firstly, only residues in coil and turn as defined by DSSP[129] were selected. Secondly, these residues

were eliminated if any of the backbone heavy atoms had B factors larger than 30. Biopython[130] was used to apply the two filters against 7957 PDB files.

2.3.18 Average relative energy error (REE) calculation

Unless otherwise noted, average REE between two sets of energies were calculated as following:

$$average\ REE = \frac{2}{N*(N-1)}\sum_{i}^{N-1}\sum_{j>i}^{N}\left|\left(E_i^a - E_j^a\right) - \left(E_i^b - E_j^b\right)\right| \quad (2.6),$$

where N is the number of conformations. E_i^a and E_j^a are energies calculated in method "a" (QM, MM, etc) of conformation i and j. E_i^b and E_j^b are energies calculated in method "b" (QM, MM, etc) of conformation i and j.

2.4 Results and Discussion

2.4.1 Backbone rotational energies in ff19SB compared to ff14SB

Alanine and Glycine energetics. Backbone φ/ψ rotational energy profiles were analyzed for QM, ff19SB, ff14SB and CMAP (derived by subtracting ff14SB00 from QM energies on the 2D grid, see **Methods**). Ala and Gly are discussed first because they are the simplest with no significant side chain degrees of freedom. We performed 2D backbone rotation scans for the capped Ala and Gly dipeptides, followed by restrained minimization and energy evaluation with implicit solvent for QM and MM. The CMAPs were derived by subtracting MM from QM energies on the 2D grid. The ff19SB energies were obtained by adding the CMAP-based bicubic function to ff14SB00 (see **Methods: Molecular mechanics (MM) optimization and energy calculations** and **CMAP fitting**). As shown in **Figure 2.9**, the ff19SB energy profiles are nearly identical to the QM reference data, which was anticipated based on the training method. However, significant differences between ff14SB and QM are apparent. In ff14SB, the overall energy profiles are highly

symmetric with little φ/ψ coupling, likely due to the lack of coupling between the ff14SB dihedral correction parameters. This coupling may arise from polarization changes as the amide dipoles become aligned in the helical conformation. The shape and location (the bin having lowest energy in the basin defined in **Table 2.11**) of the α basins from QM are poorly reproduced by ff14SB for both Ala and Gly. Importantly, the diagonal shape of the left- and right-handed α helical basins as observed in QM and ff19SB is poorly reproduced in ff14SB, which instead samples too deeply into negative φ for $\psi < 0$. In addition, for Ala, the C_7^{eq} local minimum between ppII and α_R in QM (**Figure 2.9**) is absent in ff14SB, but reproduced with ff19SB. For Gly, the QM energy barrier at φ = -120 / ψ = -60 is more accurate with ff19SB (**Figure 2.10**).

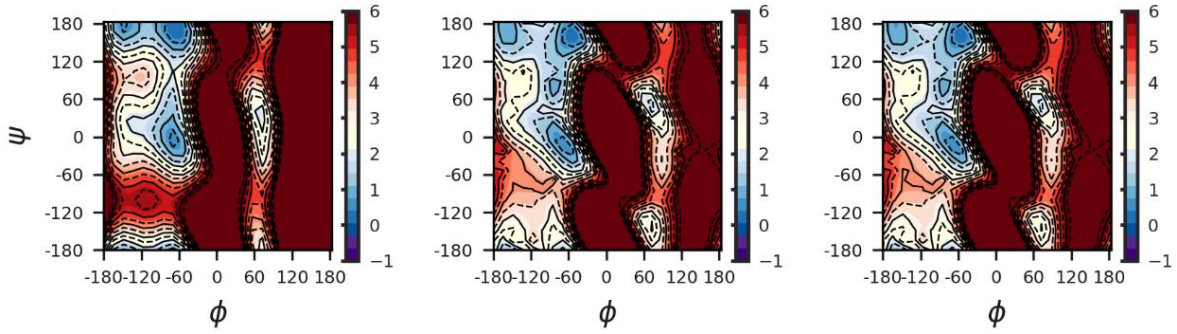


Figure 2.9 Ala dipeptide Ramachandran energy (kcal/mol) surfaces calculated in (left) ff14SB+GBSA, (middle) QM+SMD and (right) ff19SB+GBSA. All energies were zeroed relative to the lowest energy in the ppII region (defined in **Table 2.6**). The values above the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

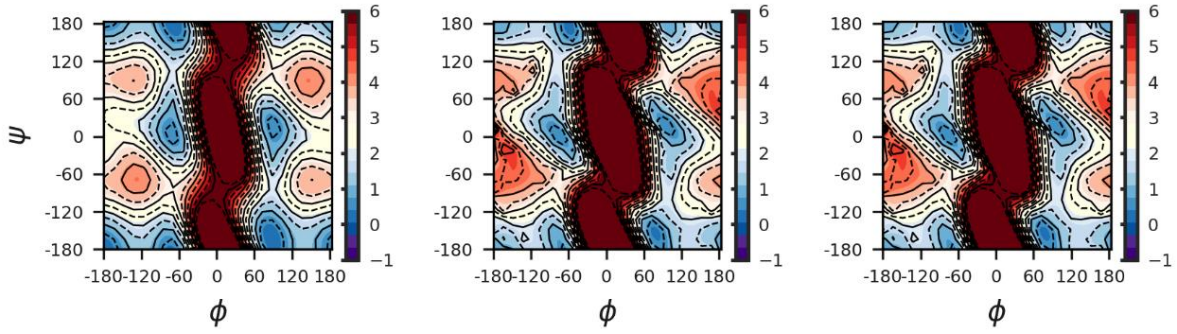


Figure 2.10 Gly dipeptide Ramachandran energy (kcal/mol) surfaces calculated in (left) ff14SB+GBSA, (middle) QM+SMD and (right) ff19SB+GBSA. All energies were zeroed relative

to the lowest energy at ppII region (defined in **Table 2.6**). The values above the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

Table 2.11 The location (φ, ψ) of α basins (α_R and α_L) for Ala and Gly dipeptide in QM+SMD, ff14SB+GBSA and ff19SB+GBSA.

	ff14SB+GBSA	QM+SMD	ff19SB+GBSA
Ala	(-75,-15)/(60,15)	(-75,-15))/(60,45)	(-75,-15))/(60,45)
Gly	(-75,-15)/(75,0)	(-75,-15))/(75,15)	(-75,-15))/(75,15)

Overall, the deviation of ff14SB from QM for Ala and Gly is notable despite the use of QM data for multiple conformations of Ala_3 and Gly_3 during training of ff14SB/ff99SB backbone parameters. This relative weakness in ff99SB/ff14SB is likely a result of the use of only gas-phase energy minima for training (thus lacking the compulsion to reproduce the entire basin shape, or even the locations of aqueous-phase minima), along with dihedral correction terms that lack φ/ψ coupling, resulting in an overly symmetric energy map. Use of the CMAP approach for ff19SB results in improved reproduction of the overall energy surfaces for both amino acids.

We tested the impact of using QM in gas-phase as the target data. We fit an Ala dipeptide CMAP (same protocol as in **CMAP fitting**) against the entire surface of gas-phase QM energy instead of in solution QM (**Figure S7A**), and ff14SB00 was used as the MM model for CMAP fitting. The resulting energy surface, applied in solution (**Figure S7C**), has an unusual shape of the α_R basin (extending much farther into ϕ<-120°) and the α_L energy basin is unexpectedly deep. We conclude that fitting CMAPS using solution QM & MM calculations is important for good results here.

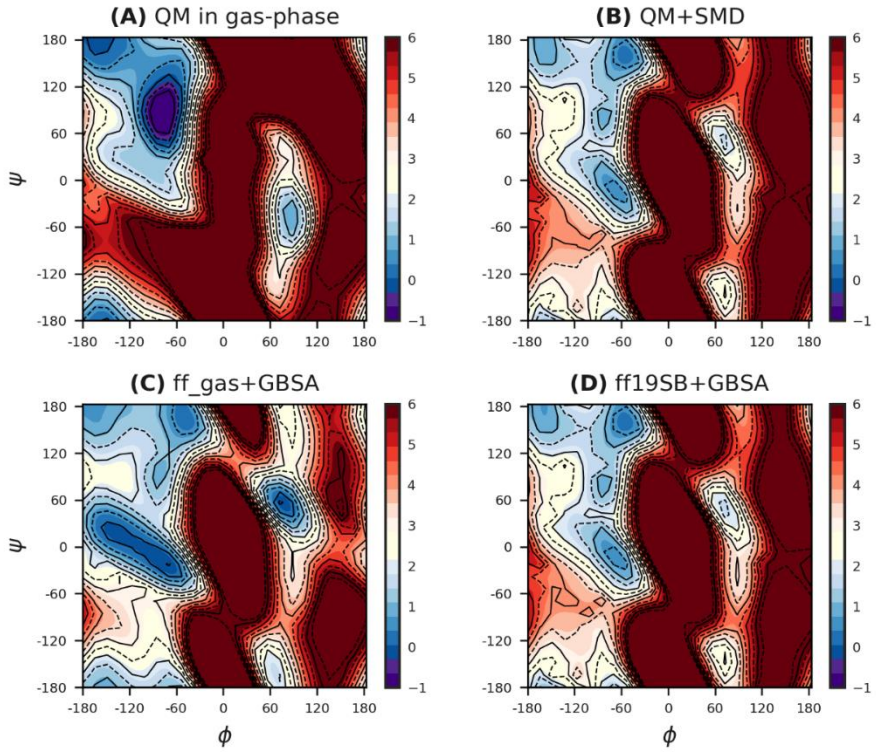


Figure 2.11 Ala dipeptide Ramachandran energy (kcal/mol) surfaces calculated in (A) QM in gas-phase, (B) QM in SMD, (C) ff_gas+GBSA, and (D) ff19SB+GBSA. QM in gas-phase was calculated using the same QM method as in ff19SB training, but excluding SMD solvation. The ff_gas model was derived by following the CMAP fitting protocol but using gas-phase QM as reference data and ff14SB00 as MM in fitting CMAPgas. CMAPgas was trained by subtracting ff14SB00 from QM in gas-phase. All energies were zeroed referenced to the lowest energy at ppII region (defined in **Table 2.6**). The values beyond color bar range are depicted in dark red. Solid contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies.

Amino acids with multiple side chain rotamers. The 2D CMAP training provides a "perfect" fit against the 2D reference QM data for Ala and Gly since no other significant rotational degrees of freedom are present. However, all other amino acids have longer side chains with additional degrees of freedom, and the situation becomes more complex since the energies (and their errors) depend on rotational degrees of freedom not sampled explicitly in the CMAP. While 3D fitting might accommodate some amino acids such as Val or Ser, this rapidly becomes intractable. We first compare alanine and valine using the valine rotamer used in training, then evaluate the transferability of the Val CMAP to alternate Val rotamers.

Our strategy to improving rotamer dependence extends the approach to improving transferability in the side chain parameters we used when developing ff14SB, where we assumed that the largest contribution to poor transferability of dihedral parameters arises from including structures in the training set that expose inaccuracies in the MM short-range nonbonded model that depend on degrees of freedom outside those being trained. Therefore, rotamer dependency was addressed here by initializing all structures on each CMAP training ϕ/ψ grid at the same rotamer conformation, then locally relaxing the side chain conformations to relieve any backbone:sidechain steric clashes that were likely to be inaccurately modeled in MM. If corrections for training set structures with inaccurate backbone:rotamer MM energies were to be incorporated into the backbone parameter for that φ/ψ grid point, the CMAP would have poor transferability to structures with the same φ/ψ values but with alternate rotamers that lack these inaccurately modeled interactions (see **Methods: Molecular mechanics (MM) optimization and energy calculations**).

Comparison of Alanine and Valine Energy Surfaces. For Val, we selected the *trans* rotamer for CMAP training (**Figure 2.12 first row**). As shown in **Figure 2.12B** and **Figure 2.12B**, the QM profiles are qualitatively different between Ala and Val. Val prefers a flatter β/ppII transition region with a U-shape, while Ala has a higher barrier, a stronger preference of ppII over β, and a lower transition barrier between αR and ppII. The C_7^{eq} local minimum between ppII and αR observed in Ala is absent in Val. In addition, the elongated diagonal shape of the αR and αL basins in Ala (indicating strong φ/ψ coupling) is quite different from the narrow circular minimum in Val. The energy minimum at φ = 60 and ψ -150 in Ala is shifted upwards at φ = 70 and ψ = -60 in Val. Importantly, these differences in the Ala/Val QM surfaces are reproduced poorly in ff14SB where the Ala and Val surfaces are generally too similar; both Ala and Val prefer ppII over β and have similar symmetric α_R/α_L basins (**Figure 2.9A** vs. **Figure 2.12A**).

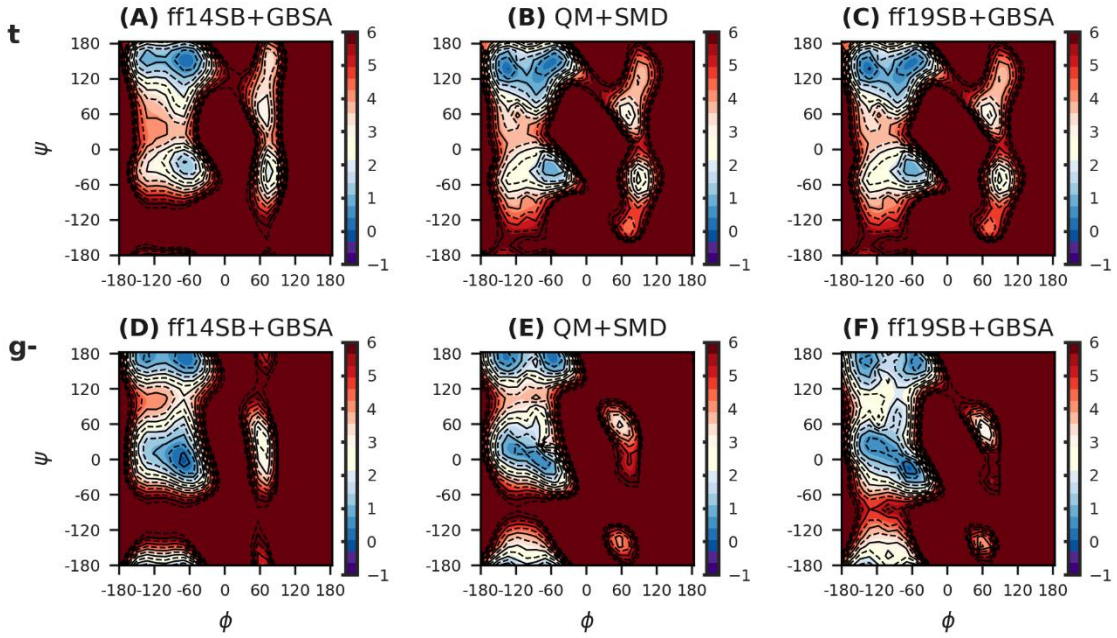


Figure 2.12 Val dipeptide Ramachandran energy surfaces using the *trans* (t) rotamer, calculated in (A) ff14SB+GBSA, (B) QM+SMD and (C) ff19SB+GBSA, and using the *gauche(-)* (g-) rotamer, calculated in (D) ff14SB+GBSA, (E) QM+SMD and (F) ff19SB+GBSA. The *trans* rotamer was used for ff19SB training. All energies were zeroed relative to the lowest energy at ppII region (**Table 2.6**). The values beyond the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol and dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

Transferability of ff19SB backbone parameters to different side chain rotamers

We tested the ability of our approach to provide reasonable transferability of CMAPs between alternate rotamers using valine, for which the side chain rotamer is known to significantly influence backbone populations[53, 74, 131]. We switched the Val rotamer from *trans* to *gauche(-),* calculating QM and MM φ/ψ energies for *gauche(-)* conformations, but keeping the Ala-based ff14SB and *trans*-based Val ff19SB MM parameters (**Figure 2.12, bottom row**). Even though ff19SB was fit using the *trans* rotamer, it reasonably reproduces the changes in the Val QM data from *trans* to *gauche(-).* For example, moving from *trans* to *gauche(-),* the α basins become more diagonal, α_L extends farther into the upper left quadrant, the barrier between ppII and β increases, and the minimum at (90°, -60°) disappears. As seen with the Ala/Val comparison, ff14SB poorly reproduces each of these changes, and the overall energy profiles are generally much too similar between the two rotamers, inconsistent with the QM results. Even though the α basin is stabilized

more and becomes wider from *trans* to *gauche(-)* for ff14SB, the energy profiles are still highly symmetric in both rotamers and the notable difference in the shape of α basins reflected by QM and ff19SB is poorly reproduced in ff14SB, along with a too-flat barrier between ppII and β. Furthermore, rather than the disappearance of the (90°, -60°) minimum as seen in QM and ff19SB, the two minima with positive φ values merge into a single minimum in the wrong location with ff14SB. Thus even though ff19SB was trained using a single rotamer for Val, it does a better job than ff14SB at reproducing the rotamer-dependent backbone profiles from the QM calculations. The results also demonstrate that the high quality match between QM and ff19SB is not simply the result of empirical fitting to an energy map with a single rotamer, but that the accurate reproduction of the QM profiles is maintained even when the map is qualitatively different for an alternate rotamer. To quantify the changes, we calculated average REE (see **2.3.18 Average relative energy error (REE) calculation**) between QM and MM for *trans* and *gauche(-)* as a function of QM energy range above the minimum (**Figure 2.16**). For structures having QM energy within 7 kcal/mol above the minimum, the average REE for the training rotamer *trans* are 1.78 kcal/mol and 0.03 kcal/mol for ff14SB and ff19SB respectively. The average REE for the test rotamer *gauche(-)* are 1.39 kcal/mol and 0.89 kcal/mol for ff14SB and ff19SB. Reasonable transferability is observed for other amino acids as well; examples include Ser and Glu. For Ser (**Figure 2.13**), ff19SB was trained against *gauche(+)*, but is able to reproduce reasonable QM surfaces for both *gauche(+)* and *gauche(-)*, such as the diagonal shape of α_R and α_L basin for both rotamers and the local minimum between ppII and α_R for *gauche(-)*. For structures having QM energy within 7 kcal/mol above the minimum, the average REE for *gauche(+)* are 1.80 kcal/mol and 0.06 kcal/mol for ff14SB and ff19SB. The average REE for *gauche(-)* are 1.98 kcal/mol and 1.01 kcal/mol for ff14SB and ff19SB. For Glu (**Figure 2.14**), ff19SB was trained against rotamer mt-10 (using naming conventions from literature[74]) (*gauche(-)* for χ1, *trans* for χ2 and -10° for χ3) and reproduces reasonably the QM surfaces for both mt-10 and tt0 (*trans* for χ1, *trans* for χ2 and 0° for χ3). In contrast, ff14SB merges the two minima into one at $\phi = 60°$ for mt-10, and poorly reproduces the barrier height at $\phi = -120°$ and $\psi > 30°$ for tt-0[74]. For structures having QM energy within 7 kcal/mol above the minimum, the average REE for mt-10 are 2.05 kcal/mol and 0.08 kcal/mol for ff14SB and ff19SB. The average REE for tt10 are 1.82 kcal/mol and 0.72 kcal/mol for ff14SB and ff19SB.

The QM, ff14SB and ff19SB energy maps for all 16 amino acid dipeptides in the training set are shown in **Figure 2.15**.

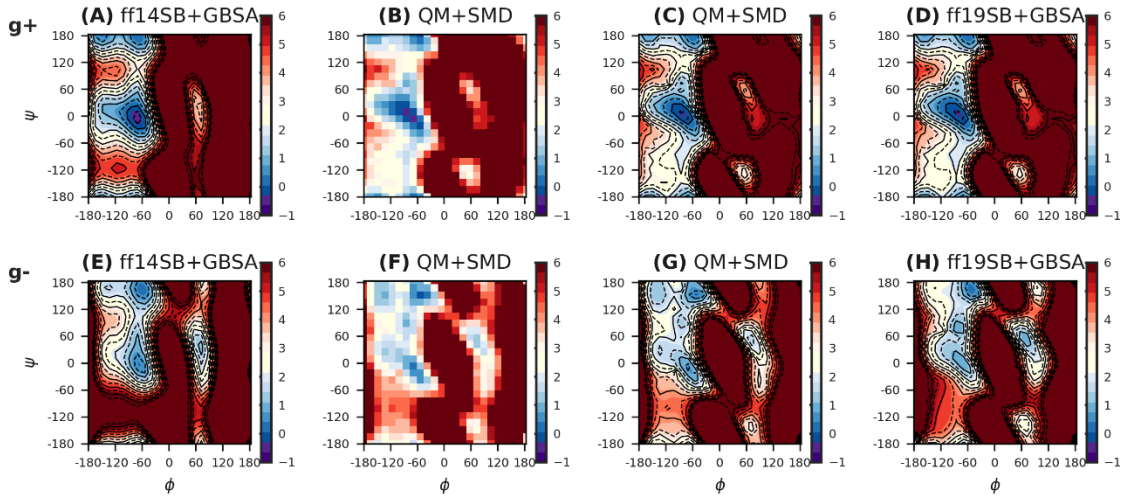


Figure 2.13. Ser dipeptide Ramachandran energy surfaces on *gauche(+)* rotamer calculated in (A) ff14SB+GBSA, (B) QM+SMD with no interpolation, (C) QM+SMD with bicubic interpolation and (D) ff19SB+GBSA, and on *gauche(-)* rotamer calculated in (E) ff14SB+GBSA, (F) QM+SMD with no interpolation, (G) QM+SMD with bicubic interpolation and (H) ff19SB+GBSA. All energies were zeroed referenced to the lowest energy at ppII region (**Table 2.6**). The values beyond color bar range are depicted in dark red. Solid contours indicate integer energy values in kcal/mol and dashed contours indicate half integer energies.

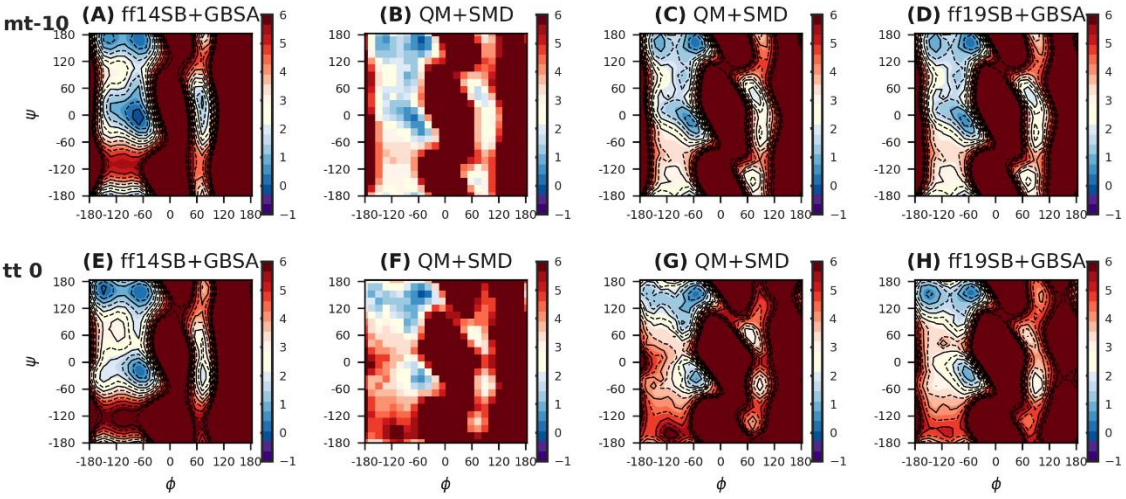
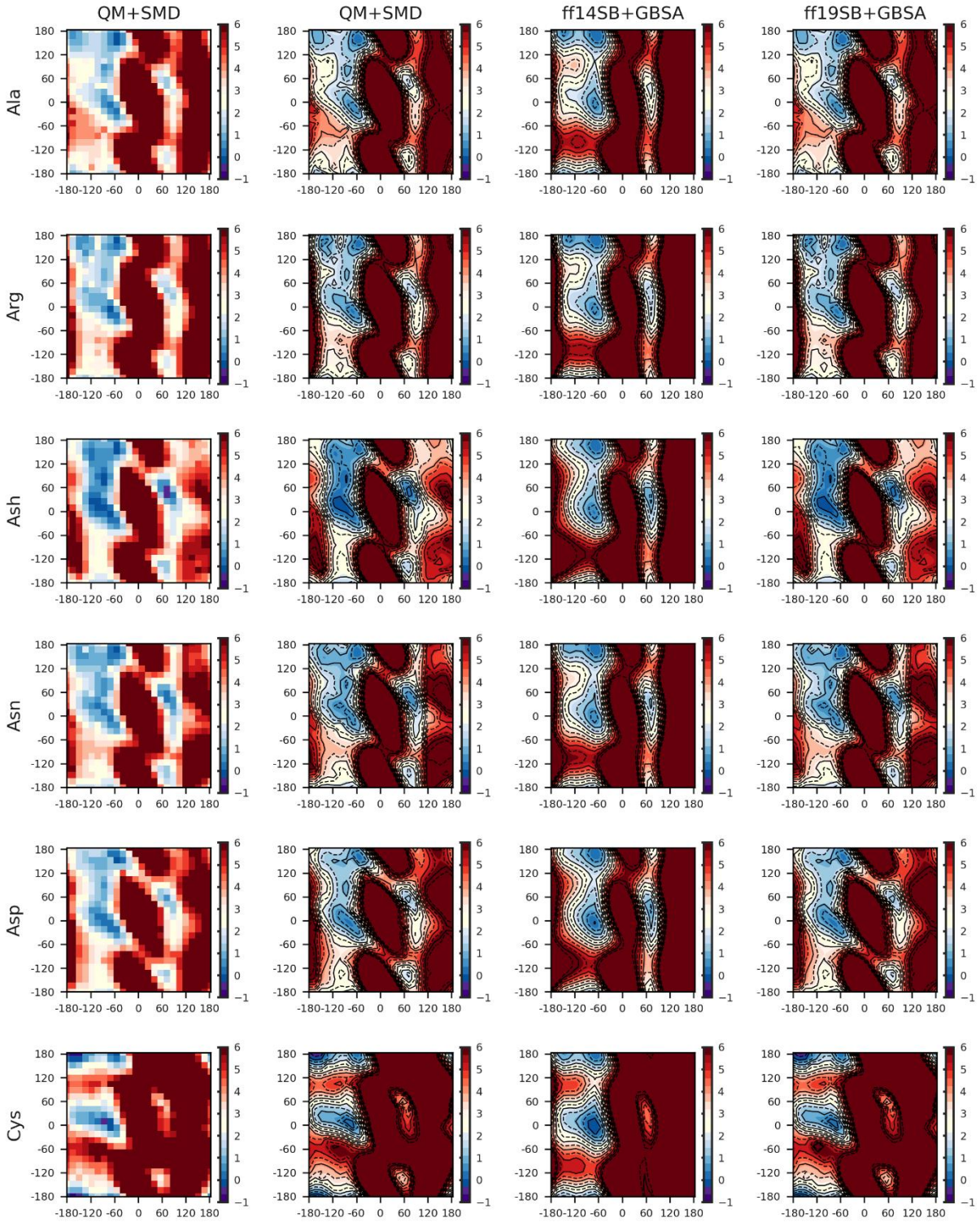
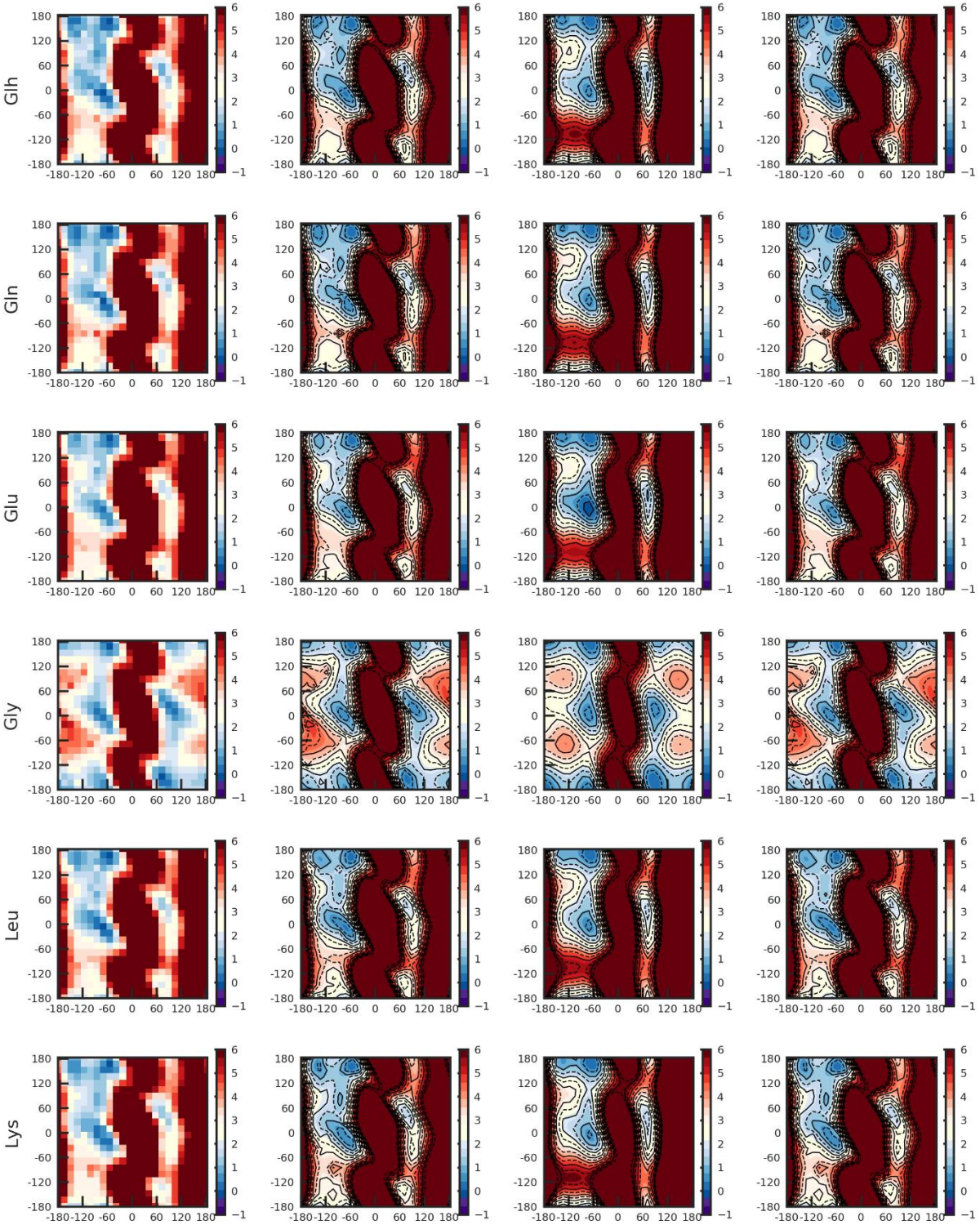


Figure 2.14. Glu dipeptide Ramachandran energy surfaces on mt-10 rotamer calculated in (A) ff14SB+GBSA, (B) QM+SMD with no interpolation, (C) QM+SMD with bicubic interpolation and (D) ff19SB+GBSA, and on tt 0 rotamer calculated in (E) ff14SB+GBSA, (F) QM+SMD with

no interpolation, (G) QM+SMD with bicubic interpolation and (H) ff19SB+GBSA. All energies were zeroed referenced to the lowest energy at ppII region (**Table 2.6**). The values beyond color bar range are depicted in dark red. Solid contours indicate integer energy values in kcal/mol and dashed contours indicate half integer energies.


QM+SMD
QM+SMD
ff14SB+GBSA
ff19SB+GBSA
Ala
Arg
Ash
Asn
Asp
Cys
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180
6
5
4
3
2
1
0
−1


Glh
Gln
Glu
Gly
Leu
Lys
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180
6
5
4
3
2
1
0
−1

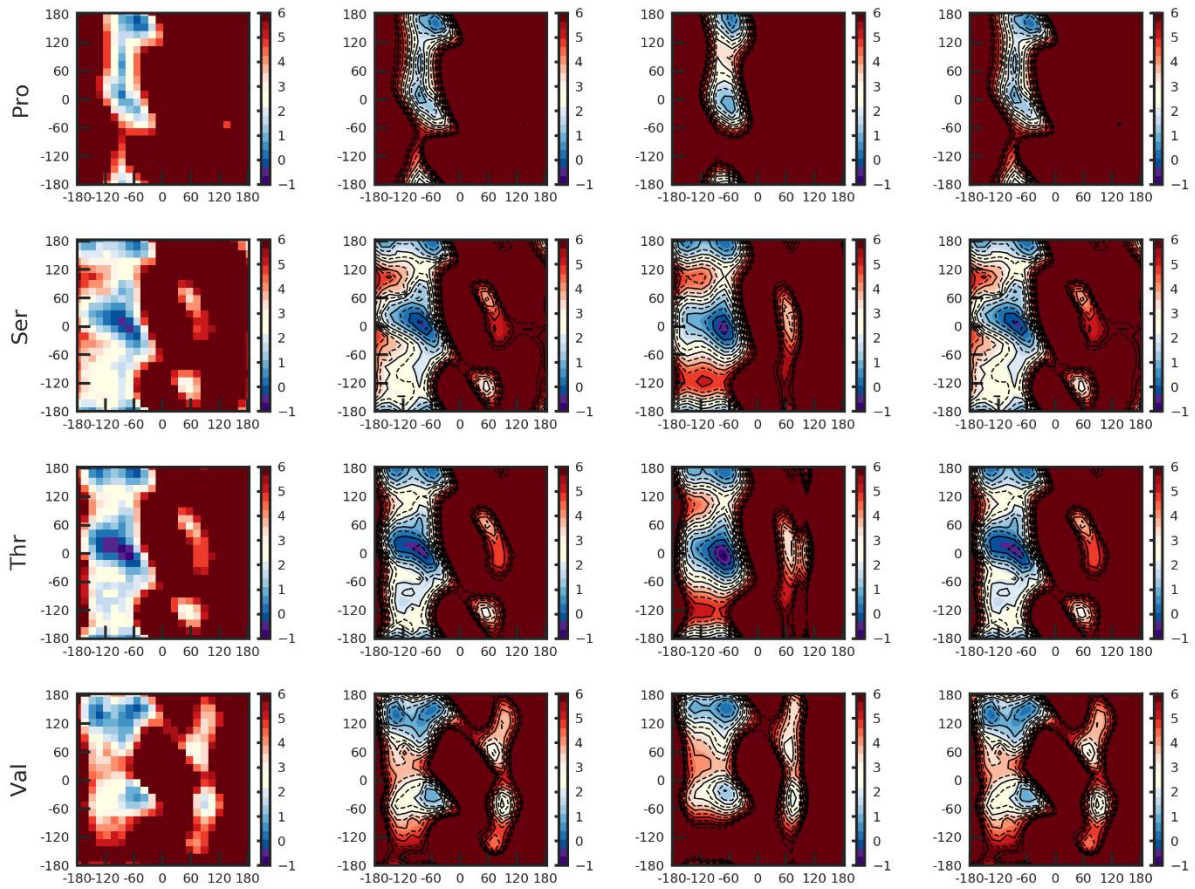


Figure 2.15. Ramachandran energy surfaces calculated for 16 training dipeptides, where the X and Y axes of each plot are φ and ψ, respectively. QM energy surfaces with no interpolation are shown in the 1st column; QM energy surfaces with bicubic spline interpolation implemented in Python are shown in the 2nd column; ff14SB+GBSA energy surfaces are shown in the 3rd column; ff19SB+GBSA energy surfaces are shown in the 4th column.

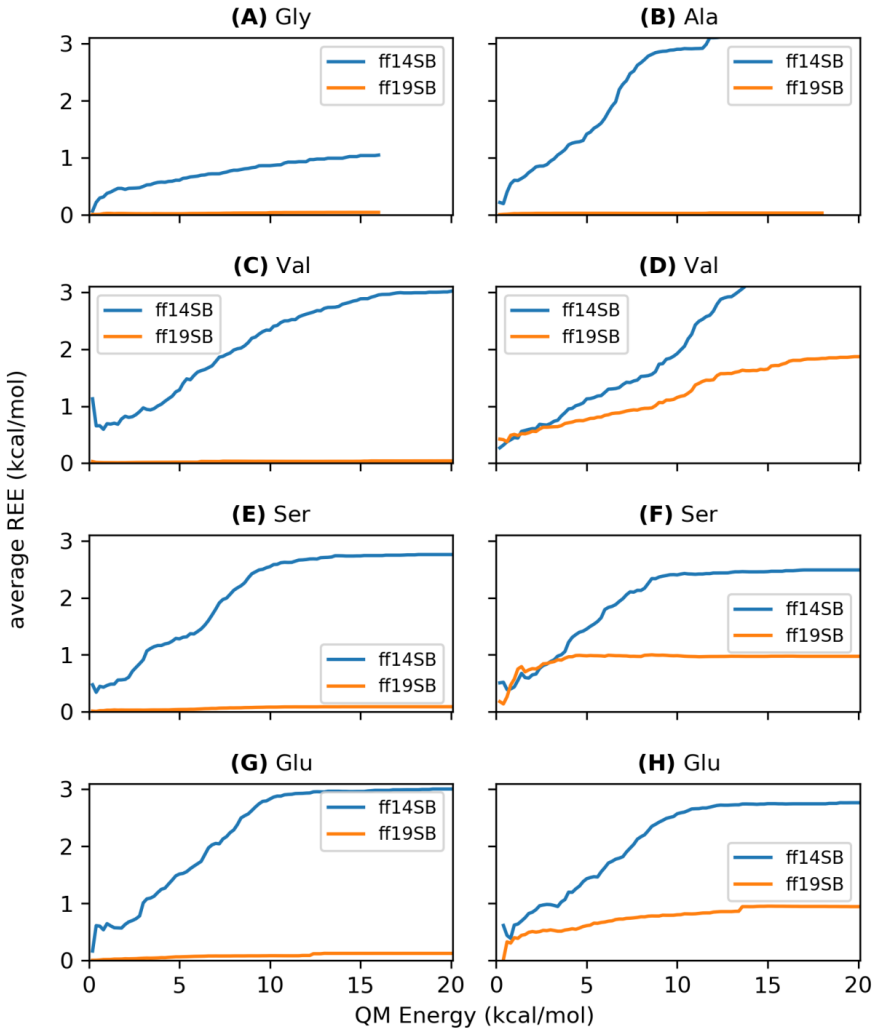


Figure 2.16. The average REE between QM and ff14SB, QM and ff19SB as a function of QM energy range above the minimum for (A) Gly dipeptide, (B) Ala dipeptide, (C) Val dipeptide in trans rotamer, (D) Val dipeptide in gauche(-) rotamer, (E) Ser dipeptide in gauche(+) rotamer, (F) Ser dipeptide in gauche(-) rotamer, (G) Glu dipeptide in mt-10 rotamer and (H) Glu dipeptide in tt10 rotamer.

2.4.2 Amino-acid specific Ramachandran sampling from PDB is reproduced better with ff19SB

As shown above, the CMAP procedure allows the MM 2D φ/ψ energy surfaces to quantitatively match the QM 2D training data. Furthermore, we showed that using CMAPs improves the ability of MM to reproduce changes in QM φ/ψ basin shapes and locations for different χ rotamers. An important question, though, is whether these QM-based training data for

dipeptides in solution provide good reference states for longer peptides in solution, or larger proteins with more complex structures and interactions. In order to explore the relevance of the differences seen between the ff19SB and ff14SB energy maps for different amino acids, we sought out high-quality PDB data[74, 128] on each amino acid and compared them to dipeptide φ/ψ sampling in MD using ff19SB. As discussed above in the context of statistical potentials, such comparisons have significant flaws, largely arising from the imperfect assumption that the distribution of backbone conformations for an amino acid across different proteins in a crystal environment (at different and typically low temperatures) corresponds to the MD-sampled Boltzmann distribution for the unconstrained peptide in solution at room temperature. Here, we restrict the use of the PDB data to a comparison of qualitative differences between amino acids from the same data source, such as from PDB or MD simulations. We expect that comparison of general features such as simulation and crystallographic basin shapes could provide valuable feedback that is independent of the dipeptide QM training data. However, we avoid assessment of quantitative features such as basin energies, for the reasons discussed above.

Distributions from the high resolution crystal structures[104] ("PDB"), dipeptide MD in ff14SB+OPC and dipeptide MD in ff19SB+OPC are shown in **Figure 2.17** for Ala, Val and Leu (with all amino acids shown in **Figure 2.18**). The OPC solvent model was selected for this test since this model was developed by optimizing the charge distribution to match QM data and vdW parameters to reproduce water density. Neither ff14SB nor ff19SB parameters were empirically adjusted with this model (ff14SB used TIP3P data in training). Because the dipeptide is fully exposed to the solvent, the results are more sensitive to the protein force field than to the solvent model; similar distributions are observed between ff14SB+OPC and ff14SB+TIP3P, and also between ff19SB+OPC and ff19SB+TIP3P (**Figure 2.18**) for each amino acid. However when comparing between force fields and PDB, as expected, the PDB distributions indicate that each of these amino acids samples unique features on the Ramachandran map. The ff14SB approach is clearly overly simplistic; when the same uncoupled Ala-based parameters are applied to all three amino acids, the peptides exhibit very similar φ/ψ sampling during MD, with the only apparent difference being slight changes to the population of the β basin (**Figure 2.17**). This result is consistent with the ff14SB potential energy maps (**Figure 2.9** and **Figure 2.12**) where only subtle differences in β basins are observed between Ala and Val. The ff14SB population maps also lack the diagonal shape of the α basin that is seen in the PDB data (and was also apparent in the

dipeptide QM data discussed above). In contrast, using amino-acid specific training against QM data with solvent polarization, the differences in Ramachandran maps are reproduced much better with ff19SB CMAPs. For instance in PDB, Val and Leu both have a flatter β-ppII transition region than Ala, with Val preferring greater population in this transition region. Compared to Ala, Leu has a broader diagonal α basin extending into the positive ψ region; these differences are reproduced more faithfully with ff19SB than ff14SB. The relative insensitivity of ff14SB backbone sampling to amino acid identity also explains its poor ability in modeling sequence dependence as discussed in the Introduction. Overall, given the fact that PDB data was not used in ff19SB training, this agreement between ff19SB and PDB shows a remarkable improvement in reproducing sequence-dependent behavior obtained using physics-based training, and highlights that these trends can be recapitulated without problematic empirical fitting against PDB data.

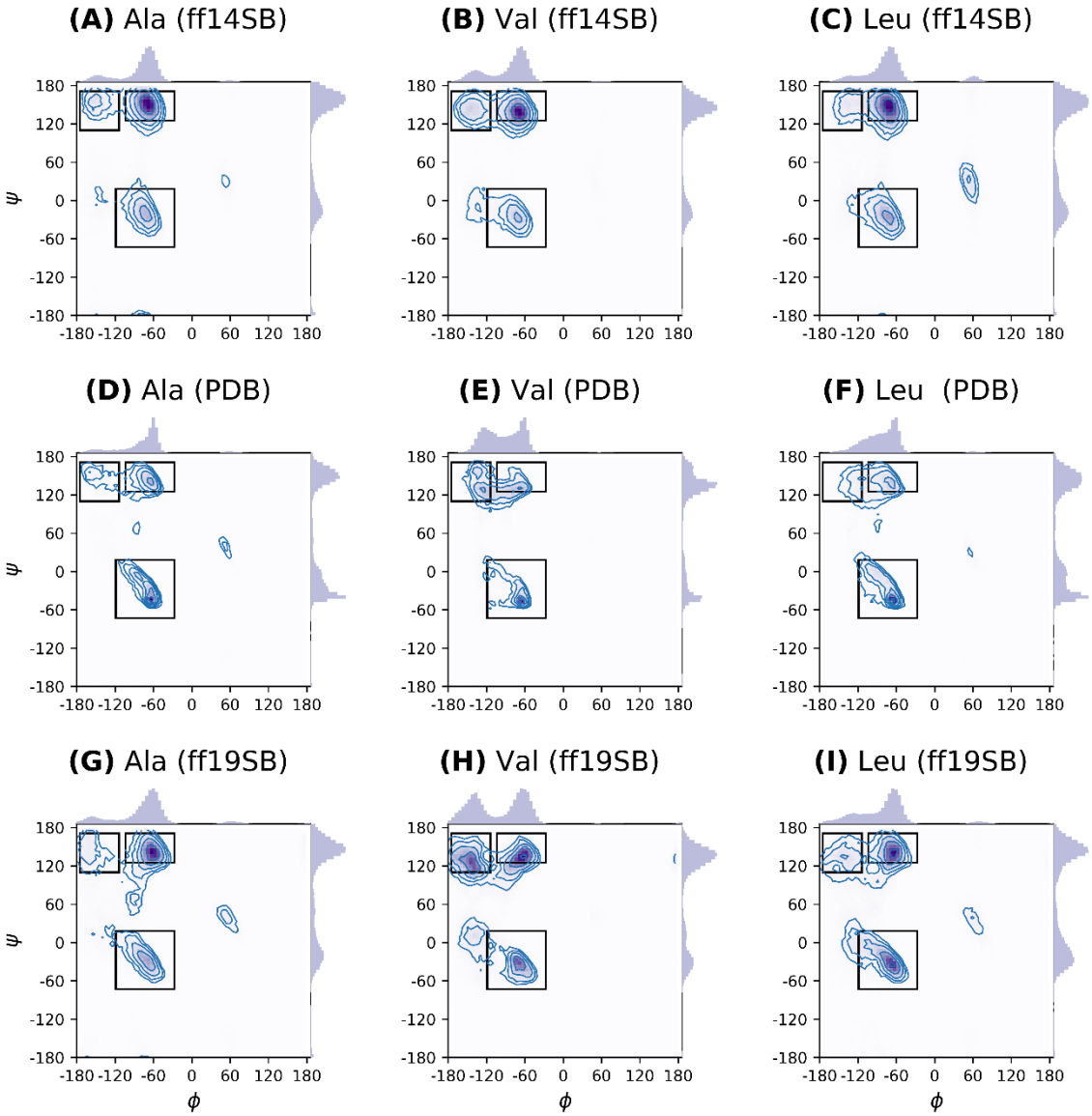
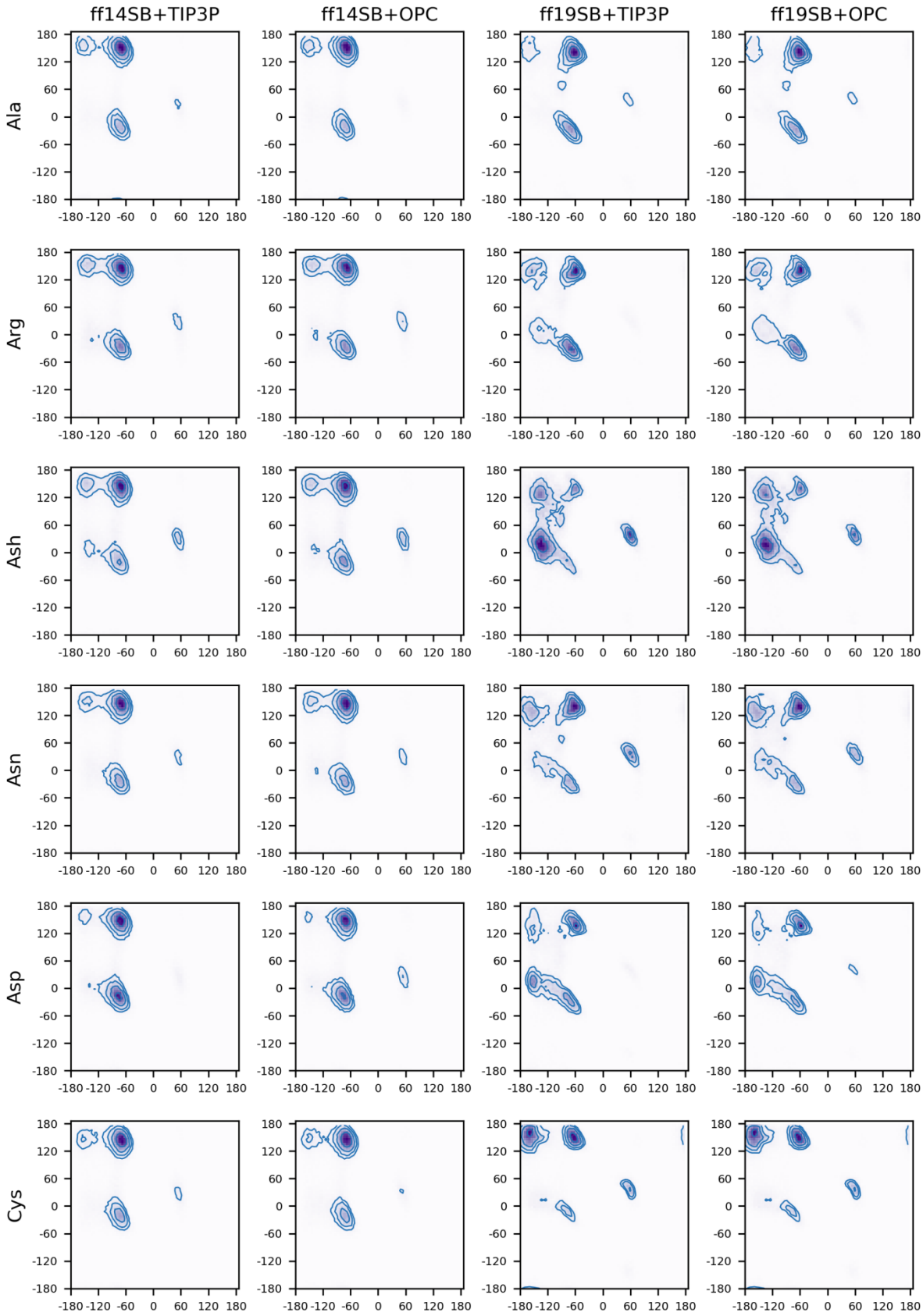
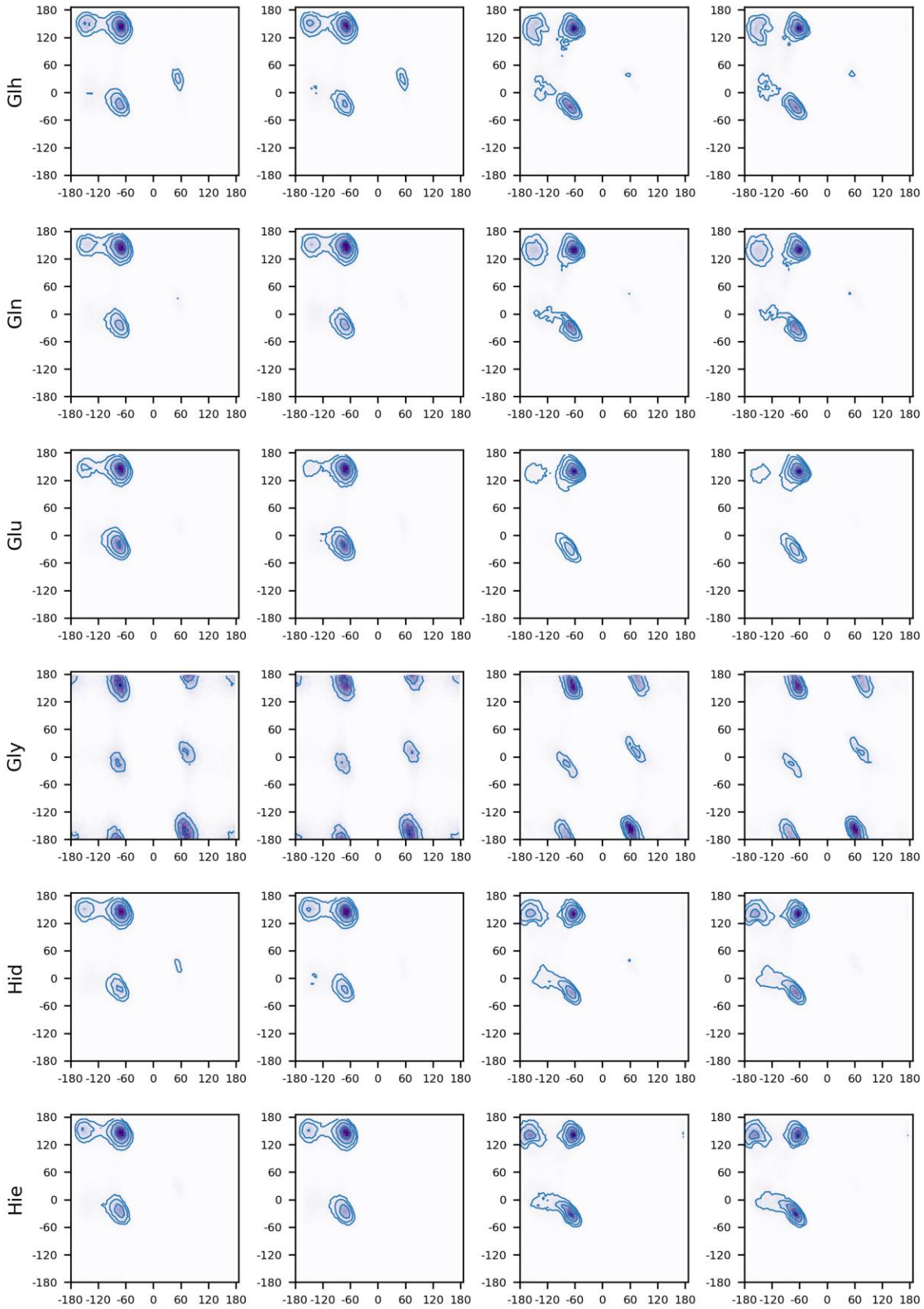
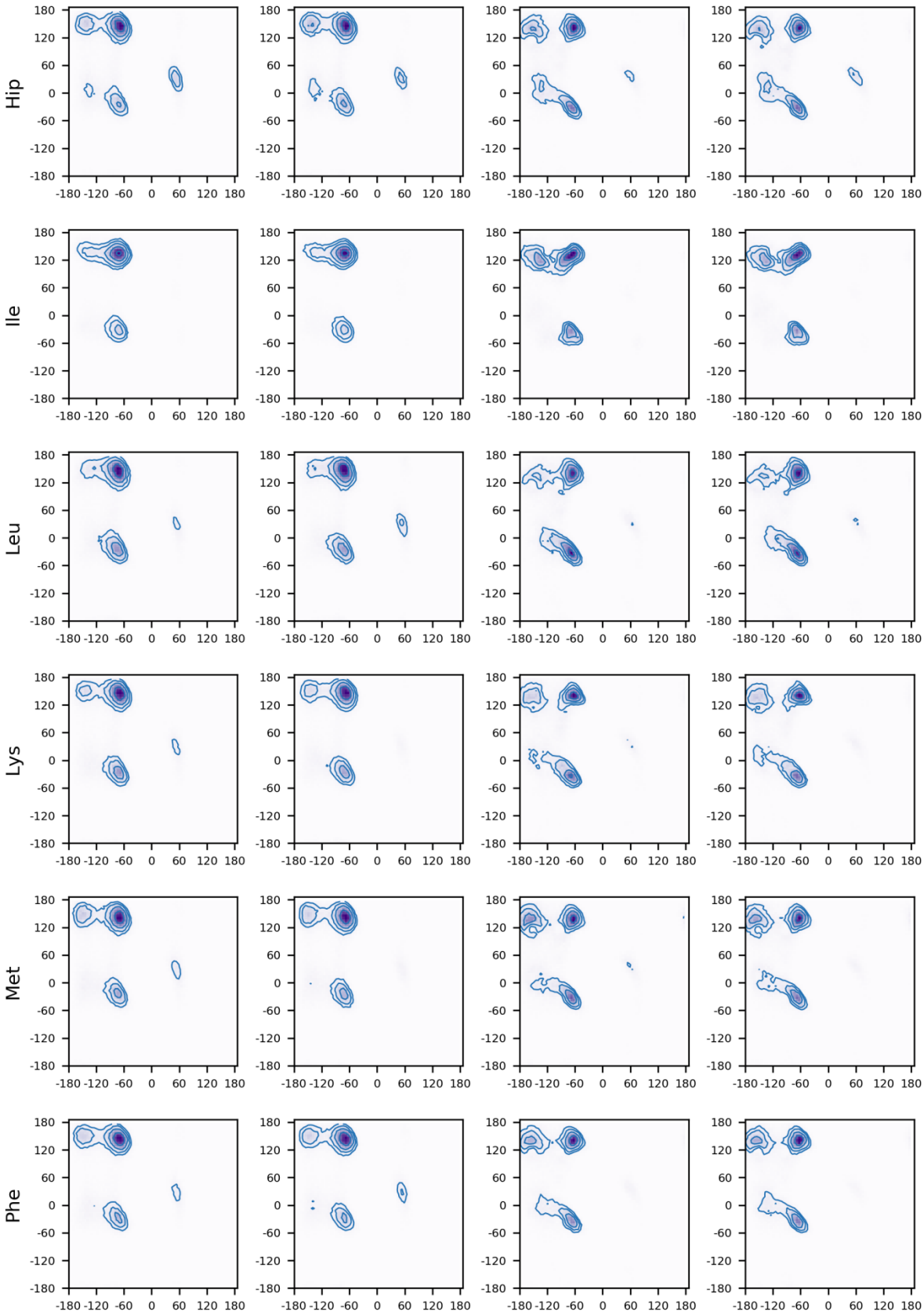


Figure 2.17 Ramachandran sampling shown for Ala, Val and Leu in dipeptide simulations with OPC water and ff14SB (A)-(C), in PDB (by Lovell et al.[74, 128]) (D)-(F), in dipeptide simulation with OPC water and ff19SB (G)-(I). Each contour line represents a doubling in population. Density is also shown as grids filled with light (no density) to dark (maximum density). Side histograms on each subplot represent independent distributions on φ and ψ. The box was defined in **Table 2.6** α, β and ppII. The MD simulations were run at 300K for a total of ~10 μs for all data shown.


ff14SB+TIP3P
ff14SB+OPC
ff19SB+TIP3P
ff19SB+OPC
Ala
Arg
Ash
Asn
Asp
Cys
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180


Glh
Gln
Glu
Gly
Hid
Hie
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180


Hip
Ile
Leu
Lys
Met
Phe
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180

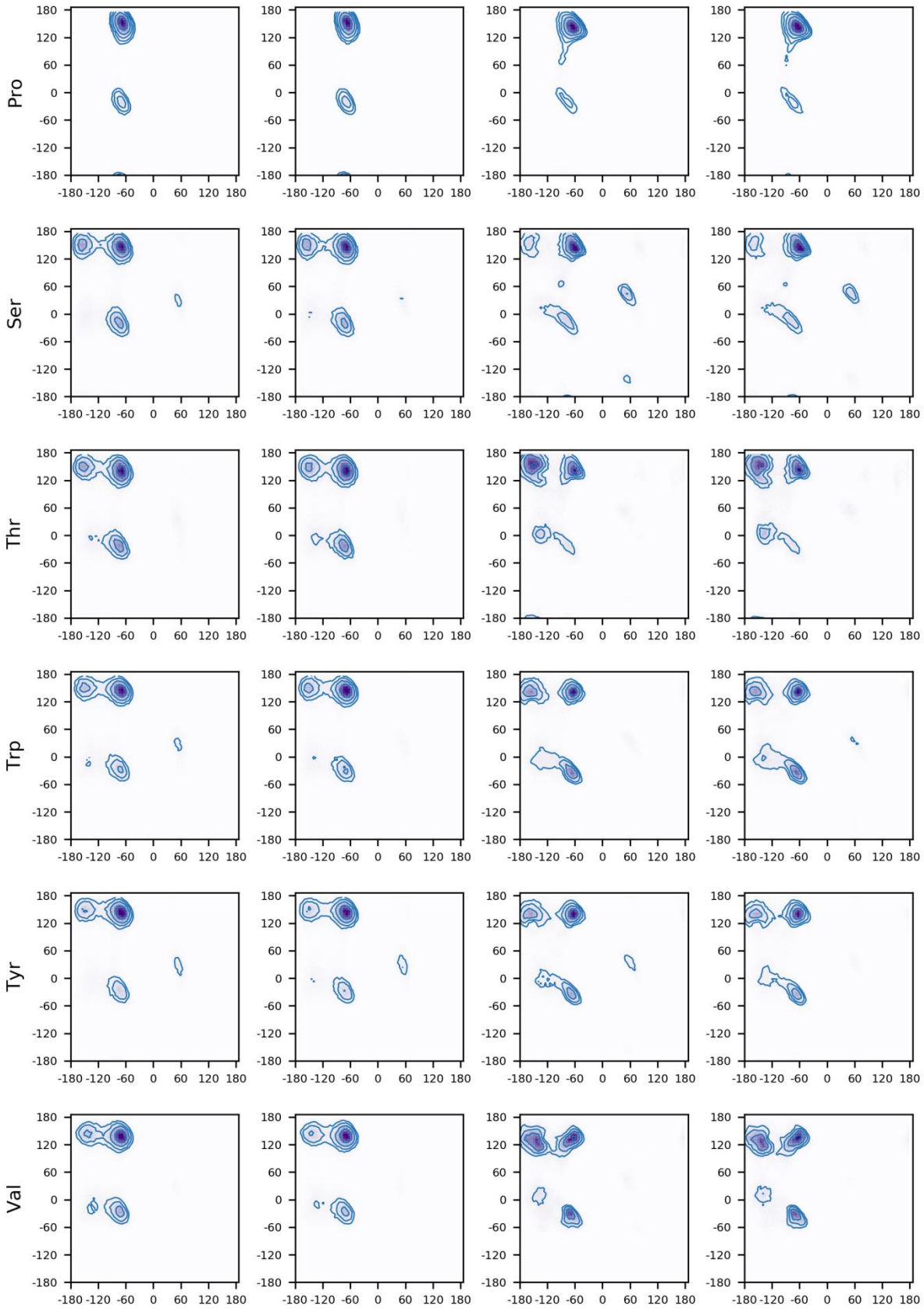

Pro
Ser
Thr
Trp
Tyr
Val
180
120
60
0
-60
-120
-180
-180 -120 -60 0 60 120 180

Figure 2.18 Ramachandran sampling maps, where the X and Y axes of each plot are ϕ and ψ, respectively, from ff14SB+TIP3P (1st column), ff14SB+OPC (2nd column), ff19SB+TIP3P (3rd column) and ff19SB+OPC (4th column) simulation for 24 dipeptides including alternate protonation states for Asp, Glu and His. The distributions were used for χ^2 analysis. Each contour line represents a doubling in population. Density is also shown as grids filled with white (no density) to purple (maximum density).

2.4.3 Improved reproduction of NMR $^3J(HNHA)$ scalar couplings on blocked dipeptides

Another way to examine the ability of ff19SB to improve amino-acid specific behavior in solution is through quantitative comparison against NMR data probing backbone dihedrals, which have been reported[121] for each of the amino acids in a dipeptide form (except Pro which lacks HN). As explained (**Methods: CMAP fitting groups**), a total of 16 CMAPS were fit and then applied to 20 natural amino acids (also including alternate side chain protonation states) in ff19SB. We compared the performance of ff19SB and ff14SB by simulating blocked dipeptide systems (**Methods: Structure preparation & simulations**) in both OPC and TIP3P solvent models. We then calculated the $^3J(HNHA)$ from each MD trajectory based on the Karplus equation[119] and "Orig" parameter set[120] and quantified the agreement by calculating the χ^2 error following Best et al[29a] and us[26a]. The χ^2 error was also used as an empirical target in ff14SB backbone training[2c]. The χ^2 value quantifies the agreement between experimental and MD ensemble average J value(s), also taking into account the uncertainty of the theoretical model being used. In theory, smaller χ^2 errors correspond to better agreement between MD and experiment. However, χ^2 values below one only indicate that the error is smaller than the uncertainty of the model and do not necessarily indicate continued improvement vs. experiment. Further details of the calculations and precision estimates are provided in Methods (**Methods: NMR scalar coupling calculations**).

The calculated $^3J(HNHA)$ values for each amino acid, using four different combinations of FF (ff14SB and ff19SB) and water model (OPC and TIP3P), are provided in **Table 2.9**, with the χ^2 errors for OPC shown in **Figure 2.19** and TIP3P shown in **Figure 2.20**. Though we observed differences among force fields for the Ramachandran sampling maps, the χ^2 errors and actual ^{3}J(HNHA) values appear relatively insensitive to force field. For a given force field, neither Ramachandran sampling maps nor the χ2 errors and actual 3J(HNHA) values are sensitive to solvent model. For instance, for either ff14SB or ff19SB, the average χ^2 errors are similar and

mostly below 0.5 for both OPC and TIP3P (**Figure 2.19** and **Figure 2.20**). In this respect, the performance of ff19SB is not significantly improved over ff14SB for dipeptide NMR data, as ff14SB already showed reasonable behavior with few amino acids having errors larger than 1.0 (His^+ and Cys) for both solvent models. In addition, the histograms of χ^2 errors are similar regardless of the force field and solvent model (**Figure 2.21**). Together with the fact that $^3J(HNHA)$ in the Karplus calculation is sensitive only to the φ dihedral, this test seems insufficient to examine the specificity of parameters for different amino acids and the quality of parameters across the full Ramachandran space. However, this is a good indicator that the QM fitting is reasonable and ff19SB introduced no spurious outliers.

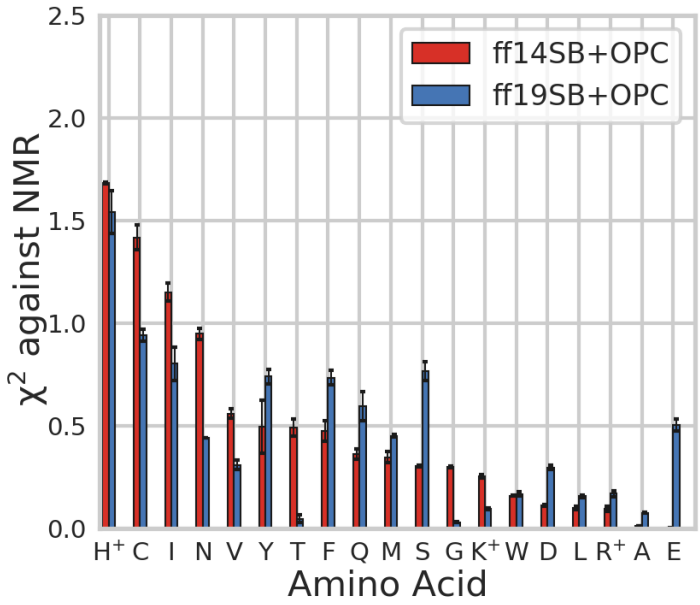


Figure 2.19 χ^2 errors in reproducing NMR $^3J(HNHA)$ coupling data for all non-Pro amino acids (using single letter codes on X axis), with data for ff14SB+OPC (red) and ff19SB+OPC (blue). The MD simulations were run at 300K for a total of ~60 μs for all data shown.

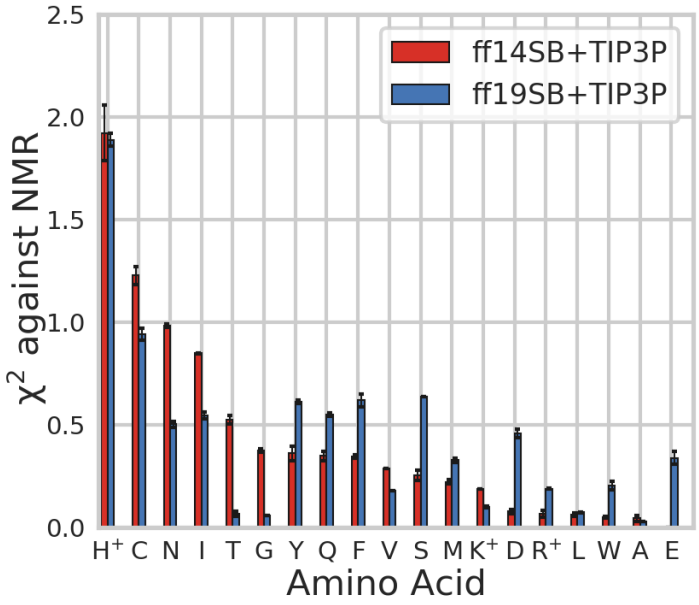


Figure 2.20 χ^2 errors in reproducing NMR ^{3}J(HNHA) coupling data for all non-Pro amino acids (using single letter codes on X axis), with data for ff14SB+OPC (red) and ff19SB+OPC (blue). The MD simulations were run at 300K for a total of ~60 μs for all data shown.

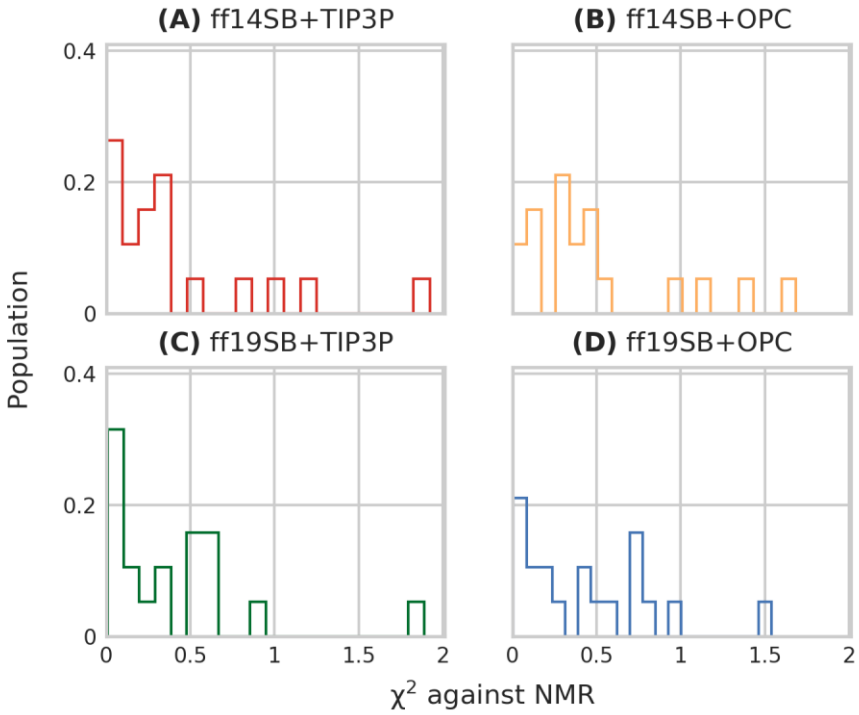


Figure 2.21 Histogram on χ^2 errors for all non-Pro amino acids with data for (A) ff14SB+TIP3P, (B) ff14SB+OPC, (C) ff19SB+TIP3P and (D) ff19SB+OPC.

As shown in **Figure 2.19**, ff19SB+OPC gave a slightly larger error for Glu, but since the pH used in the NMR experiment (4.9) was close to the Glu side chain pKa (~ 4.25), a simulation using either a protonated or deprotonated state of Glu may not adequately model the experimental ensemble. To address this ambiguity, we ran constant pH simulation (pH=4.9) on Glu dipeptide (**Methods: Constant pH simulation**), and obtained the carboxyl group protonated state ratio for each force field + solvent model combinations (**Table 2.12**). Next, we performed regular MD for both protonated and deprotonated Glu. The combined trajectory weighted by protonation state ratio (**Methods: Constant pH simulation**) was used so that our calculated χ^2 more accurately reflected the protonation states in the experiment.

Table 2.12 Averaged side chain protonation state ratio of Glu and Asp dipeptide from constant pH simulation using ff14SB+TIP3P, ff14SB+OPC, ff19SB+TIP3P and ff19SB+OPC. Error bars were calculated from two independent runs starting from either helical or extended conformation.

	Glu	Asp
ff14SB+TIP3P	0.48±0.01	0.17±0.01
ff14SB+OPC	0.46±0.01	0.17±0.01
ff19SB+TIP3P	0.46±0.01	0.22±0.01
ff19SB+OPC	0.43±0.01	0.21±0.01

For deprotonated Glu, the ppII region is the most populated in both ff14SB and ff19SB and the shape of energy basins are similar between ff14SB and ff19SB regardless of the solvent model (**Figure 2.15**). However, ff19SB samples the ppII basin extending farther towards $\phi > -60°$ than ff14SB. This subtle change causes the $^3J(HNHA)$ to deviate significantly from experiment (χ^2 = 1.31±0.03). This shift, however, is much less pronounced in the protonated state MD with ff19SB (**Figure 2.15**), resulting in a much smaller χ^2 error of 0.031±0.01. Overall, the χ^2 value from the re-weighted population at pH 4.9 was calculated to be 0.50±0.03, indicating that the scalar coupling calculated with ff19SB is in reasonable agreement with experiment once the protonation state is taken into account.

We also performed constant pH simulation at pH=4.9 for Asp, obtaining the side chain carboxyl protonation ratio for different force field + solvent model combinations (**Table 2.12**). The χ^2 values from Asp simulation with deprotonated side chain and pH-weighted ensemble were

calculated to be 0.01±0.01 and 0.30±0.01, respectively, with both indicating reasonable agreement with experiment for ff19SB. In addition, for both Asp and Glu, in either ff14SB or ff19SB simulations, using TIP3P vs. OPC has little effect on the χ^2 results with average χ^2 errors all below 0.5.

In summary, both ff19SB and ff14SB provided reasonable results in reproducing NMR scalar coupling when using either OPC or TIP3P solvent, indicating that this test is relatively insensitive to the sampling differences that are apparent in the Ramachandran surfaces (**Figure 2.15**). It is encouraging, however, that ff14SB includes an empirical adjustment to improve agreement with the same type of NMR data as used here, while the QM-trained ff19SB achieves similar or better accuracy without empirical adjustment.

2.4.4 Accurate reproduction of Ala_5 NMR scalar couplings is maintained in ff19SB

We next tested ff19SB by simulating Ala_5 in both OPC and TIP3P solvents, and compared to ff14SB. A total of six NMR scalar couplings have been measured on this peptide[28]. Following Best et al.[29a] and us[2c, 26a] previously, we calculated the scalar couplings from each MD trajectory as discussed above, and quantified the agreement between simulations and NMR by calculating the χ^2 error (**Methods: NMR scalar coupling calculations**). The NMR data, calculated scalar couplings for ff14SB and ff19SB in both OPC and TIP3P water and the systematic error σ[26a, 29a] used in χ^2 calculations are provided in **Table 2.10**, with the χ^2 errors in OPC shown in **Figure 2.22** and TIP3P shown in **Figure 2.23**. Overall, the average χ^2 errors are smaller than one regardless of force field and solvent model, indicating a reasonable reproduction of NMR data for ff14SB and ff19SB with both OPC and TIP3P. Specifically, ff19SB has smaller averaged χ^2 compared to ff14SB for both OPC (0.77±0.03 vs. 0.93±0.10) and TIP3P (0.77±0.03 vs. 0.88±0.09) solvent model. The measurement of $^3J(HNCA)$ is correlated with the ϕ dihedral as well as the ψ dihedral of the preceding amino acid[28, 119]; this is the only coupling we examined that depends on two dihedrals instead of one. This Karplus correlation has the smallest σ among all of these scalar coupling types, making it more sensitive to error than other scalar coupling types. Even though the χ^2 value is large (**Figure 2.22** and **Figure 2.23**), the difference between simulation and NMR in

actual $^{3}J(HNCA)$ value is as small as 0.2 across all models, suggesting reasonable agreement between simulation and NMR across different models (**Table 2.10**).

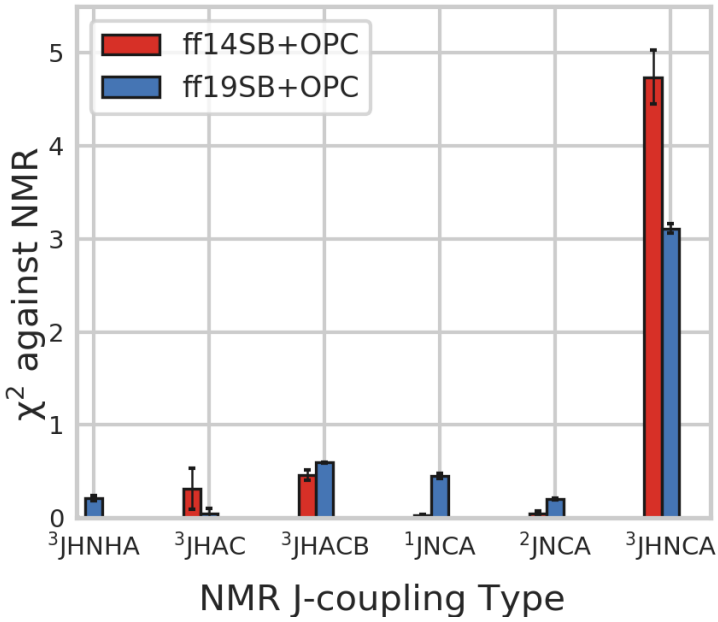


Figure 2.22 χ^2 errors in reproducing six NMR scalar coupling data for Ala_5, with data for ff14SB+OPC (red) and ff19SB+OPC (blue). The MD simulations were run at 300K for a total of ~3 μs.

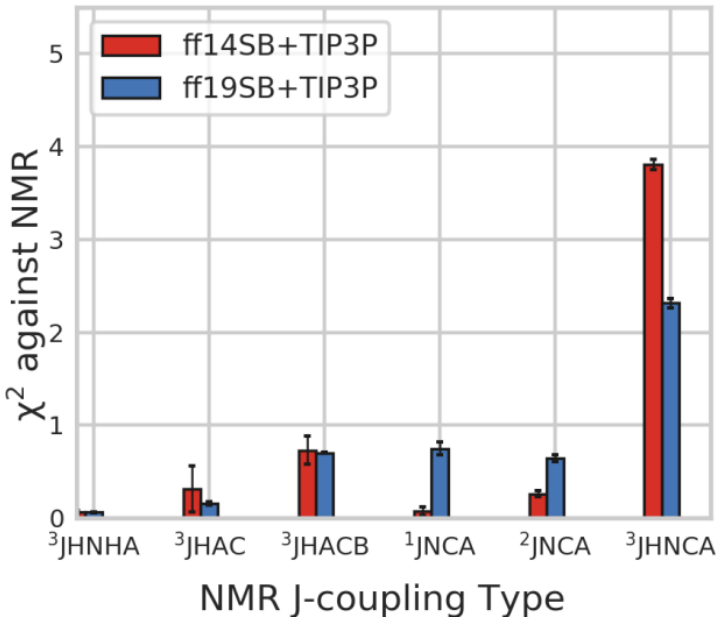


Figure 2.23 χ^2 errors in reproducing multiple NMR scalar coupling data for Ala_5, with data for ff14SB+TIP3P (red) and ff19SB+TIP3P (blue). The MD simulations were run at 300K for a total of ~3 μs.

2.4.5 Amino-acid specific helical propensities are significantly improved in ff19SB

Since the scalar coupling χ^2 analysis presented above was relatively insensitive to the updated residue-specific parameters, additional tests were performed to further validate the new model. The *$^3J(HNHA)$* analysis is only sensitive to the distribution for φ; thus, we calculated amino-acid specific helical propensities to probe ψ dihedral sampling. We focus both on the absolute helical propensity in the force field as well as the ability to reproduce known differences between amino acids. We performed multiple MD simulations on model peptides with sequence Ace-A_4**X**A_4-NH_2 with varying **X**, and fit helical propensity parameters *w* through Lifson-Roig[116] theory implemented in a genetic algorithm (**Methods: Helical propensity**). Different from having three substitutions in Best et al.'s system[32], our model peptides only have a single substitution, as was done for the experimental system[72], to avoid possible interaction between the substitutions across turns of helix. The sensitivity to the peptide length was tested by comparing propensities calculated using A_4**X**A_4 and A_9**X**A_9 in ff14SB + GBneck2; calculated helical propensities for all amino acids with ff14SB + GBneck2 are well correlated between A_4**X**A_4 and A_9**X**A_9 (**Figure 2.24**), justifying the use of the shorter peptide in the more computationally expensive explicit solvent simulations.

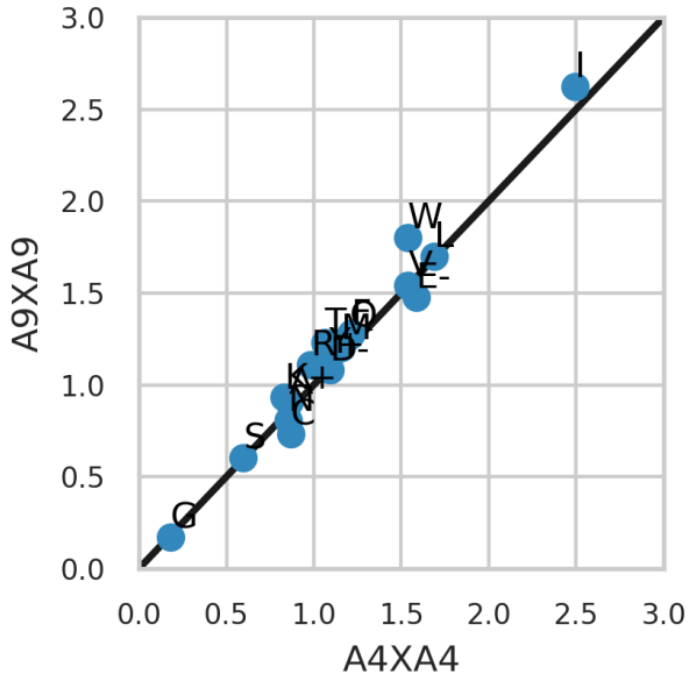


Figure 2.24 Comparison of helical propensity w from simulations of $A_4\underline{\mathbf{X}}A_4$ and the longer $A_9\underline{\mathbf{X}}A_9$ with ff14SB+GBneck2. The MD simulations were run at 300K for a total of ~1008 μs.

We also calculated the sensitivity of the results to the exact definition of the helical region of overall φ/ψ space (defined in **Table 2.6**) using ff14SB and ff19SB, in both OPC and TIP3P. The calculated helical propensities for each force field and solvent model show little sensitivity to the α basin definition, especially for ff19SB+OPC (**Figure 2.6**).

Helical propensities were calculated for $A_4\underline{\mathbf{X}}A_4$ with ff14SB and ff19SB, in TIP3P, TIP4P-Ew[57] and OPC, and also for ff19SB in OPC3[90]. The results of the MD simulations are compared to values based on experiments[72]. Data for ff14SB+TIP3P, ff14SB+OPC, ff19SB+TIP3P and ff19SB+OPC are shown in **Figure 2.25**, with data for TIP4P-Ew and ff19SB+OPC3 in **Figure 2.26**. Histidine was excluded from plots, see **Methods: Helical propensity** for details. Numerical values are provided in **Table 2.7** and **Table 2.8**. For TIP4P-Ew and ff19SB+OPC3 runs, a subset of 12 representative amino acids were selected due to the computational expense of the calculations. Ala, Leu, Ile, Gln and Trp were selected since helicities for these are significantly overestimated in ff19SB+TIP3P (**Figure 2.25C**). Charged amino acids Glu, Arg and Lys were selected as well. In addition, several amino acids having low (Gly and Asn) and medium (Val and Phe) experimental helical propensity were selected.

In general, ff14SB has difficulty reproducing the trend from NMR experiments regardless of solvent model. In TIP3P, Ala should be the most helical amino acid but is distinctly underestimated, while most other amino acids have significantly overestimated helical propensities, and the overall residue-specific correlation with NMR is poor at $R^2 = 0.38$ (**Figure 2.25A**). Although OPC is arguably a better water model[66] than TIP3P, combining it with ff14SB produces worse results than in TIP3P ($R^2 = 0.27$, **Figure 2.25B**), with helical propensities being underestimated for most amino acids. There is very little sequence dependence, with a slope of 0.49. The amino acids with negatively charged side chains (Asp and Glu) are outliers in both solvent models for ff14SB. Results in TIP4P-Ew are similar, with $R^2 = 0.41$ and somewhat lower overall helical propensities than in TIP3P (**Figure 2.26**).

This poor correlation with experiment appears to be due to ff14SB rather than weaknesses in these solvent models; the correlation is significantly higher when comparing the helical propensities of ff14SB in 2 water models (OPC vs. TIP3P $R^2 = 0.84$ as shown in **Figure 2.26**, with TIP3P giving higher helical propensities). These results suggest that the ff14SB force field would be unable to reliably model quantitative changes to secondary structure or protein stability due to point mutations, despite its ability to successfully fold large proteins to near-native structures[70]. Protein folding tests are likely less sensitive to sequence-specific energetics since the overall fold can be maintained even when a large fraction of the protein sequence is varied[132].

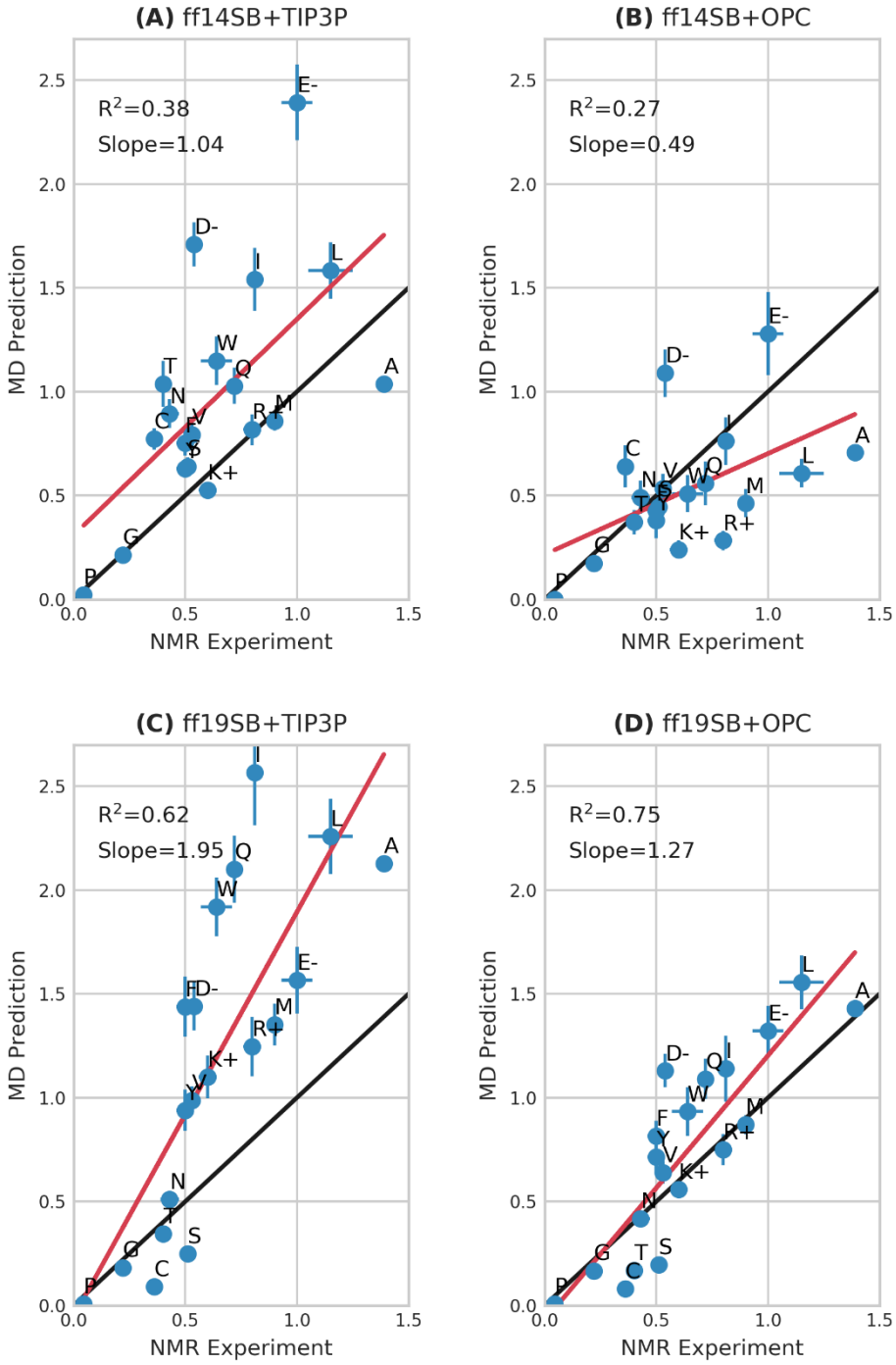


Figure 2.25 Correlation between helical propensities *w* from experiment[72] and simulations using (A) ff14SB+TIP3P, (B) ff14SB+OPC, (C) ff19SB+TIP3P and (D) ff19SB+OPC. Amino acids are indicated using single letter codes. Values on the X-axis represent the data based on NMR[72] and the reported standard deviations. Values on Y-axis represent the helical propensities fit against the combined trajectory (3.2 μs * 12), with error bars calculated via bootstrapping analysis. Black lines represent perfect agreement. Linear regression (red lines) was performed against the data points,

with R^2 and slope quantifying the goodness of fit. The MD simulations were run at 300K for a total of ~3225 μs.

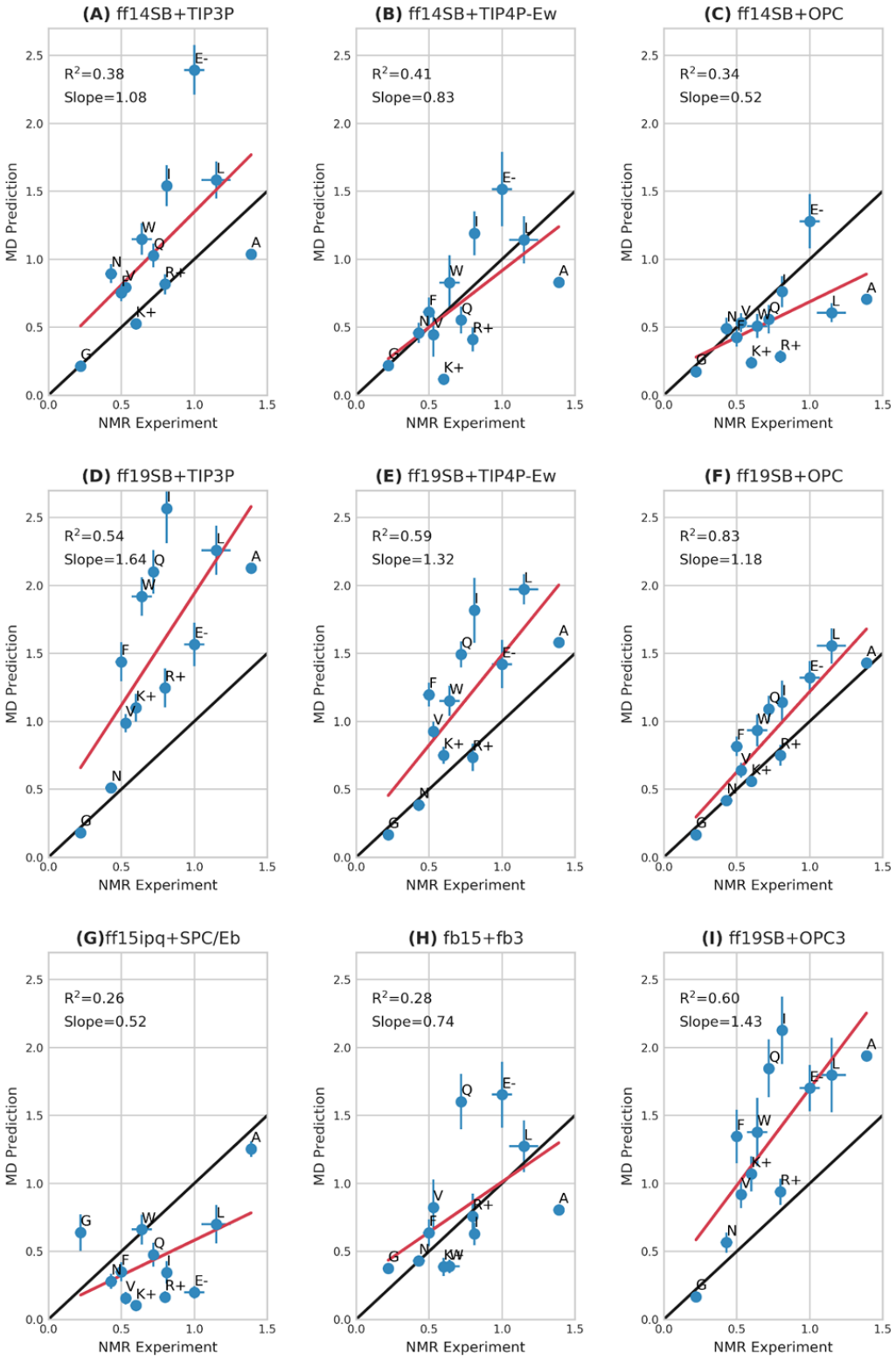


Figure 2.26 Correlation between helical propensities *w* from experiment and simulations using (A) ff14SB+TIP3P, (B) ff14SB+TIP4P-Ew, (C) ff14SB+OPC, (D) ff19SB+TIP3P, (E) ff19SB+TIP4P-Ew, (F) ff19SB+OPC, (G) ff15ipq+SPC/E_b, (H) fb15+fb3 and (I) ff19SB+OPC3 Only 12 amino acids were calculated in TIP4P-Ew, ff15ipq+SPC/E_b, fb15+fb3 and ff19SB+OPC3, thus only these 12 were included in all plots for comparison. Amino acids are indicated using single letter codes. Values on Y-axis represent the fitted helical propensities from the original combined trajectories (3.2 μs * 12), with error bars calculated via bootstrapping analysis. Values on X-axis represent the reported NMR data and the standard deviation on these values. Linear regression was performed against the data points, with R^2 and slope quantifying the goodness of fit. Black lines represent a perfect linear correlation. Red lines represent a best-fit line via linear regression. The MD simulations in TIP4P-Ew, ff15ipq+SPC/E_b, fb15+fb3 and ff19SB+OPC3 were run at 300K for a total of ~2304 μs.

Ideally, the ff19SB residue-specific training against QM data should improve modeling of sequence-dependent behavior and give improved correlation to experimental residue-specific differences. Consistent with this expectation, we find that using ff19SB+TIP3P reproduces the experimental trend much better than ff14SB+TIP3P (R^2 = 0.62 vs 0.38, respectively, **Figure 2.25C** vs. **A**). However for ff19SB+TIP3P we also observe substantially higher sensitivity to amino acid variation than in experiment (slope = 1.95, **Figure 2.25C**). The source of this high slope and amplified sensitivity may be weaknesses in TIP3P (see **Introduction**), in particular the bias favoring compact structures like helices.

When ff19SB is combined with the better OPC water model (**Figure 2.25D**), the correlation between simulated and experimental helical propensities is further improved (R^2 = 0.75 vs. 0.62 in TIP3P) and the sensitivity to amino acid is also improved (slope = 1.27 vs. 1.95 in TIP3P). The sensitivity of the model still seems slightly overestimated, with slope modestly larger than unity. The remaining deviations from a perfect linear correlation may not be highly significant, since small disagreements also exist among various experimental measurements (**Figure 2.7**). In OPC, the helical propensity for Ala remains slightly too low with ff19SB, and Leu is similar to Ala within uncertainties (**Table 2.7**). Ser, Thr and Cys are all predicted to have helical propensity somewhat lower than experiment; all have short, polar side chains that could compete with backbone hydrogen bonding and reduce helical content. This will be investigated in more detail in the future.

These results show that ff19SB has significantly improved capability to differentiate amino acid properties and thus should have better predictive power for modeling sequence-specific

behavior, protein mutations, and also rational protein design which requires quantitative sequence-structure accuracy.

In addition to ff14SB and ff19SB, we also considered several other recent Amber-related force fields in combination with their recommended water model. In ff15ipq[58]+SPC/E_b[59], Ala shows good agreement with experiment, but otherwise there is poor overall correlation and weak sensitivity among the remaining amino acids (R^2 = 0.26 and slope = 0.52, **Figure 2.26G**). In fb15[86]+fb3[94], Ala helical propensity is much lower than NMR, and the overall correlation is also poor (R^2 = 0.28 and slope = 0.74, **Figure 2.26H**). As with the TIP4P-Ew and ff19SB+OPC3 runs, 12 representative amino acids were included for these Amber-related force fields tests.

Best et al. reported helical propensity benchmarks for 20 amino acids, showing that the overall trend from experiments[72] was poorly reproduced by two force field + water combinations (ff03w[133]+TIP4P/2005[134] and ff99SB*[29b]+TIP3P[89]) with correlation coefficients R^2 being 0.01 and 0.22 respectively[32]. Therefore, they performed an empirical adjustment of a few amino acids, together with the updated parameters in the ILDN[31] variants of ff99SB*, to better match helix-coil transition data. They refit partial charges of Cα and side chain atoms on charged amino acids (D, E, K, R) while forcing the charges on amide N, H, C, O to have same values as all the other residues. The helical propensities[32] using these charge-refit residues were better correlated with experiment (R^2 = 0.51 and slope = 0.68 for all amino acids, **Figure 2.28**) than the original ff03w and ff99SB*, but even with this empirical fitting the overall trend for the 20 amino acids is still notably worse than ff19SB+OPC (R^2 = 0.75 and slope =1.27, **Figure 2.28**).

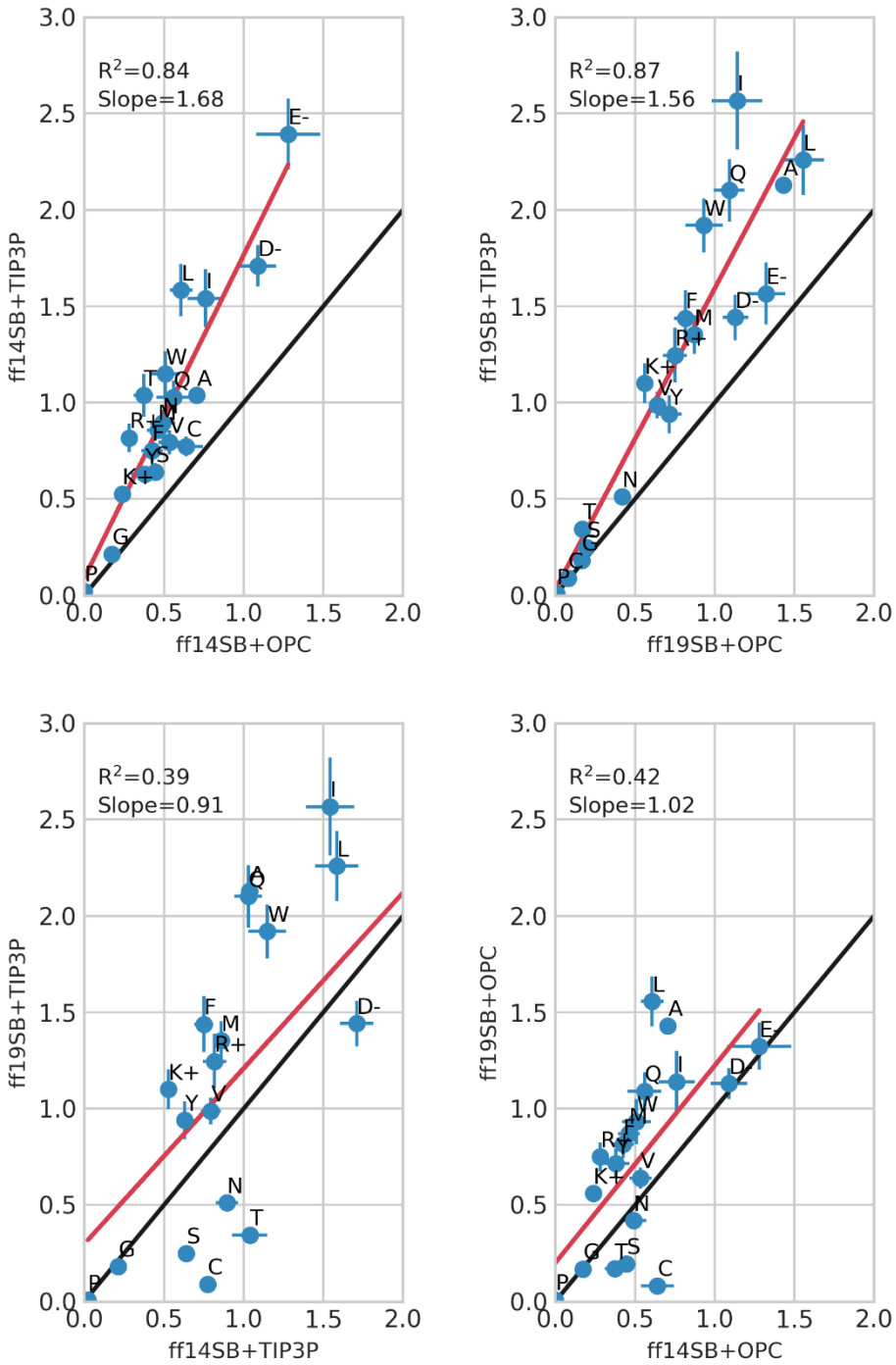


Figure 2.27 Correlation between helical propensities w from simulations. Amino acids are indicated using single letter codes. Values on Y-axis represent the fitted helical propensities from the original combined trajectories (3.2 μs * 12), with error bars calculated via bootstrapping analysis. Values on X-axis represent the reported NMR data and the standard deviation on these values. Linear regression was performed against the data points, with R^2 and slope quantifying the goodness of fit. Black lines represent a perfect linear correlation. Red lines represent a best-fit line via linear regression.

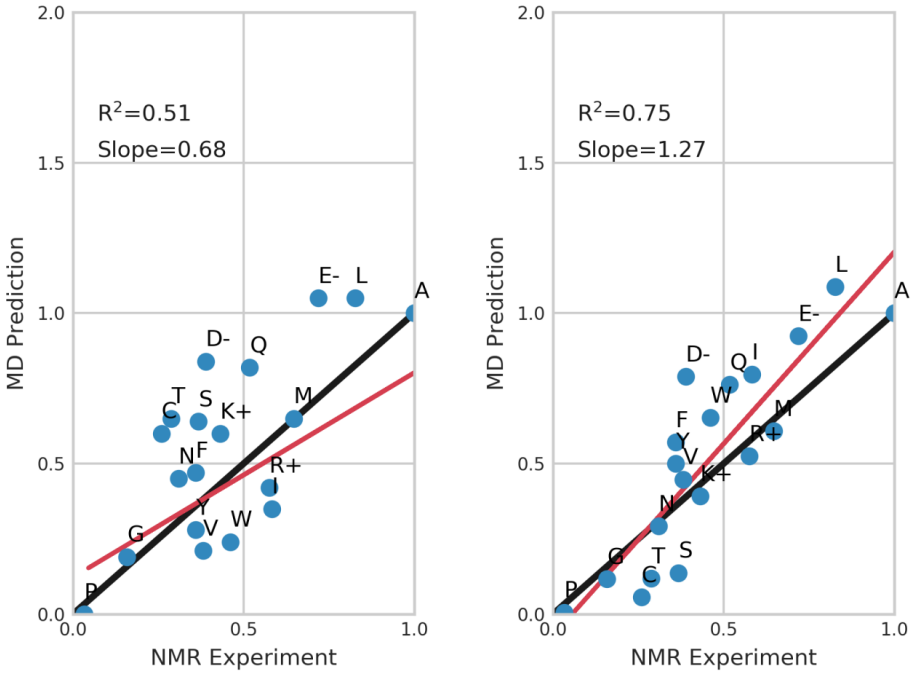


Figure 2.28 Correlation between helical propensities w from experiment and simulations using (left) Best et al[32], (right) ff19SB+OPC. Amino acids are indicated using single letter codes. Values on both X-axis and Y-axis represent the helical propensities normalized here by dividing by the Ala helical propensity, consistent with what was done by Best et al[32]. Linear regression was performed against the data points, with R^2 and slope quantifying the goodness of fit. Black lines represent a perfect linear correlation. Red lines represent a best-fit line via linear regression.

The helical propensity for Cys, Ser and Thr are notably low in ff19SB comparing to ff14SB regardless of solvent models (**Figure 2.27**). There are two possible reasons. (1) The solvation difference between QM and MM in ff19SB training lead to over-correction on backbone potential. (2) Because all these three amino acids have short polar side chain, the intra-molecular hydrogen bond between side chain and backbone amide group in adjacent residues might compete with hydrogen bond stabilizing α-helical structure. The first assumption is difficult to test, hence we focus on the second one. In order to investigate the correlation between these intra-molecular hydrogen bond with low helical propensity, we performed hydrogen bond analysis on the A4XA4 data. The fraction of these intra-molecular hydrogen bond in MD simulations are provided in **Table 2.13**. The side chain of Cys (-SH group) doesn't form any hydrogen bond with backbone probably because of so weak polarity of sulfur. The hydroxyl group (-OH) in both Ser and Thr can form hydrogen bond with amide groups in A4**X**A4 backbone for both TIP3P and OPC. Thr is

notably stronger than Ser in forming this type of intra-molecular hydrogen bond. The β-branched side chain structure in Thr makes it more rigid than Ser side chain and might form a hydrogen bond more easily with costing less entropic penalty.

Table 2.13 The cumulated average fraction of intra-molecular hydrogen bond between side chain and all backbone amides in A4**X**A4 during MD simulations. The error bar is calculated as standard deviation across all 12 independent MD runs.

hbond%	Cys	Ser	Thr
ff19SB+TIP3P	0.0%	18.9%±3.0%	40.8%±4.5%
ff19SB+OPC	0.0%	9.8%±1.7%	17.7%±3.3%

2.4.6 Evaluating helical content in the K19 peptide

In order to assess the ability of ff19SB to model α-helices in more complex systems, we employed the Ala-rich Baldwin-type[135] peptide K19[95] that was previously simulated[2c] using ff14SB. Experimental measurements[95] on K19 using NMR chemical shift deviations (CSDs) suggest that the fraction helix at 300 K of four central residues and two residues near the C-terminus are ~0.38 and ~0.17, respectively (**Figure 2.29**). Simulations with ff14SB+TIP3P exhibited an average 0.30±0.05 (central four) and 0.19±0.03 (two near C terminus) fraction helix, in close agreement with our previously reported[2c] value of 0.30±0.05 and 0.20±0.04 using the same force field and solvent model. Both values are in good agreement with the experiment, likely reflecting the inclusion of K19 data generated using TIP3P in the empirical adjustment of ff14SB backbone parameters.

In order to better separate the accuracy of the solute force field from that of the solvent model, we ask: does the good match come from a good modeling of protein and water separately, or from training-based error cancellation between the force field and solvent model? As shown in **Figure 2.29**, after substituting TIP3P with a better model for water (OPC), ff14SB MD resulted in significantly reduced helicity, with 0.08±0.02 (central four) and ~0.08±0.01 (two near C terminus) helical content for the 6 measured residues. Given OPC's excellent agreement with water properties, the worsened agreement with experiment for K19 supports a fortuitous cancellation of error in the combination of ff14SB+TIP3P. Since overly weak solvent-solute dispersion in TIP3P[87a, 87c] may introduce a bias in favor of compact structures, it seems reasonable that this bias

may also enhance helical content to maintain hydrogen bonding in compact states. This hypothesis is supported by data in **Figure 2.30**, which shows an inverse correlation between helical content and radius of gyration of K19, indicating that more compact structures tend to be more helical, and also **Figure 2.25**, which shows a dramatic increase in helical propensities when combining ff19SB with TIP3P vs. OPC. We conclude that an inherent underestimation of helicity is present in ff14SB, which is (inexactly) compensated by an increase in helical content driven by the TIP3P bias toward overly compact structures.

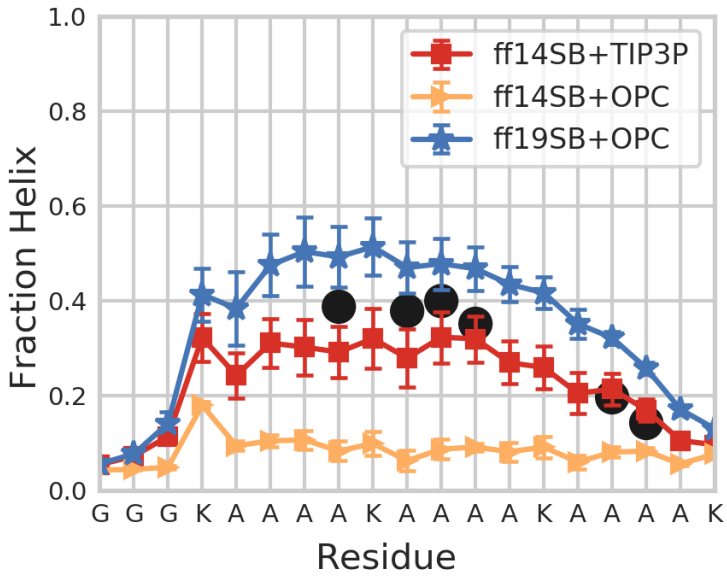


Figure 2.29 The fraction helix of each amino acid in K19 sampled in simulations using ff14SB+TIP3P (red), ff14SB+OPC (yellow) and ff19SB+OPC (blue). Uncertainties reflect the standard deviation of 10 independent runs. The black dots represent values reported in NMR experiments at 300 K[95]. The MD simulations were run at 300K for a total of ~96 μs.

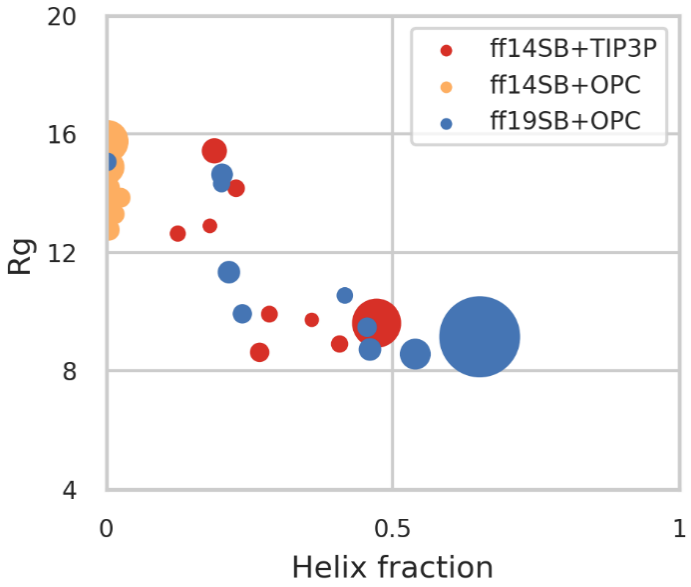


Figure 2.30 Correlation between radius of gyration (Rg) and helix fraction (calculated via DSSP) on K19 for ff14SB+TIP3P (red), ff14SB+OPC (yellow) and ff19SB+OPC (blue). Each circle represents a cluster from cluster analysis of simulation, with marker size representing cluster size. Only top 10 clusters are shown here. The cluster analysis was done on the combined trajectories from all independent runs for a given force filed + solvent model. Different from **Methods: Cluster analysis**, a cutoff of 4 Å was used for clustering to ensure a fixed average distance between clusters. All the Rg and helix fraction shown in the plot were averaged over structures within the cluster. Both Rg and DSSP calculation were performed with Cpptraj in Amber v16 software[101]. Only backbone atoms C, N, CA were used for both calculations.

Simulation of K19 with ff19SB+OPC resulted in modestly increased helical content vs. ff14SB+TIP3P, with 0.48±0.05 (central four) and ~0.29±0.01 (two near C terminus) average helicity. These values are also somewhat higher than those from experiment (~0.38 and ~0.17 respectively), but the deviation in MD corresponds to an error of only 0.24 and 0.35 kcal/mol free energy, respectively. Furthermore, uncertainties were not reported for the NMR-based data for K19, and ff19SB+OPC is in quantitative agreement with experiment for helical propensities for Lys and Ala that make up the majority of K19 (**Figure 2.25**). The simulations also agree with the trend from experiments, with the helical content falling off towards the C-terminus, with the two measured Ala in this region being less helical than the central four. Overall, we conclude that the QM-based ff19SB is in reasonable agreement with experiment when combined with an accurate solvent model, while ff14SB performs poorly with the same solvent model and relies on

cancellation of error with the less accurate TIP3P model in order to reproduce the helical content of this alanine-based peptide.

2.4.7 β-hairpin stability

We next tested whether the improvements in modeling helical content with ff19SB (and perhaps a slight overestimation of helical content) were obtained at the cost of less accurate performance on β systems. Following our previous work[2c] with ff14SB, we used CLN025[96], an engineered fast-folding hairpin that is a thermally optimized variant of Chignolin[96]. CLN025 contains four aromatic side chains, including three Tyr and one Trp. This system presents a challenge due to the relatively slow folding of β-sheets compared to the helical systems (though T-jump IR experiments[97] estimate a 100-ns folding time for CLN025). Because of the computational cost in obtaining highly precise estimates of β hairpin population in MD simulations with explicit water, we limit our testing here to a qualitative view of whether ff19SB's improved helical propensity prediction may compromise β stability. We again tested ff14SB with TIP3P and OPC, and ff19SB with OPC.

We performed four MD runs, each of 7 μs in length, at 300K (each was) starting from the NMR structure, and four additional 7 μs runs starting from a fully extended structure (56 μs total for all ff+water combinations). As measured by backbone RMSD against the NMR structure (PDBID: 2RVD[96]), folding was reversible in every simulation using each of the three combinations of the force field + solvent model (**Figure 2.31**). The histogram of RMSD values shows that both ff14SB+TIP3P and ff19SB+OPC predominantly sample the NMR structure (**Figure 2.4**). The average fraction of native population (±standard deviation) across all MD runs for ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC are 0.75±0.23, 0.33±0.19 and 0.50±0.17, respectively, compared to the experimental estimate[96] of 0.9 based on CD spectra. These populations suggest that ff14SB+TIP3P might stabilize the β-hairpin to a greater extent than the other combinations, but the differences are within the uncertainties of the populations. It is interesting that with ff14SB, MD in TIP3P appears to prefer more β-hairpin structure than with OPC. The same preference for the native structure in TIP3P was seen with K19, perhaps indicating that the weaker solute-solvent dispersion in TIP3P generally stabilizes compact structure (such as native folds) consistent with previous studies[40d, 64a, 87], rather than a specific secondary helical bias such as the K19 stability increase discussed above.

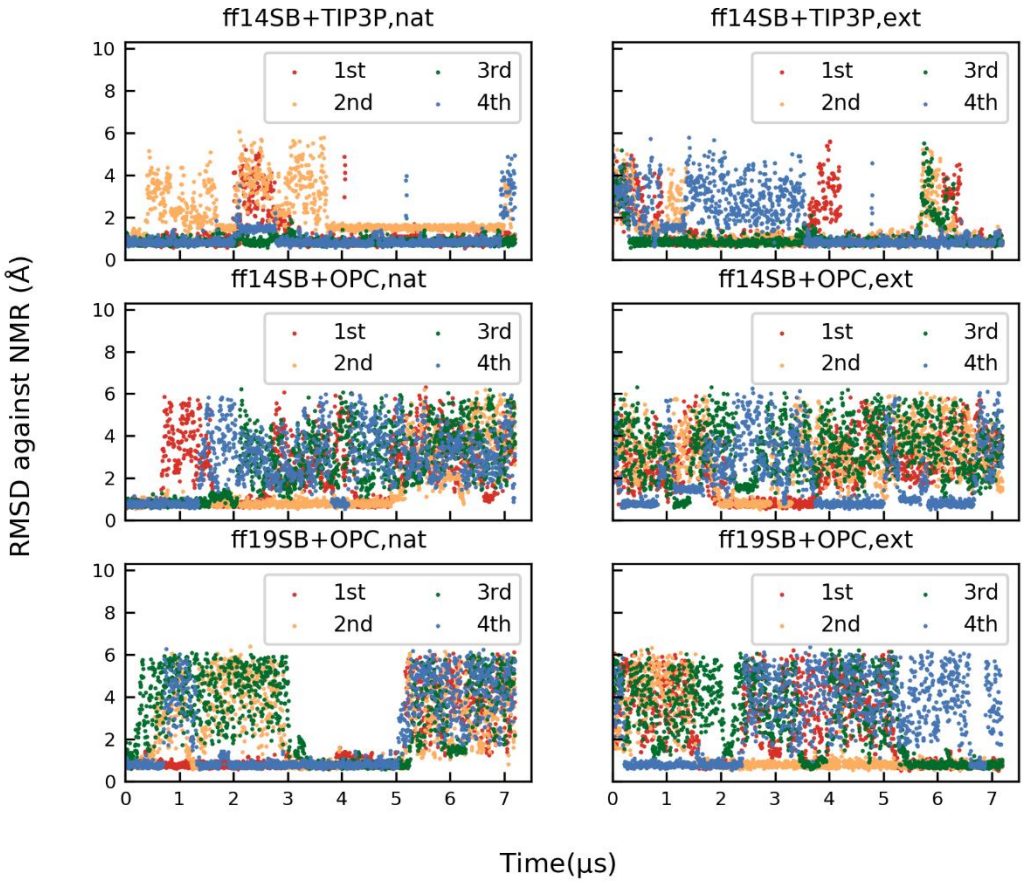


Figure 2.31 Backbone RMSD to the NMR structure (PDBID: 2RVD[96]) vs. time for the four extended (ext) and four native (nat) runs of CLN025 with ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC. The MD simulations were run at 300K for a total of ~172 μs.

2.4.8 High quality backbone dynamics vs. NMR is maintained with ff19SB

NMR S^2 order parameters reflect the internal protein dynamics that are helpful to validate MD trajectories. These internal motions need to be separated from global tumbling on time scale of pico to nanosecond. Therefore, a choice for the window size of the MD trajectory needs to be made over which S^2 values are computed and averaged, which remains challenging[125-126]. As reported in our previous work, ff14SB+TIP3P maintained ff99SB's overall good reproduction of NMR S^2 order parameters.[2c] Here, we also evaluated the ability of ff19SB to recapitulate local dynamics in well-folded proteins. As shown in **Figure 2.35**, the NMR data were reasonably reproduced by the different force field + solvent model combinations, with average absolute

difference between NMR S^2 and calculated S^2 over all amino acids close to 0.04. The overall differences were not statistically significant, however, we note some instances where all three force field + solvent models deviate from experiment and also some instances where ff19SB results are in worse agreement with experiment than is ff14SB. These residues typically have overestimated flexibility in MD as compared to NMR for Gly (smaller S^2 in MD). Examples include Gly14 in GB3, with $S^2 = 0.58$ in ff19SB+OPC and $S^2 = 0.74$ in NMR (**Table 2.14**), Gly10 in Ubiquitin with $S^2 = 0.54$ in ff19SB+OPC and $S^2 = 0.73$ in NMR (**Table 2.15**) and Ser85 in Lysozyme with $S^2 = 0.55$ in NMR and $S^2 = 0.75$ in all three force field + solvent models (**Table 2.16**). Other outliers in Lysozyme are C-terminal residues (residues 126 to 129) that are overly flexible in all three force field + solvent model combinations (**Table 2.16**). Interestingly, ff19SB+OPC sample structures with even lower RMSD against native crystal structure than either ff14SB+TIP3P or ff14SB+OPC (**Figure 2.32**, **Figure 2.33**, **Figure 2.34**).

Table 2.14 Order parameters of GB3 for NMR[136], ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC. Error bars represent the uncertainties of MD simulation, calculated from four independent MD runs.

Residue ID	NMR	ff14SB+TIP3P	ff14SB+OPC	ff19SB+OPC
3	0.83	0.85±0.01	0.87±0.01	0.80±0.03
4	0.86	0.9±0.01	0.9±0.01	0.89±0.01
5	0.88	0.91±0.01	0.91±0.01	0.91±0.01
6	0.87	0.9±0.01	0.9±0.01	0.9±0.01
7	0.83	0.9±0.01	0.9±0.01	0.89±0.01
8	0.85	0.89±0.01	0.88±0.01	0.9±0.01
9	0.83	0.74±0.04	0.78±0.01	0.85±0.02
10	0.8	0.81±0.01	0.81±0.01	0.83±0.02
11	0.76	0.76±0.01	0.78±0.01	0.8±0.01
12	0.66	0.7±0.02	0.7±0.02	0.61±0.01
13	0.76	0.71±0.03	0.76±0.01	0.77±0.02
14	0.74	0.68±0.03	0.66±0.03	0.65±0.04
15	-	-	-	-
16	0.83	0.9±0.01	0.9±0.01	0.89±0.01

17	0.8	0.86±0.01	0.86±0.01	0.88±0.01
18	0.85	0.89±0.01	0.89±0.01	0.9±0.01
19	0.77	0.84±0.01	0.84±0.01	0.83±0.01
20	0.76	0.86±0.01	0.85±0.04	0.81±0.01
21	0.84	0.89±0.01	0.9±0.01	0.9±0.01
22	0.87	0.86±0.01	0.87±0.01	0.86±0.0
23	0.92	0.91±0.01	0.91±0.01	0.91±0.01
24	0.81	0.88±0.01	0.87±0.01	0.88±0.01
25	-	-	-	-
26	0.91	0.91±0.01	0.91±0.01	0.91±0.01
27	-	-	-	-
28	0.89	0.92±0.01	0.91±0.01	0.92±0.01
29	0.9	0.9±0.01	0.9±0.01	0.91±0.01
30	0.89	0.92±0.01	0.9±0.02	0.92±0.01
31	0.91	0.93±0.01	0.92±0.01	0.93±0.01
32	0.88	0.91±0.01	0.9±0.01	0.91±0.01
33	0.9	0.91±0.01	0.91±0.01	0.92±0.01
34	0.91	0.92±0.01	0.92±0.01	0.92±0.01
35	-	-	-	-
36	0.89	0.89±0.01	0.89±0.01	0.9±0.011
37	0.83	0.82±0.01	0.81±0.01	0.84±0.01
38	0.79	0.85±0.01	0.84±0.01	0.85±0.01
39	0.84	0.83±0.01	0.83±0.01	0.82±0.01
40	0.73	0.84±0.01	0.82±0.02	0.82±0.01
41	0.5	0.66±0.02	0.67±0.03	0.56±0.04
42	0.83	0.88±0.01	0.87±0.02	0.87±0.01
43	0.86	0.88±0.01	0.88±0.01	0.88±0.01
44	0.86	0.9±0.01	0.89±0.02	0.9±0.01
45	0.82	0.89±0.01	0.88±0.01	0.89±0.01
46	0.85	0.87±0.01	0.87±0.01	0.87±0.01

47	0.81	0.87±0.01	0.88±0.01	0.89±0.01
48	0.75	0.88±0.01	0.88±0.01	0.89±0.01
49	0.82	0.83±0.01	0.83±0.01	0.86±0.01
50	0.88	0.88±0.01	0.88±0.01	0.88±0.01
51	0.86	0.89±0.01	0.89±0.01	0.91±0.01
52	0.87	0.92±0.01	0.92±0.01	0.92±0.01
53	0.84	0.92±0.01	0.92±0.01	0.92±0.01
54	0.89	0.91±0.01	0.91±0.01	0.91±0.01
55	0.82	0.89±0.01	0.89±0.01	0.9±0.01
56	0.85	0.88±0.01	0.86±0.03	0.88±0.01

Table 2.15 Order parameters of Ubiquitin for NMR[137], ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC. Error bars represent the uncertainties of MD simulation, calculated from four independent MD runs.

Residue ID	NMR	ff14SB+TIP3P	ff14SB+OPC	ff19SB+OPC
2	0.84	0.89±0.01	0.89±0.01	0.88±0.01
3	0.88	0.91±0.01	0.91±0.01	0.9±0.02
4	0.89	0.91±0.01	0.91±0.01	0.89±0.01
5	0.82	0.91±0.0	0.91±0.01	0.9±0.01
6	0.85	0.89±0.01	0.9±0.01	0.9±0.01
7	0.86	0.81±0.02	0.86±0.01	0.81±0.01
8	0.77	0.82±0.01	0.85±0.01	0.82±0.01
9	0.73	0.76±0.02	0.77±0.01	0.71±0.01
10	0.73	0.72±0.02	0.69±0.01	0.55±0.02
11	0.71	0.68±0.02	0.73±0.04	0.68±0.05
12	0.76	0.8±0.01	0.82±0.02	0.79±0.02
13	0.84	0.83±0.01	0.86±0.01	0.82±0.02
14	0.82	0.86±0.01	0.86±0.01	0.83±0.01
15	0.82	0.9±0.01	0.91±0.01	0.89±0.01
16	0.78	0.85±0.01	0.86±0.01	0.84±0.02
17	0.88	0.89±0.01	0.9±0.01	0.89±0.01

18	0.86	0.88±0.01	0.9±0.01	0.87±0.01
19	-	-	-	-
20	0.84	0.88±0.01	0.89±0.01	0.87±0.01
21	0.9	0.9±0.01	0.91±0.01	0.86±0.02
22	0.85	0.89±0.01	0.89±0.0	0.87±0.01
23	-	-	-	-
24	-	-	-	-
25	-	-	-	-
26	0.85	0.91±0.01	0.92±0.01	0.92±0.01
27	0.91	0.93±0.01	0.93±0.01	0.93±0.01
28	0.9	0.92±0.01	0.92±0.01	0.92±0.01
29	0.89	0.91±0.01	0.92±0.01	0.92±0.01
30	0.88	0.91±0.01	0.91±0.01	0.92±0.01
31	-	-	-	-
32	0.89	0.9±0.01	0.91±0.01	0.91±0.01
33	0.85	0.78±0.01	0.77±0.01	0.85±0.01
34	0.84	0.85±0.01	0.85±0.01	0.87±0.01
35	0.87	0.87±0.01	0.88±0.01	0.88±0.01
36	0.79	0.8±0.01	0.8±0.01	0.78±0.01
37	-	-	-	-
38	-	-	-	-
39	0.85	0.88±0.01	0.89±0.01	0.89±0.01
40	0.86	0.87±0.01	0.87±0.01	0.87±0.01
41	0.85	0.84±0.01	0.84±0.01	0.83±0.01
42	0.83	0.91±0.0	0.9±0.01	0.87±0.02
43	0.82	0.88±0.01	0.89±0.01	0.87±0.02
44	0.83	0.91±0.01	0.92±0.01	0.91±0.01
45	0.87	0.92±0.01	0.92±0.01	0.91±0.01
46	0.83	0.87±0.01	0.88±0.01	0.86±0.01
47	0.82	0.84±0.01	0.84±0.01	0.82±0.01

48	0.84	0.79±0.01	0.76±0.01	0.79±0.02
49	0.75	0.87±0.01	0.88±0.01	0.86±0.01
50	0.83	0.86±0.01	0.87±0.01	0.85±0.02
51	0.81	0.86±0.01	0.88±0.01	0.85±0.02
52	0.8	0.87±0.01	0.88±0.01	0.87±0.02
53	-	-	-	-
54	0.87	0.83±0.01	0.79±0.01	0.76±0.04
55	0.87	0.89±0.01	0.89±0.01	0.85±0.01
56	0.9	0.92±0.01	0.92±0.01	0.92±0.01
57	0.87	0.9±0.01	0.91±0.01	0.91±0.01
58	0.89	0.9±0.01	0.9±0.01	0.89±0.01
59	0.86	0.86±0.01	0.86±0.01	0.87±0.01
60	0.88	0.89±0.01	0.89±0.01	0.87±0.01
61	0.85	0.89±0.01	0.89±0.01	0.88±0.01
62	0.7	0.84±0.01	0.83±0.01	0.8±0.01
63	0.82	0.89±0.01	0.89±0.01	0.88±0.01
64	0.87	0.91±0.01	0.91±0.01	0.9±0.02
65	0.87	0.86±0.01	0.87±0.01	0. 86±0.01
66	0.83	0.88±0.01	0.88±0.01	0.86±0.04
67	0.84	0.89±0.01	0.89±0.01	0.88±0.02
68	0.87	0.89±0.01	0.9±0.01	0.88±0.01
69	0.84	0.84±0.01	0.87±0.01	0.86±0.01
70	0.91	0.89±0.01	0.9±0.01	0.88±0.01
71	0.79	0.87±0.01	0.87±0.01	0.85±0.01
72	-	-	-	-
73	0.56	0.64±0.06	0.65±0.02	0.75±0.07

Table 2.16 Order parameters of Lysozyme for NMR[138], ff14SB+TIP3P, ff14SB+OPC and ff19SB+OPC. Error bars represent the uncertainties of MD simulation, calculated from four independent MD runs.

Residue ID	NMR	ff14SB+TIP3P	ff14SB+OPC	ff19SB+OPC

2	0.83	0.82±0.01	0.81±0.03	0.84±0.01
3	0.83	0.84±0.01	0.83±0.01	0.86±0.01
4	0.83	0.85±0.01	0.84±0.01	0.87±0.01
5	0.85	0.87±0.01	0.88±0.01	0.89±0.01
6	0.86	0.87±0.01	0.88±0.01	0.90±0.01
7	0.88	0.88±0.01	0.88±0.01	0.88±0.01
8	0.89	0.91±0.01	0.91±0.01	0.92±0.01
9	0.93	0.91±0.01	0.92±0.01	0.92±0.01
10	0.89	0.91±0.01	0.91±0.01	0.92±0.01
11	0.89	0.91±0.01	0.91±0.01	0.91±0.01
12	0.91	0.92±0.01	0.92±0.01	0.92±0.01
13	0.92	0.91±0.01	0.91±0.01	0.92±0.01
14	0.82	0.91±0.01	0.91±0.01	0.91±0.01
15	0.84	0.88±0.01	0.87±0.01	0.87±0.01
16	-	-	-	-
17	0.89	0.87±0.01	0.85±0.02	0.83±0.02
18	0.86	0.84±0.01	0.84±0.01	0.78±0.01
19	0.84	0.84±0.02	0.84±0.01	0.83±0.01
20	0.85	0.89±0.01	0.88±0.01	0.87±0.01
21	0.89	0.9±0.01	0.89±0.01	0.9±0.01
22	0.99	0.86±0.01	0.86±0.01	0.86±0.01
23	0.88	0.84±0.01	0.82±0.01	0.82±0.01
24	0.89	0.88±0.01	0.86±0.02	0.9±0.01
25	0.87	0.9±0.01	0.9±0.01	0.91±0.01
26	0.91	0.91±0.01	0.9±0.01	0.92±0.01
27	0.94	0.9±0.01	0.91±0.01	0.91±0.01
28	0.87	0.9±0.01	0.9±0.01	0.9±0.01
29	0.9	0.91±0.01	0.91±0.01	0.92±0.01
30	-	-	-	-

31	0.93	0.91±0.01	0.91±0.01	0.92±0.01
32	0.94	0.91±0.01	0.91±0.01	0.92±0.01
33	0.91	0.9±0.01	0.9±0.01	0.91±0.01
34	0.92	0.88±0.01	0.88±0.01	0.91±0.0
35	0.88	0.86±0.01	0.86±0.01	0.89±0.0
36	0.86	0.85±0.01	0.85±0.01	0.81±0.01
37	0.96	0.89±0.01	0.89±0.01	0.9±0.01
38	0.9	0.91±0.01	0.91±0.01	0.89±0.01
39	0.89	0.88±0.01	0.88±0.01	0.88±0.01
40	0.91	0.9±0.01	0.9±0.01	0.91±0.01
41	0.86	0.89±0.01	0.89±0.01	0.9±0.01
42	0.87	0.87±0.01	0.88±0.01	0.88±0.01
43	0.83	0.9±0.01	0.9±0.01	0.89±0.01
44	0.83	0.88±0.01	0.88±0.01	0.86±0.01
45	0.78	0.85±0.01	0.86±0.01	0.86±0.01
46	0.83	0.81±0.01	0.81±0.01	0.75±0.01
47	0.78	0.84±0.01	0.85±0.01	0.81±0.01
48	0.77	0.81±0.01	0.82±0.01	0.75±0.01
49	0.82	0.83±0.01	0.84±0.01	0.81±0.01
50	-	-	-	-
51	0.89	0.89±0.01	0.9±0.01	0.9±0.01
52	0.89	0.91±0.01	0.92±0.01	0.91±0.01
53	0.87	0.92±0.01	0.92±0.01	0.92±0.01
54	0.91	0.91±0.01	0.91±0.01	0.91±0.01
55	0.94	0.92±0.01	0.92±0.01	0.92±0.01
56	0.92	0.92±0.01	0.92±0.01	0.93±0.01
57	0.94	0.91±0.01	0.91±0.01	0.92±0.01
58	0.9	0.88±0.01	0.87±0.01	0.88±0.01
59	0.91	0.9±0.01	0.9±0.01	0.89±0.01
60	0.93	0.91±0.01	0.91±0.01	0.92±0.01

61	0.95	0.88±0.01	0.88±0.01	0.87±0.02
62	0.85	0.84±0.01	0.87±0.01	0.86±0.01
63	0.9	0.87±0.01	0.87±0.01	0.86±0.02
64	0.91	0.88±0.01	0.87±0.02	0.86±0.01
65	0.86	0.85±0.01	0.87±0.01	0.86±0.02
66	0.89	0.85±0.01	0.82±0.07	0.86±0.01
67	0.85	0.85±0.01	0.85±0.01	0.86±0.01
68	0.78	0.72±0.02	0.77±0.01	0.81±0.04
69	0.76	0.83±0.01	0.82±0.03	0.67±0.02
70	-	-	-	-
71	0.72	0.76±0.03	0.77±0.04	0.77±0.03
72	0.76	0.75±0.05	0.76±0.02	0.77±0.02
73	0.88	0.75±0.02	0.76±0.03	0.77±0.05
74	0.87	0.86±0.02	0.81±0.06	0.84±0.02
75	0.94	0.91±0.01	0.87±0.06	0.91±0.01
76	0.92	0.88±0.01	0.84±0.06	0.89±0.01
77	0.9	0.9±0.01	0.89±0.01	0.9±0.01
78	0.91	0.75±0.03	0.84±0.01	0.9±0.01
79	-	-	-	-
80	0.91	0.9±0.01	0.9±0.01	0.91±0.01
81	0.86	0.9±0.01	0.9±0.01	0.9±0.01
82	0.88	0.89±0.01	0.89±0.01	0.89±0.01
83	0.83	0.89±0.01	0.9±0.01	0.9±0.01
84	0.83	0.9±0.01	0.91±0.01	0.9±0.01
85	0.55	0.75±0.02	0.76±0.04	0.76±0.05
86	0.8	0.86±0.01	0.88±0.01	0.86±0.01
87	0.8	0.85±0.01	0.86±0.01	0.82±0.01
88	0.8	0.88±0.01	0.88±0.01	0.88±0.01
89	0.92	0.92±0.01	0.92±0.01	0.92±0.01
90	0.91	0.92±0.01	0.92±0.01	0.92±0.01

91	0.85	0.91±0.01	0.91±0.01	0.91±0.01
92	0.93	0.92±0.01	0.92±0.01	0.92±0.01
93	0.93	0.92±0.01	0.92±0.01	0.92±0.01
94	0.92	0.91±0.01	0.91±0.01	0.92±0.01
95	0.92	0.92±0.01	0.92±0.01	0.92±0.01
96	0.92	0.93±0.01	0.92±0.01	0.93±0.01
97	0.94	0.9±0.01	0.9±0.01	0.91±0.01
98	0.92	0.91±0.01	0.91±0.01	0.92±0.01
99	-	-	-	-
100	0.89	0.87±0.02	0.88±0.01	0.87±0.01
101	0.85	0.83±0.04	0.84±0.04	0.81±0.02
102	0.72	0.75±0.04	0.75±0.02	0.68±0.01
103	0.63	0.7±0.04	0.72±0.04	0.60±0.02
104	0.81	0.77±0.07	0.82±0.02	0.79±0.02
105	0.88	0.84±0.03	0.84±0.02	0.85±0.01
106	0.96	0.88±0.01	0.88±0.01	0.9±0.01
107	0.91	0.86±0.01	0.87±0.01	0.88±0.01
108	0.84	0.84±0.01	0.84±0.01	0.88±0.01
109	0.85	0.85±0.01	0.87±0.01	0.88±0.01
110	-	-	-	-
111	0.84	0.83±0.02	0.83±0.02	0.86±0.01
112	0.89	0.89±0.01	0.9±0.01	0.92±0.01
113	0.89	0.83±0.01	0.83±0.01	0.87±0.0
114	0.87	0.76±0.03	0.74±0.01	0.75±0.01
115	0.79	0.77±0.01	0.77±0.02	0.77±0.02
116	0.84	0.8±0.03	0.78±0.02	0.73±0.05
117	0.81	0.61±0.10	0.65±0.07	0.75±0.06
118	0.72	0.59±0.04	0.6±0.05	0.68±0.03
119	0.8	0.76±0.04	0.75±0.02	0.76±0.04
120	0.8	0.75±0.03	0.71±0.05	0.77±0.03

121	0.91	0.86±0.02	0.86±0.01	0.88±0.01
122	0.92	0.88±0.01	0.87±0.02	0.89±0.01
123	0.9	0.85±0.02	0.84±0.01	0.87±0.01
124	0.9	0.86±0.01	0.85±0.01	0.87±0.01
125	0.87	0.84±0.03	0.8±0.02	0.82±0.02
126	0.82	0.68±0.09	0.69±0.02	0.75±0.01
127	0.77	0.65±0.08	0.59±0.05	0.54±0.06
128	0.76	0.61±0.13	0.59±0.06	0.65±0.1
129	0.6	0.45±0.17	0.43±0.02	0.52±0.02

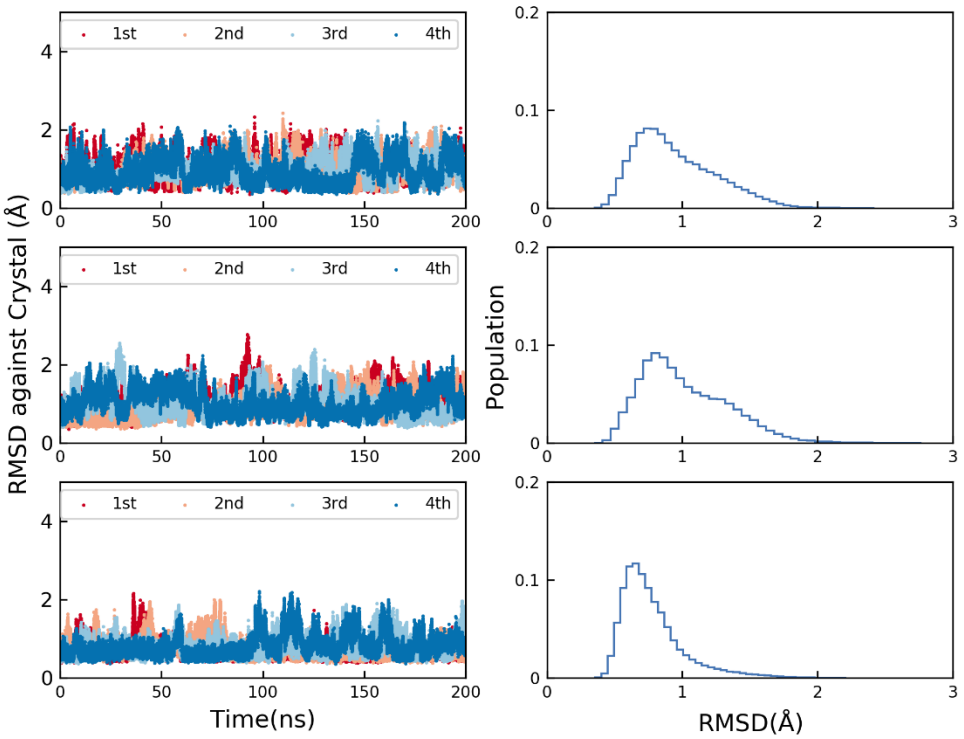


Figure 2.32 Backbone RMSD against crystal structure versus time (left) and histogram on RMSD (right) for GB3 (PDBID: 1P7E) for (top) ff14SB+TIP3P, (middle) ff14SB+OPC and (bottom) ff19SB+OPC. Four independent runs starting from different initial velocities were performed. C, N and CA atoms were used for RMSD analysis. Four runs were combined in the histogram (right).

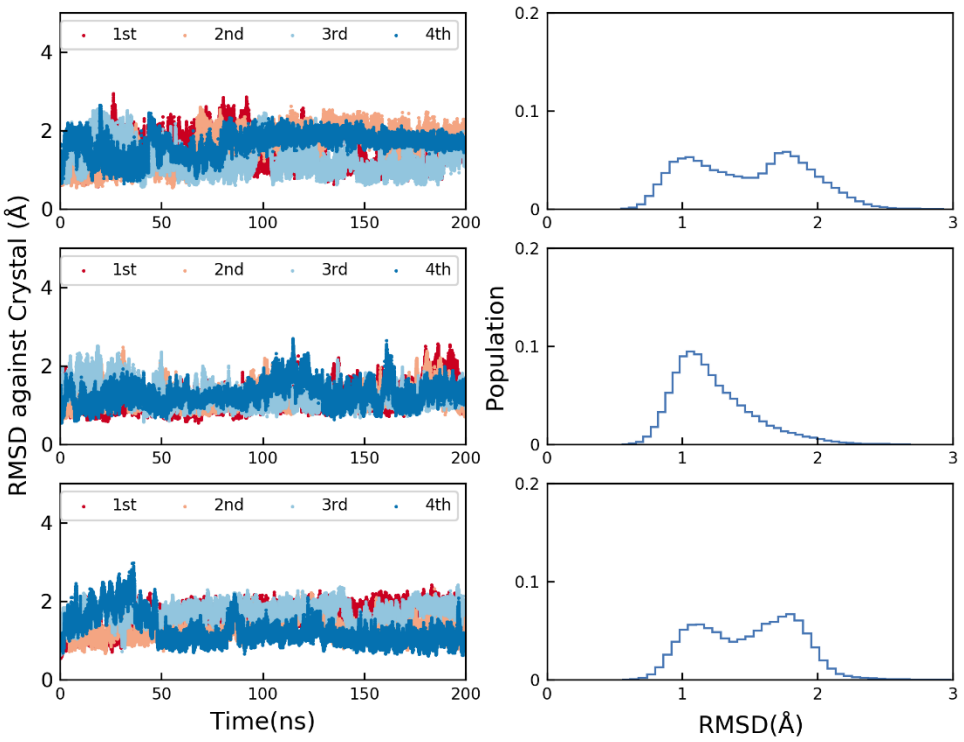


Figure 2.33 Backbone RMSD against crystal structure versus time (left) and histogram on RMSD (right) for Ubiquitin (PDBID: 1UBQ) for (top) ff14SB+TIP3P, (middle) ff14SB+OPC and (bottom) ff19SB+OPC. Four independent runs starting from different initial velocities were performed. C, N and CA atoms were used for RMSD analysis. Four runs were combined in the histogram (right).

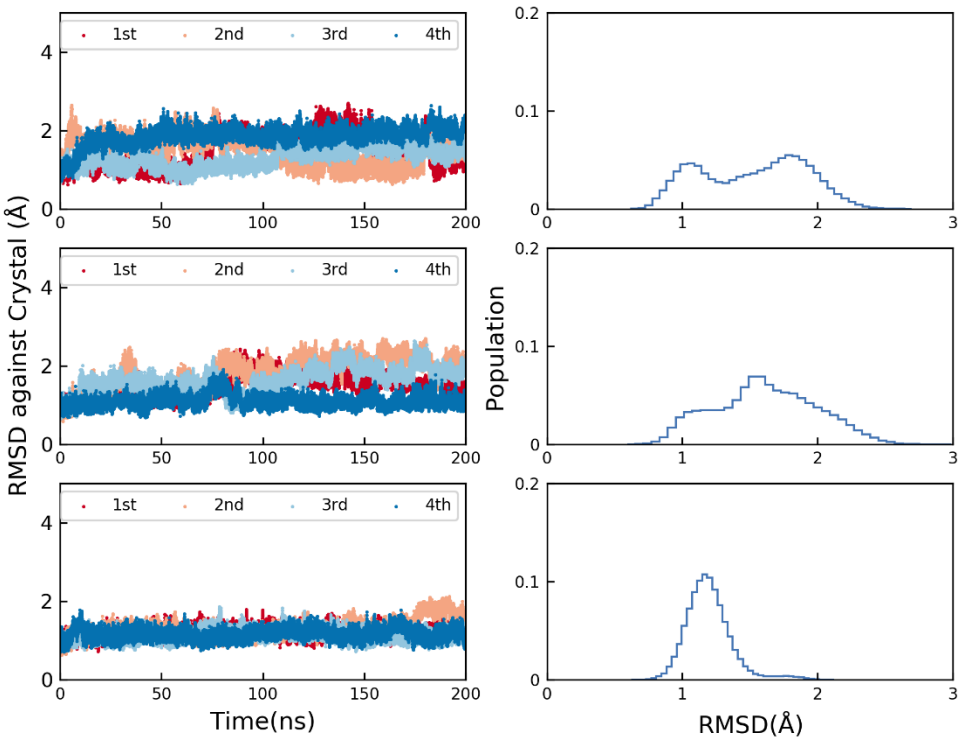


Figure 2.34 Backbone RMSD against crystal structure versus time (left) and histogram on RMSD (right) for Lysozyme (PDBID: 6LYT) for (top) ff14SB+TIP3P, (middle) ff14SB+OPC and (bottom) ff19SB+OPC. Four independent runs starting from different initial velocities were performed. C, N and CA atoms were used for RMSD analysis. Four runs were combined in the histogram (right).

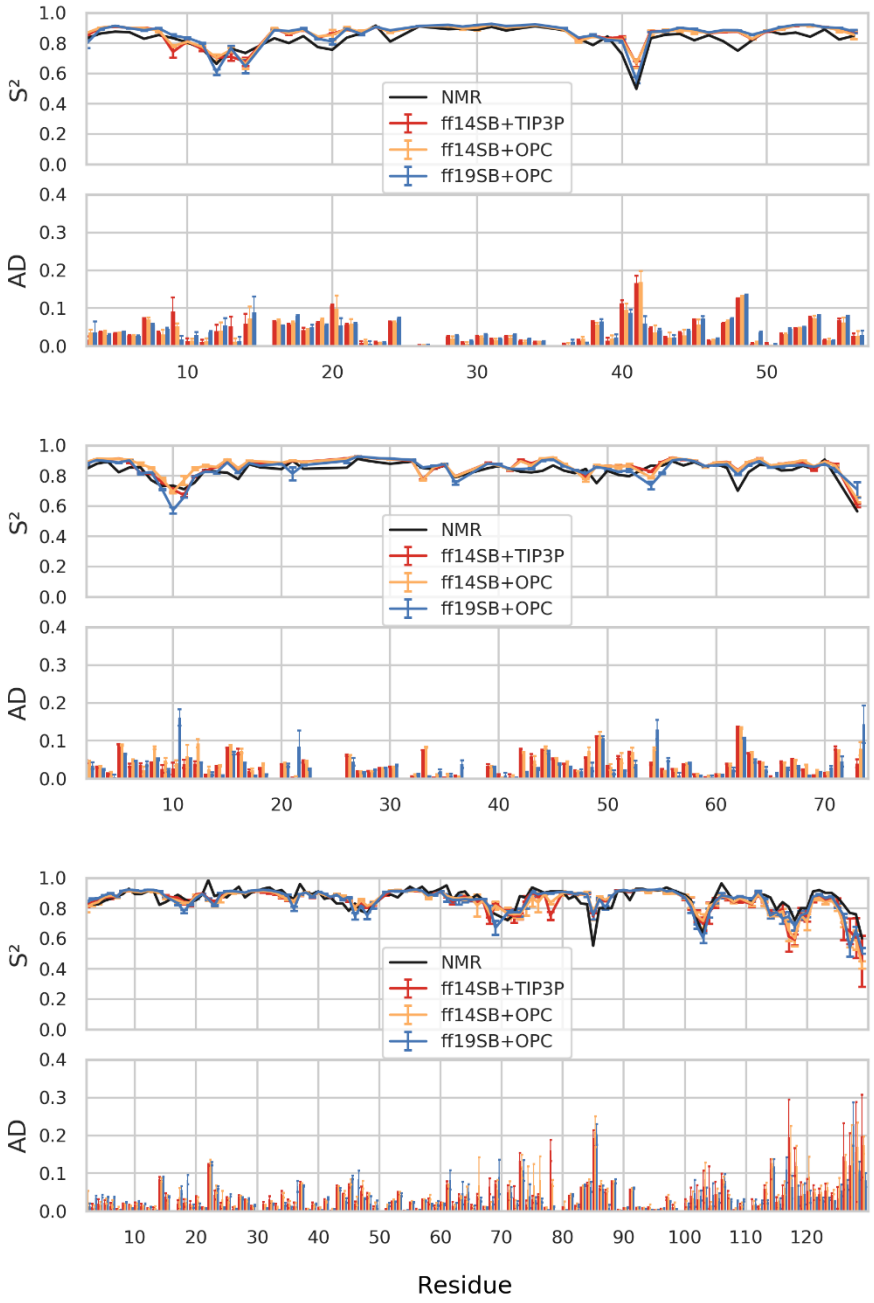


Figure 2.35 Per-residue order parameters (S^2) from NMR compared to simulations using ff14SB+TIP3P (red), ff14SB+OPC (yellow) and ff19SB+OPC (blue) of (top) GB3[136], (middle) Ubiquitin[137] and (bottom) Lysozyme[138]. AD is the absolute difference between NMR and MD simulation. MAD is mean absolute difference over all residues. For each subplot, error bars represent the standard deviation from four independent runs. Some residues are missing experimental values as indicated in the original NMR papers[136-138]. The MD simulations were run at 300K for a total of ~1.8 μs.

In GB3, Gly14 was reported to have a high S^2 (0.74) using NMR, likely due to its intermolecular hydrogen bond in a β-sheet secondary structure. However, Gly14 is similarly more flexible with ff19SB+OPC (0.65±0.04), ff14SB+TIP3P (0.68±0.03) and ff14SB+OPC (0.66±0.03). This may not reflect problems in ff19SB Gly parameters since this trend is reversed for Gly41 in the loop region connecting a β-strand to an α-helix. S^2 from NMR is quite low for Gly41 (0.50) due to loop flexibility, and this flexibility is reproduced much better with ff19SB+OPC (0.56±0.04) than ff14SB+TIP3P (0.66±0.02) and ff14SB+OPC (0.67±0.03).

In Ubiquitin, Gly10 flexibility is overestimated in ff19SB+OPC (0.55±0.02), but not in ff14SB+TIP3P (0.72±0.02) and ff14SB+OPC (0.69±0.01) compared to NMR (0.73). Except for the slightly worsened performance on Gly10, ff19SB+OPC yields the best overall agreement with NMR compared to ff14SB with either TIP3P or OPC solvent model.

In Lysozyme, Ser85 lies in a loop region connecting two α helices, and is overly rigid with all three simulation models (~0.75 in MD vs 0.55 in NMR). However, Ser85 backbone (ϕ/ψ) and side chain (χ1/χ2) sampling in all three force field + solvent model combinations reproduces that seen in the crystal structure.

In spite of subtle disagreements with NMR, we concluded that ff19SB generally maintained the overall performance of ff14SB and ff99SB in order parameter reproduction, with a few outliers that do not appear to follow any systematic trend that could be attributed to the CMAPs.

2.5 Conclusion

In the updated ff19SB (with the "SB" models indicating Stony Brook) protein force field presented here, we have developed new backbone dihedral parameters with amino-acid specific CMAP functions. We trained the parameters to match solution phase QM data using full 2D φ/ψ scans, instead of the gas-phase minima used for training uncoupled φ and ψ cosine terms in ff99SB. Use of energies calculated from QM in solution provides better consistency with the pre-polarized partial atomic charges used by the MM model, as compared to gas-phase energies that were used

previously. Fitting of dihedral corrections against QM in solution also allows the model to incorporate (to some extent) conformation-dependent polarization energy that is not present explicitly in a fixed-charge MM model such as the one used here.

A total of 16 CMAPs were fit, with applicability to all amino acids using a grouping approach based on side chain size, branching and polarity. Leu was used as a general model for other amino acids, in contrast to Ala that has traditionally been used as a protein backbone model. We also investigated whether CMAP functions fit using a single side chain rotamer could remain accurate for other rotamer states, and found good transferability as measured by the ability of the model to reproduce rotamer-dependent differences in Ramachandran space QM energetics and PDB-based statistics.

One possible weakness to our approach was the use of simple implicit water models during training, such as the GB model in the MM component. Older GB models exhibit secondary structure biases for longer peptides[139], but here we have used our GBneck2 model[92] that much more accurately reproduces secondary structure preferences. Furthermore, we have shown that the solvation energy of dipeptides (which we used here for the CMAP training in GB) is largely insensitive to specific GB model used[140]. Nevertheless, our use of GB during training could be a limitation, which is one reason we carried out extensive testing here using a variety of fully explicit water models.

We performed a total of ~6 milliseconds MD simulations in explicit solvent to extensively validate ff19SB against experiments. The results show that our new FF more accurately reproduces amino-acid specific NMR properties such as scalar coupling and helical propensity, as well as structure and stability of a Baldwin-type helical peptide and a small hairpin. Folded proteins show good agreement with NMR S^2 order parameters, and modestly improved RMSD values as compared to ff14SB.

We make the important observation that the performance of the QM-based ff19SB model improves as the quality of the water model is improved (going from TIP3P to TIP4P-Ew to OPC), suggesting lack of fortuitous cancellation of error with a particular water model, and that the water model is likely the limiting factor in these comparisons of ff19SB to experiment. Currently, our best results are obtained using ff19SB with OPC water, and we recommend that combination. Biomolecular force fields such as ff19SB that are not tied to a specific water model through empirical adjustment will be in a stronger position to take advantage of future, better-quality water

models. In contrast, use of a better model for water does not lead to improved match with experiment for ff14SB, supporting that both a good water model and good protein force field are needed for an accurate simulation. We also conclude that weaker solute-solvent dispersion in TIP3P not only leads to overly compact unfolded states as has been reported previously, but also overstabilizes native helical and hairpin structures as compared to OPC.

If water models can be sufficiently improved, there is in principle no need for specialized "IDP" force fields, as suggested in recent work[40d] by Robustelli et al. Our belief is that physics-based protein FFs trained against short peptides should be quite capable of modeling IDPs and unfolded ensembles, which are more similar to the peptide training data than are folded proteins. Amber's OPC 4-point water model not only better reproduces liquid water properties as compared to most other models[66], but IDP simulations with OPC result in much less compact ensembles as compared to simulations using the same FFs in older water models.[67] This provides additional evidence that the current problems with modeling IDPs are likely to be related to the water models, and further improvement of physics-based protein FFs is warranted, independent of water model development going on in parallel. While the studies here of flexible peptides using ff19SB+OPC are promising, future studies using this combination for IDPs will be carried out in the future.

Chapter 3

Further investigate the physical cause of errors that are corrected by CMAP in ff19SB

"All models are wrong, but some are useful"

--- George Box

3.1 Introduction

As discussed in the previous two chapters, many approximations are made in fitting FF parameters. In ff19SB, we revisited several weaknesses that may be dominant factors limiting accuracy of classical AMBER force field. To overcome the weaknesses, we developed amino-acid specific protein backbone dihedral parameters by employing CMAP functions and utilized quantum mechanics energy in solution as reference data. This strategy overcomes the major issues in previous AMBER force field such as ff99SB[2b] and ff14SB[2c] as reflected by extensive test in MD simulations (Results and Discussion section in **Chapter 2**). However, the physical motivation behind dihedral corrections is that the rest of the FF is purely classical, and therefore lacks quantum effects such as the increased energy barrier for rotation around a double bond. In practice, these

corrections are used broadly to empirically optimize force fields during training, accounting for quantum effects as well as other weaknesses in the simple model, such as lack of conformation-dependent polarization that could impact electrostatic interaction profiles, or even to remedy lack of agreement with experiments. The need for dihedral parameters such as CMAP could be reduced if the accuracy of the rest of the force field is greatly improved. The transferability of FF can be improved if dihedral parameters are not responsible for errors in the rest of the force field.

Nonbonded interactions including electrostatics and van der Waals (vdW) interactions have well-understood physical origins (see **Chapter 1**). They are represented with simple functional forms in traditional force fields, for instance point charges for electrostatics and Lennard-Jones 12-6 functional for vdW for the sake of computational efficiency. Increasingly, these simple forms are shown to be insufficient to fully describe the underlying physics; hence it is impossible for their parameters to be both accurate and transferable. What is even worse, the error arisen from non-bonded terms might be unintentionally compensated by dihedral terms in practice. As computer power continues to grow and new hardware (such as GPU) to emerge, computing cost will become less an obstacle to more accurate modelling. Historical compromises must be revisited. It is thus crucial to understand the error of nonbonded interactions in classical model and provide insights for developing next generation force field with higher accuracy and transferability.

A few examples related to nonbonded errors were already mentioned in **Chapter 2**. As shown in **Figure 2.12** and **Figure 2.16**, the Val CMAP trained against *trans* rotamer gives near-zero error for *trans* structures, but ~0.89 kcal/mol for *gauche(-)* ones (see **2.4.1 Backbone rotational energies in ff19SB compared to ff14SB**). The transferability issue arises because of the well-established physical mechanism. For Val, some side chain rotamers clash with backbone carbonyl oxygen atoms in the helical conformation, reducing overall helical propensity. The error gets larger when the inaccurate MM short-range repulsion (from *trans* rotamer) is totally folded into the CMAP correction (because CMAP fitting is "perfect"), and the CMAP is applied to backbone of *gauche(-)* rotamer where the clash is not present. This is the main transferability issue of ff19SB. Due to the fact that Val has non-polar β-branched side chain, the short-range vdW interaction errors might be most likely the cause. Another example is Ser, theoretically, the disagreement between QM and MM in intramolecular hydrogen-bond (H-bond) can be perfectly corrected by CMAP for a particular rotamer, but the CMAP becomes imperfect when applied to

another rotamer Ser where the H-bond is not present. The physical problem will not be fixed if the non-bonded errors are corrected by dihedral terms. It is thus advantageous to know the physical reason behind the overall FF errors and this is conducive to understanding FF limitation and further improving the model.

In this chapter, we investigate the fundamental source of errors that are most likely corrected in ff19SB with the use of amino-acid specific CMAP functions. We will prioritize the major FF issues by quantifying the magnitude of errors from various FF terms. The assumptions we made in the error investigation include: 1. The lack of hydrogen bond in training will worsen the transferability of CMAP parameters to large biomolecules. 2. Retaining the old empirical 1-4 scaling factor might limit the accuracy of short-range interactions. 3. Retaining the old RESP charges[8] especially for charged amino acids might cause problem in hydrogen bond formation. 4. Using GB model in CMAP training might only partially cancel out the solvation energy and leave the solvation energy in dihedral parameters. When the dihedral parameters are applied in MD that include models for solvation environment, the solvation energy is over-counted. The last one is aimed to test the assumption made in ff19SB training (making dihedral fitting and partial charge fitting consistent) and its effect on resulted ff19SB parameters, rather than identifying errors corrected by ff19SB.

Certainly, the errors of force field can be attributed to other terms as well such as bonds, angles, improper dihedrals, etc. However, we believe that the non-bonded parameters are the most problematic and most important, because they are determinant in secondary structure prediction in MD simulation. Specifically, the 1-4 scaling factor is purely empirical and haven't been systematically revisited for decades except a few but limited efforts[141]. The partial charges are also retained from models developed decades ago[8]. After the errors are quantified and prioritized, a systematic refitting of certain term such as dihedral is crucial for future force field development. Because dihedral potentials are often fit to make up the total energy profile, by improving other terms such as non-bonded, we can essentially improve dihedral potentials as well. In this manner, the force field can be fundamentally improved with more physically relevant terms and be more transferable.

3.2 Methods

3.2.1 Geometry scanning and energy calculation on Ala tetrapeptide

Acetyl and N-methyl capped tetrapeptide of alanine (Ace-Ala-Ala-Ala-Nme) was adopted. The first and second Ala were restrained onto α conformation (φ=-60° and ψ=-45°) or ppII conformation (φ=-60° and ψ=150°). For either α or ppII, backbone geometry scans were performed on the third Ala to generate multiple structures. The 2D scan was performed on φ and ψ dihedrals over ranges of -180° to 165° with an interval of 15°. All scans were carried out via the LEaP module of AmberTools in Amber v16 software[101].

Tetrapeptide structures were minimized after geometry scanning including restraints on φ and ψ values with harmonic force constant of 1000 *kcal mol^{-1} rad^{-2}*. MM optimization and energy calculations were performed using ff14SB[2c] and GBneck2[92] implicit solvent model with the mbondi3 radii set[92] for polar solvation and SASA-based nonpolar solvation (default 0.005 kcal mol^{-1} Å^{-1} surface tension was adopted)[103]. Structures were minimized for a maximum of 10,000 cycles in ff14SB+GBSA with no cutoff on non-bonded interactions. Steepest descent was employed for the first 10 cycles in the minimization and conjugate gradient for the following cycles. Single point energies were calculated for the MM-optimized structures using ff14SB00 (**Table 2.2**) + GBSA. The convergence criterion for energy gradient is when the root-mean-square of the Cartesian elements of the gradient is less than 10^{-4} kcal mol^{-1} Å^{-1}.

QM calculations were performed with Gaussian 09[142]. Geometry optimizations and single point energy calculations were performed at the M05-2X/6-311G**/SMD level of theory[108]. Grimme's dispersion correction with the original D3 damping function was used to correct for long-range dispersion. This combination was adopted in ff19SB and is proved to be accurate (see accuracy for each QM methods in **Figure *2.5***). The solvation environment was represented as a self-consistent reaction field, with exterior dielectric set to default 78.3553, using SMD[106] with consideration of both polar and nonpolar solvation energy components. Very tight optimization convergence criterion was used to generate data for fitting. To maintain the structure on the φ/ψ grid, we followed the protocol of "**2.3.7 QM optimization and energy calculations**". The first and second Ala were also restrained to the values from the structures taken from the last step of

MM optimization throughout the QM calculations. Following "**2.3.5 CMAP fitting**", a CMAP was then fit using the tetrapeptide data.

3.2.2 1-4 scaling factor scanning

Acetyl and N-methyl capped dipeptide of Val (Ace-Val-Nme) was used in the 1-4 scaling factor scanning. The conformations were generated following "**2.3.2 Geometry scanning**" and QM energies were calculated following "**2.3.7 QM optimization and energy calculations**". For each conformation of the φ/ψ scanned grid, either in *trans* or *gauche(-)* rotamer, a series of MM calculations were performed with scanned values on 1-4 electrostatics (14scee) and 1-4 vdW (14scnb) over ranges of 0.6 to 3.0 with an interval of 0.2 (13 points in each dimension). In this manner, for each of the 576 conformations on the φ/ψ grid, a following up 13*13 MM calculations were performed. MM calculation is comprised of optimization and energy calculation. Geometry minimization was performed using GBSA and original ff14SB with restraints on φ and ψ dihedrals with harmonic force constant of 1000 $kcal\ mol^{-1}\ rad^{-2}$. Energy calculation was performed using GBSA and modified ff14SB00 (with modification to 14 scee and 14 scnb). The original ff14SB00 is defined in **Table *2.2***. The average relative energy error (REE) between QM and MM for each 14scee and 14scnb combination is calculated as:

$$avgREE = \frac{2}{N(N-1)} \sum_i \sum_{j \neq i} \left| QM_i - QM_j - (MM_i - MM_j) \right| \quad (3.1),$$

where N is the number of conformations, QM is the QM+SMD energy and MM is ff14SB00+GBSA plus the 1-4 energy calculated using adjustable 14scee and 14scnb.

3.2.3 Refitting of atomic partial charges

The partial charges for acetyl and N-methyl capped dipeptide of Asp and Glu (Ace-X-Nme) were refitted. We took the charges from Best et al[32] derived by performing the following procedures. An approach similar to that adopted by Kollman and co-workers in deriving the original Amber ff94 charge set[23] was used, except that the charges on the amide N, H, C, O were fixed to have the same (fixed) values as all the other residues, i.e., 0.4157e, 0.2719e, 0.5973e, and 0.5679e, respectively. The RESP method[8] was used to fit the charges to electrostatic potentials derived from amino acid dipeptides, with several conformations being used in the fit, selected to represent both α and β conformations. The backbone dihedral angles were set to φ=-165°, ψ=165°

for β conformation and φ=-60°, ψ=-45° for α conformation. For each backbone conformation, three sets of side-chain torsion angles were chosen, using the three most populated conformers from the rotamer library of Lovell et al.[74], and energy minimized within the Amber ff99SB force field, with strong restraints keeping the side chain and backbone dihedral values close to the starting values. Conformations in which the side chain was hydrogen-bonded to the backbone were eliminated. An electrostatic potential was obtained at the HF/6-31G* level of theory for each of the previous optimized geometries using the Gaussian program suite[142]. A multiple conformation RESP fit was done over six conformations (three α and three β) for each residue with the previous charge constraints.

3.2.4 MM solvation calculations

Two variants of implicit solvent GB models were tested including "GB^{OBC}" model[122, 143] (igb =5 in Amber) and GBn model[93] (igb=7 in Amber). Ala dipeptides taken from geometry scanning (**2.3.2 Geometry scanning**) were minimized using ff14SB and any selected GB model including restraints on φ and ψ values with harmonic force constant of 1000 *kcal mol^{-1} rad^{-2}*. Poisson-Boltzmann (PB)[144] model was also adopted for the same calculation. Two radii sets were adopted including mbondi3 (same to GBneck2) and Tan and Luo's radii[145]. The same geometries were used and single point GB and PB energies were calculated with pbsa implemented in Amber v16[101].

Two explicit solvent models (TIP3P and OPC) were also employed. Thermodynamics integration (TI) was performed with pmemd implemented in Amber v16[101] to obtain polar solvation free energy for each conformation on Ala dipeptide φ/ψ grid with both TIP3P and OPC solvent. All atomic partial charges on Ala dipeptide were turned off in state 0 and turned on in state 1. Each conformation was solvated with 496 TIP3P waters or OPC waters in TI calculations. The production runs for the gas-phase state (state 0) and in-solution state (state 1) were conducted. Hydrogen atoms were constrained using the SHAKE algorithm[112]. The temperature was set to 298 K and no salt ions were included. Langevin dynamics was used to control temperature and the collision frequency was set to be 0.1 ps^{-1}. Particle mesh Ewald method was used to calculate electrostatic energies[146]. The cutoff of non-bonded interactions was set to 8 Å. A time step of 2 fs was used. The φ and ψ dihedral values were restrained onto the grid with harmonic force constant of 100 *kcal mol^{-1} rad^{-2}* during TI. A total of 12 λ windows were employed with values set to

0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366 0.88495, 0.95206 and 0.99078 following protocol from our previous TI studies[147]. Trapezoidal integration was used to obtain the free energy change between gas-phase and in-solution calculations.

3.3 Results and Discussion

We will investigate the possible errors that can be improved in long term force field development. First, the stability of hydrogen bond will be investigated by comparing energy surface of Ala dipeptide with tetrapeptide. This will also be helpful to understand the necessity of replacing tetrapeptide (ff99SB training model) with dipeptide (ff19SB training model) for backbone training. Second, the empirical 1-4 scaling factor will be examined by 1-4 grid scanning on Val dipeptide MM energies and identifying more optimal 1-4 scaling factors that yield better QM/MM agreement. Third, the accuracy of backbone partial charges on charged amino acids including Asp and Glu will be studied. The ff14SB and ff19SB helical propensity data will be used as reference and compared with new helical propensity calculated with updated partial charges and/or dihedral parameters. Lastly, we test whether the representation of MM solvation in ff19SB training will impact the eventual dihedral parameters.

3.3.1 Comparing backbone rotational energies between Ala tetrapeptide and Ala dipeptide

Choice of model systems is crucial in force field training. Enabled by greater computer power, this has led to fitting FF parameters against large peptides. The transferability gets improved this way since the training model more closely reflect the situations in which the parameters will be applied. The protein backbone φ and ψ dihedral parameters can alter the energy profiles for bond rotations, and thus influence secondary structure preferences and loop conformations. These have been frequently revised over the years based on observations of secondary structure biases in prior models[47]. While early FFs used capped single amino acids

(dipeptides) to train the backbone, our ff99SB[2b] FF used tetrapeptides, allowing φ and ψ parameters to be trained in a context of conformational diversity of neighboring amino acids in a longer peptide. The improvement was significant, and ff99SB has been widely adopted. Another reason of using tetrapeptide in ff99SB is gas-phase QM was used as reference data and tetrapeptide can form helices and thus have helical minimum. The barrier height from tetrapeptide in gas-phase hence becomes meaningful when adjusting dihedral parameters during fitting. In ff19SB, in-solution QM was used and the energy surface of dipeptide has minimum in α-basin due to the aligned dipole being stabilized by interacting with solvent dipole (**Figure 2.9**). The dipeptide is also the largest peptide that enable entire φ/ψ dihedral scan and provide training structures in the full backbone dihedral space. Both of these make dipeptide amenable for dihedral fitting in ff19SB.

However, switching from tetrapeptide to dipeptide, we are interested in knowing the sensitivity of force field to the training model, and retrospective about the advantage of using tetrapeptide back in 2006[2b]. Our major concern with dipeptide is, even with in-solution QM data, if the energy surface is hypothetically not representative of dipeptide in big biomolecule form with neighboring contexts, the dihedral parameters derived from dipeptide will be poorly transferable to big biomolecules. Therefore, we designed two model systems to study the 2D energy surface of Ala in tetrapeptide (Ace-Ala-Ala-Ala-Nme) with two different context, with first and second Ala being in either ppII or α conformation. We compared the CMAP derived from the two tetrapeptides with dipeptide CMAP, and also tested whether dipeptide CMAP can be transferable to the two tetrapeptides. 2D backbone rotation scan was done for the third Ala in tetrapeptide form (Ace-Ala-Ala-Ala-Nme), followed by restrained minimization and energy evaluation with implicit solvent for QM and MM. During scanning and optimization, both first and second Ala were restrained to either ppII or α conformation. This was designed on purpose so that the tetrapeptide can form a helical turn (by forming intermolecular H-bond) when both first and second Ala are adopting α conformations, while doesn't form helical turn when both first and second are adopting ppII conformations.

Backbone φ/ψ rotational energy profiles were analyzed for QM, MM and CMAP for the third Ala in tetrapeptide. The CMAPs were derived by subtracting total MM from total QM energies on the 2D grid (see **2.3.5 CMAP fitting**). As shown in **Figure 3.2**, the QM energy profiles between dipeptide (**Figure 3.2I**) and tetrapeptide with ppII restraints (**Figure 3.2A**) are very similar at negative φ values, but quite different at α_L basin. This is likely because the first and

second Ala (both in ppII conformations) form favorable short-range interactions with the backbone of third Ala in α_L conformation (**Figure 3.1),** but the interaction doesn't exist or is not strong when the third Ala has negative φ value. In dipeptide, this is not observed because of lack of neighboring chemical context. The QM energy profiles between dipeptide and tetrapeptide with α restraints (**Figure 3.2E**) are quite different in α_R basin because tetrapeptide can form helical structure when the third Ala is also in α conformation and it is energetically more favorable than solvated dipeptide. In ff14SB+GBSA, the overall energy profiles are very similar between dipeptide (**Figure 3.2F**) and tetrapeptide with ppII restraints (**Figure 3.2B**). However, in tetrapeptide with ppII restraints, the QM energy in α_R is higher than α_L, and the trend is reversed in dipeptide QM (**Figure 3.2A vs. Figure 3.2I**). This is likely because of the first two Ala adopting ppII conformation and having medium-range interaction with the third Ala. In tetrapeptide with ppII restraints MM, the α_R is lower than α_L. The energy issue might arise from non-bonded interaction error or solvation error between QM and MM. The helical minimum is observed in MM for tetrapeptide with α restraints (**Figure 3.2E**). But the shape of α-basin is different from QM and is highly symmetric with little φ/ψ coupling, due to the uncoupled cosine terms in backbone dihedral modelling. In general, the shape and location from QM are poorly reproduced by ff14SB for Ala tetrapeptides with either ppII or α restraints.

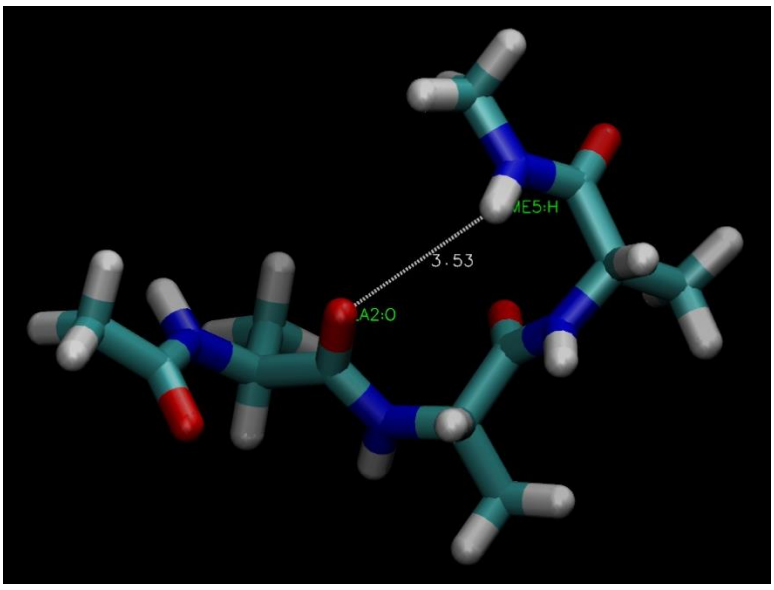


Figure 3.1 The structure of Ala tetrapeptide with first and second Ala in ppII and third Ala in α_L. The distance between C=O on second Ala and N-H on NME is labeled.

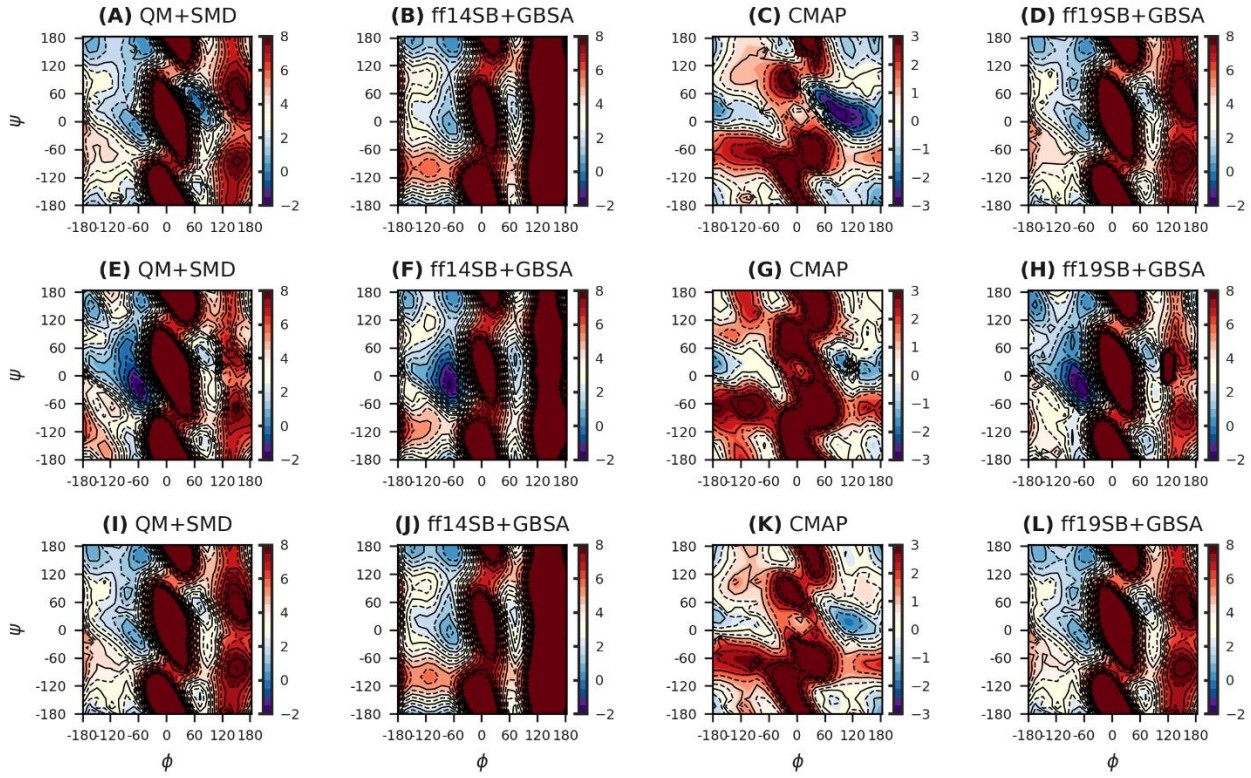


Figure 3.2 The third Ala of Ala tetrapeptide (first and second being ppII conformation) Ramachandran energy (kcal/mol) surfaces calculated in (A) QM+SMD, (B) ff14SB+GBSA, (C) CMAP (the difference between A and B) and (D) ff19SB+GBSA. The third Ala of Ala tetrapeptide (first and second being α conformation) Ramachandran energy (kcal/mol) surfaces calculated in (E) QM+SMD, (F) ff14SB+GBSA, (G) CMAP (the difference between E and F) and (H) ff19SB+GBSA. Ala dipeptide Ramachandran energy (kcal/mol) surfaces calculated in (I) QM+SMD, (J) ff14SB+GBSA, (K) CMAP (ff19SB CMAP) and (L) ff19SB+GBSA. All energies were zeroed relative to the lowest energy in the ppII region (defined in **Table 2.6**). The values above the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

The CMAP correction from dipeptide (**Figure 3.2K**) is overall similar to CMAP derived from tetrapeptide with α restraints (**Figure 3.2G**). Especially in α basin, the contours in both correction maps are very similar which indicate whether or not to include H-bond (whether dipeptide or tetrapeptide with α restraints) in training model has little effect on the resulted CMAP. In the CMAP derived from tetrapeptide with ppII restraints (**Figure 3.2C**), the α_L basin is notably over-stabilized comparing to dipeptide CMAP. The interaction between amide N on the NME capping group and carbonyl oxygen on the second Ala is medium-range with distance around 3.5 Å. There is a big disagreement between in-solution QM and MM energy for that region (α_L region

in **Figure 3.2A&B**). It is likely because of the non-bonded errors in MM including vdW, electrostatics and maybe GB solvation. Therefore, using this model system in training is risky because the observed errors are highly dependent on the context of the structures, here as, ppII conformation for both first and second Ala. When the geometry of the first and second Ala changes, the discrepancy between QM and MM would become smaller in α_L basin (such as when first and second Ala are in α) but might show up in other areas of the energy surface. The complicated characteristics of tetrapeptide makes it hard to be a ideal training model for CMAP training.

We further calculated average REE (**equation 3.1**) between dipeptide CMAP and tetrapeptide CMAP with α restraints and ppII restraints. Since big discrepancy between CMAPs exist in the α_L basin, we only considered energy surface at $\varphi < 0°$. For structures having negative φ values and having QM energy within 10 kcal/mol above the minimum, the average REE between dipeptide CMAP and tetrapeptide CMAP with α restraints is 0.78 kcal/mol, the difference between dipeptide CMAP and tetrapeptide CMAP with ppII restraints is 0.40 kcal/mol.

Based on both energy surfaces and quantitative analysis, we conclude that the variation of CMAP to model systems (dipeptide and tetrapeptide) is seemingly low. Therefore, the transferability of Ala CMAP in ff19SB (derived from dipeptide) to tetrapeptide should be reasonably accurate in theory. The energy profile of ff19SB+GBSA on the two tetrapeptides were also calculated (**Figure 3.2**). They are qualitatively agreeing better with QM profiles than ff14SB. To quantify the agreement, we further calculated average REE between QM (with SMD) and MM (ff19SB+GBSA) for tetrapeptides with both α and ppII restraints as a function of QM energy range above the minimum (similar calculation in **Figure 2.16**). Based on results, the ff19SB model works reasonably well for both tetrapeptides with average REE below 0.8 kcal/mol even at high QM energies. The first peak in ff19SB tetrapeptide ppII profile (**Figure 3.3 right**) results from the arbitrarily low QM energies in α_L region (see **Figure 3.2A**). These data provide rigorous test on transferability of ff19SB on molecules bigger (tetrapeptide) than training model (dipeptide), and can form actual secondary structure comparing to absence of secondary structure in training model. The agreement between QM and ff19SB proves that dipeptide serves as a reasonable model system when combined with the other assumptions made in ff19SB training such as including entire dihedral space of structures, use of "perfect" fitting function CMAP and employing in-solution QM data as reference.

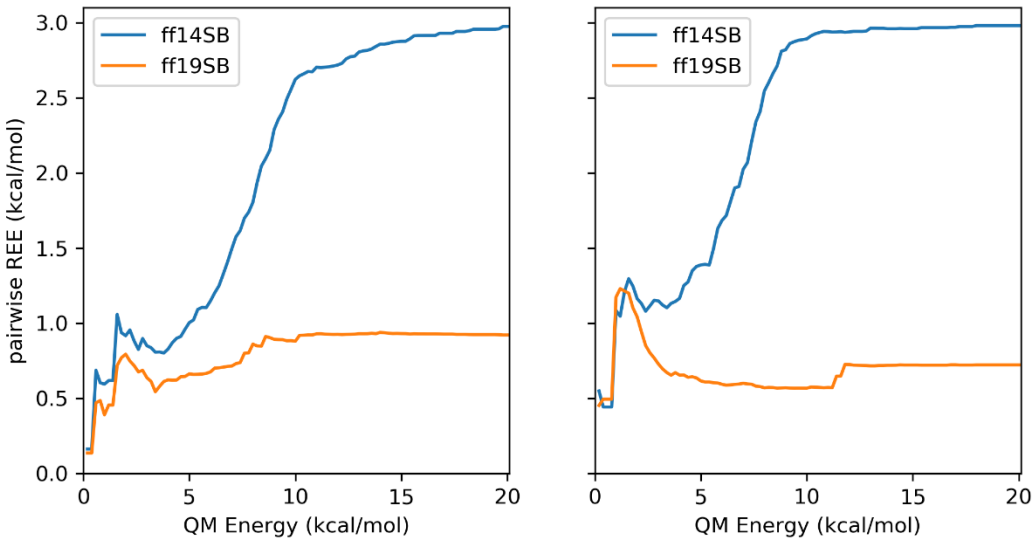


Figure 3.3 The average REE between QM and ff14SB (blue), QM and ff19SB (orange) for tetrapeptide with α (left panel) and ppII (right panel) restraints as a function of QM energy range above the minimum.

3.3.2 Improved reproduction of QM energy with ff14SB and modified 1-4 scaling factor

The Val dipeptide was used to investigate 1-4 scaling factor due to its small size overall and β-branched side chain. The scaling factor including 1-4 vdW and 1-4 electrostatics are used to empirically correct for short-range non-bonded interaction between atoms separated by three covalent bonds, and used to compensate for dihedral potential errors. As shown in **Figure 3.4**, the 1-4 distance is strongly associated with dihedral angle. Thus the scaling of 1-4 non-bonded interaction could highly change the torsion potential. Due to the simplicity of Lennord-Jones potential and the use of fixed-point charges together with Coulomb's law, these 1-4 scaling factors are practically useful to weaken the short range interactions. However, both 1-4 vdW and 1-4 electrostatics are empirically applied to any four consecutive heavy atoms regardless of the local chemical environment. The error of 1-4 scaling could vary as the bond rotates and local chemical environment changes, and the optimal value of 1-4 scaling factor has never been exhaustively tested on multiple conformations.

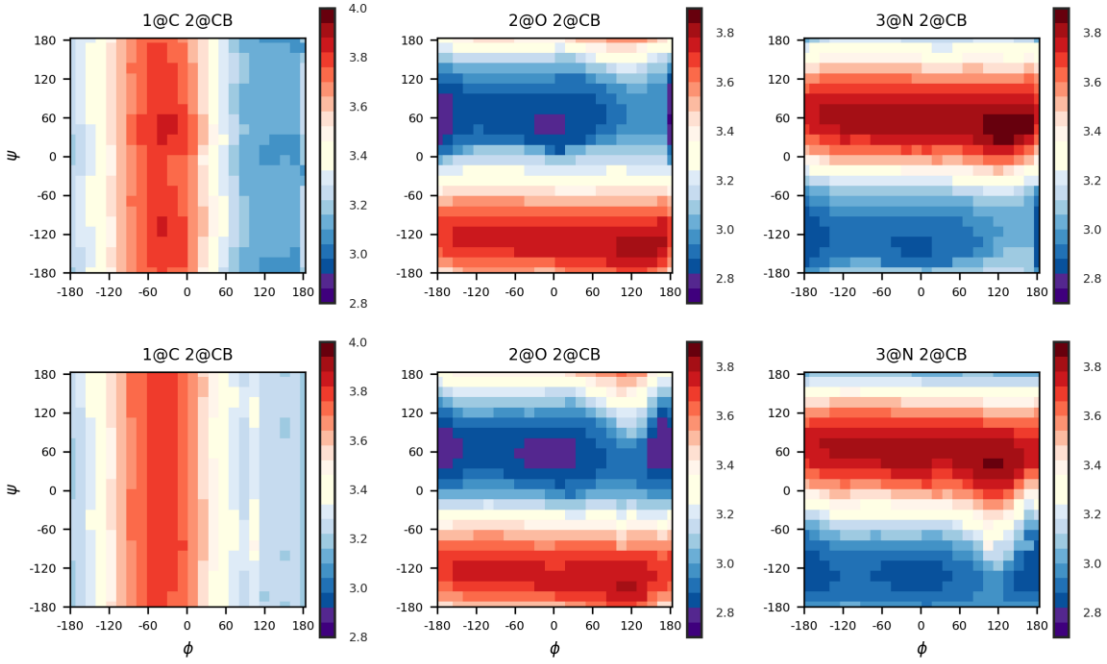


Figure 3.4 The Ramachandran map on 1-4 distance for Val dipeptide in trans rotamer (top row) and *gauche(-)* rotamer (bottom row).

In order to investigate the errors of 1-4 scaling on multiple dipeptide conformations, we tested the optimal combination of 1-4 vdW and 1-4 electrostatics by performing two-dimensional 1-4 scanning on the 576 Val dipeptides for both *trans* and *gauche(-)* rotamers. Val was chosen since it is β-branched and can form short-range intermolecular interactions between side chain and backbone. The error of MM against QM is quantified by the average relative energy error (REE) over structures that were scanned in full φ/ψ dihedral space. Only structures having QM energy within 10 kcal/mol above the minimum were included to avoid the artifacts of high energy structures on the error analysis. As shown in **Figure 3.5**, the optimal 1-4 scaling factors lie in a region that is higher than the by default value in Amber (2.0 for scnb and 1.2 for scee) for both *trans* and *gauche(-)* rotamer. Since the MM is defined as ff14SB00+adjustable scee&scnb and the backbone dihedral parameters of MM are in fact set to 0 in all the error calculation, the disagreement between QM and MM reflects isolated error in 1-4 scaling factor rather than the combined error of 1-4 scaling factor and dihedral parameters. The shift of the minimum in **Figure 3.5** to bigger values might indicate overestimated repulsion in short-range for ff14SB. The

minimum for *trans* rotamer is when scnb=2.8 and scee=1.6 and the minimum for *gauche(-)* rotamer is when scnb=2.8 and scee=1.4.

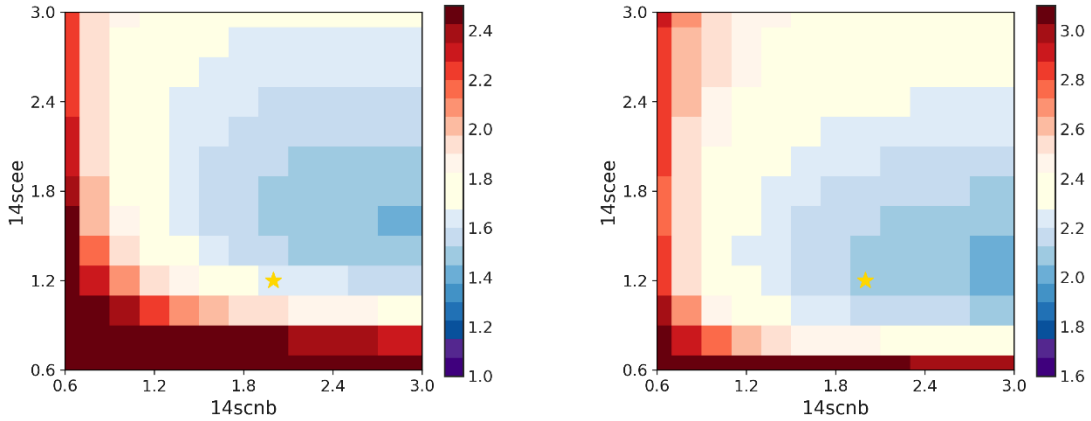


Figure 3.5 The average REE of Val dipeptide between QM and MM in 2D scanning 1-4 scnb (X-axis) and 1-4 scee (Y-axis) for (left) *trans* rotamer and (right) *gauche(-)* rotamer. Only structures having QM energy within 10 kcal/mol above the minimum were included in the error calculations. The average REE with the by default 14scee (1.2) and 14scnb (2.0) in Amber was labeled as golden star.

We replaced the default values with the optimal ones (scnb=2.8 and scee=1.4) shown in and combined them with the original ff14SB00 to create the force field ff14SB00_new (ff14SB00+scee+scnb). We re-calculated average REE between QM (with SMD) and MM (ff14SB00_new+GBSA) for both *trans* and *gauche(-)* as a function of QM energy range above the minimum (similar calculation as used in **Figure 2.16**) shown in **Figure 3.6**. We compared ff14SB00_new with the original ff14SB00. As shown in **Figure 3.6**, the error is slightly smaller with ff14SB00_new in general, especially for *trans* rotamer at high QM energy (>5 kcal/mol). The energy profiles are very similar between ff14SB00 and ff14SB00_new for *gauche(-)* rotamer and subtle improvement is observed with the modified 1-4 scaling factors, likely because steric clash happens less frequently in φ/ψ dihedral space when *gauche(-)* is adopted.

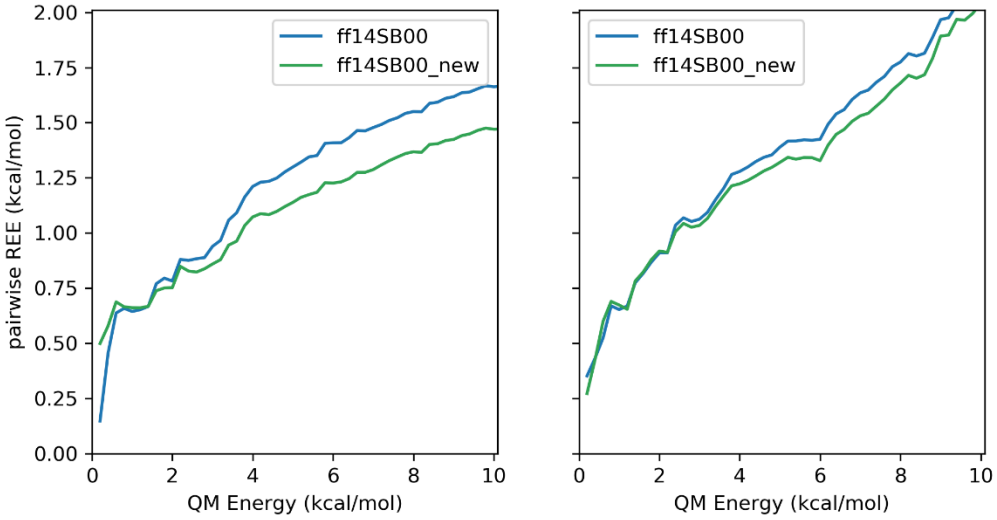


Figure 3.6 The average REE between QM and ff14SB00, QM and ff14SB00_new (14scnb=2.8, 14scee=1.4) as a function of QM energy range above the minimum for Val dipeptide in (left) *trans* rotamer and (right) *gauche(-)* rotamer. The blue curves are ff14SB00 and the green curves are ff14SB00_new.

Due to the fact that Val dipeptide is such a small system with non-polar β-branched side chain, we expect the 1-4 vdW is more sensitive than 1-4 electrostatics to the overall potential energy. Based on our preliminary data, the difference between QM and MM is mildly dependent on the choice of empirical 1-4 scaling factors. Other short-range interactions such as 1-5 (between Cγ and carbon on amide carbonyl) and 1-6 (between Cγ and oxygen on amide carbonyl) interactions might also contribute to the discrepancy between QM and MM and need further investigation. Previous studies have shown that 1-4 scaling factor is sensitive to the conformational equilibrium between secondary structures[141], and a reduced scaling factor might be beneficial for protein folding. However we provide contradictory results that exaggerated 1-4 scaling factors (14scnb=2.8, 14scee=1.4) can mildly improve the agreement between force field and QM energies (approximately by 0.25 kcal/mol). Our preliminary investigation on 1-4 scaling factor was done with small non-polar peptide (Val), and our proposed favorable modification to scaling factors may not be transferable to other amino acids having charged or polar side chains. Besides, further extensive MD simulations in polypeptides are also necessary to validate our findings.

3.3.3 Helical propensities worsen for charged amino acids with updated partial charges in ff14SB

In ff14SB, the backbone partial charges are different between non-charged amino acids (Gly, Ala, Val, Thr, etc) and charged amino acids (Asp, Glu, Arg, Lys and double-protonated His), and the backbone dihedral parameters are shared broadly among all of the amino acids. In this case, the same dihedral parameters are applied to segments (defined using four consecutive atom types) that have different partial charges. A single dihedral term is unlikely to be an equally accurate correction in situation where the charge distribution is different. The overly broad application of dihedrals is a significant inconsistency and weakness in current models. To overcome this, we can either differentiate dihedral parameters or make charge/dihedral consistent by equalizing the charge distribution. In ff19SB development, we are certain that diversified amino-acid specific backbone parameters are required to achieve better force field model (see **Chapter 2**). Thus, we employed amino-acid specific dihedral parameter (CMAP) to better model the backbone profiles of 20 amino acids. However, with employing CMAP and retaining old partial charges, the errors which the CMAP is actually correcting for might arise directly from the error of partial charges and the change of error of partial charges across amino acids. If the error of partial charge and the inconsistency between partial charge and dihedral in ff14SB is the source of error, then fixing partial charges in ff14SB might be more meaningful. Our assumption here is the backbone partial charges and backbone dihedral parameters should be both same across amino acids, and the amino-acid specific behavior is only modelled by side chain parameters and non-bonded parameters.

Based on our helical propensity results (**2.4.5 Amino-acid specific helical propensities are significantly improved in ff19SB**), we observed the overestimation of helical propensity for charged amino acid such as Asp and Glu in ff14SB and both are outliers relative to the rest of amino acids regardless of in TIP3P or OPC (**Figure 2.25**). The backbone charges of Asp and Glu are different from the non-polar amino acids because of the assumption made decades ago that charge distribution of charged side chain has notable and different effects on backbone charge distribution against non-charged ones. However, as mentioned, this assumption could be wrong, and it is okay having the same backbone charges for all amino acids and same backbone dihedral parameters for all amino acids. In addition, the H-bond stability is greatly affected by the

magnitude of backbone charges that model how strong the dipole moment of amide is. In ff19SB, instead of revising this assumption, we updated backbone dihedral parameters for each amino acid without changing backbone partial charges. The results show that Asp and Glu are no longer outliers and are well correlated with experimental data (**Figure 2.25**). We suspect there is error of cancellation and the charge errors in Asp/Glu are compensated by CMAP unintentionally.

Best et al.[32] revisited the charges and refit partial charges of Cα and side chain atoms on charged amino acids (D, E, K, R) while forcing the charges on amide N, H, C, O to have same values as all the other residues. They reported helical propensity benchmarks for 20 amino acids, showing that the overall trend from experiments[72] was better reproduced by the new force field ff99SB*_ILDN_Q. The refitting of partial charges was performed with the old force field ff99SB*_ILDN[31] in which several iterations of parameter tweaking was done to improve model matching to NMR chemical shifts and J-coupling data. In ff14SB, we systematically improved the side-chain parameters and the training is more rigorous. Therefore, we revisit same assumption as Best et al. years ago by examining the errors of backbone partial charges of the outliers Asp and Glu and their effects on helical propensity, but with the context of ff14SB.

Following Best et al.'s protocol, we unset the backbone charges to the same ones as in Ala and refit side chain for Asp and Glu (see **3.2.3 Refitting of atomic partial charges**). We combined the new charges of Asp and Glu with the original ff14SB and made a new force field ff14SB_Q. The helical propensity simulations were reran with ff14SB_Q+OPC and were compared against original ff14SB+OPC simulations. The OPC is used since it better models solvation than TIP3P. As shown in **Figure 3.7**, the helical propensity of Glu decreases with ff14SB_Q+OPC and agrees better with NMR data, while the helical propensity of Asp increases with ff14SB_Q+OPC and deviates more from the NMR data. The change of helical propensity with new charges indicate that the magnitude and distribution of partial charges are quite sensitive to the helical propensity. And the contradictory results (Asp vs. Glu) indicate that more consistent and rigorous charge refitting is necessary. Best et al's [32] refitting is not consistent across amino acids in terms of dihedral fitting, etc, and in ours, both charges and dihedral parameters match Ala. But our training is still not rigorous due to lack of dihedral refitting with the updated partial charges.

To improve the consistency of partial charges and dihedral parameters, we refit CMAP for Asp with new charges (same to ff14SB_Q) and created a new force field ff14SB_Q_CMAP following the protocol in **2.3.5 CMAP fitting**. We simulated A4DA4 with the new force field and

fit helical propensity for Asp (**Figure 3.7**). There is no difference between ff14SB_Q_CMAP and ff19SB even though both partial charges and CMAP are different between the two force fields.

Based on the improvements of Glu on helical propensity, we believe there is room for improvements on the old Amber partial charges (especially charged amino acids) that were developed decades ago[8]. The improvement of Asp is not significant. The interplay between partial charges and dihedral parameters on Asp could be more complicated than Glu because of the shorter Asp side chain and stronger electrostatic interactions. In the future, systematic refitting on backbone partial charges followed by backbone dihedral refitting is necessary to fundamentally improve the accuracy of the force field.

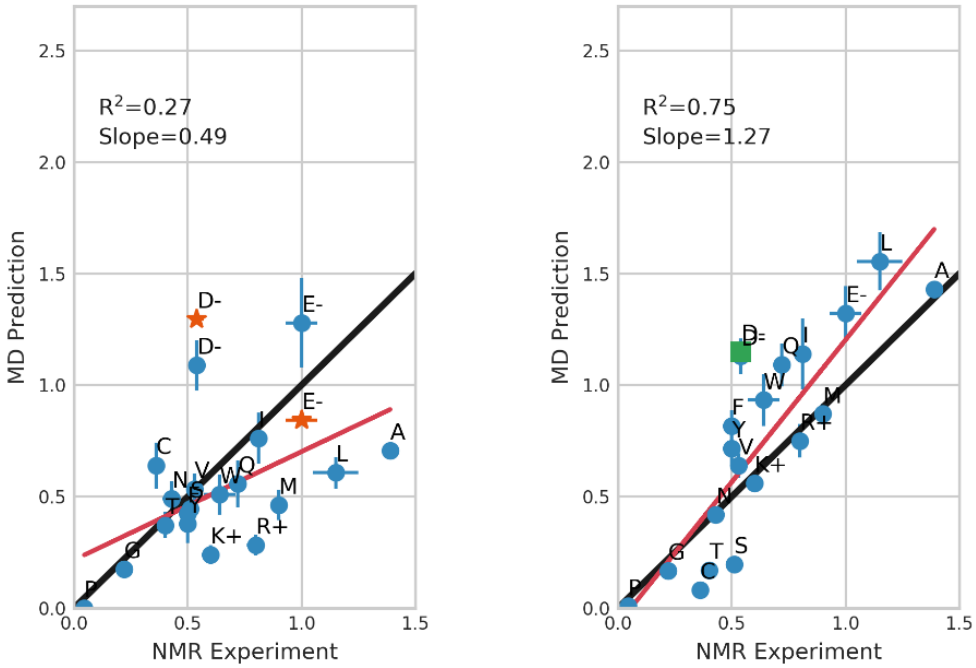


Figure 3.7 Correlation between helical propensities *w* from experiment[72] and simulations using (left) ff14SB+OPC (blue dots) and ff14SB_Q+OPC (orange stars), (right) ff19SB+OPC (blue dots) and ff14SB_Q_CMAP+OPC (green square). Amino acids are indicated using single letter codes. Values on the X-axis represent the data based on NMR[72] and the reported standard deviations. Values on Y-axis represent the helical propensities fit against the combined trajectory (3.2 μs * 12, blue dots), with error bars calculated via bootstrapping analysis. No error bars are reported for ff14SB_Q+OPC and ff14SB_Q_CMAP+OPC data. Black lines represent perfect agreement. Linear regression (red lines) was performed against the data points (only blue dots), with R^2 and slope quantifying the goodness of fit.

3.3.4 ff19SB training is insensitive to MM solvent model for dipeptide

One possible weakness to our approach in ff19SB development was the use of simple implicit water models during training, such as the GB model in the MM component. Older GB models exhibit secondary structure biases for longer peptides[139], but here we have used our GBneck2 model[92] that much more accurately reproduces secondary structure preferences. Furthermore, we have shown that the solvation energy of dipeptides (which we used here for the CMAP training in GB) is largely insensitive to specific GB model used[140]. The assumption is reasonable in general. But our use of GB during training could still be a limitation, and it is one reason we carried out extensive testing with a variety of explicit water models. In this section, we revisited this assumption in ff19SB development, on using the combination of QM+SMD and GBneck2 + SASA for CMAP training. This assumption was made to address the inconsistency issue in dihedral fitting and partial charge fitting in the previous force fields. The GBSA model was used in MM calculation to cancel out the solvation energy in SMD. The cancelation is a must to avoid including solvation errors in the backbone parameters (see **2.3.3 Molecular mechanics (MM) optimization and energy calculations**). However, there is no evidence to rigorously prove that the solvation energy is exactly cancelled out with our QM/MM combination in ff19SB training even though SMD and GB use similar theory (Onsager model[148] and Born model[149]) in solvation calculation and are mathematically equivalent according to previous study[150]. More tests are required to show the variation of MM solvation and its sensitivity to CMAP.

We performed additional calculations to explore the variation and sensitivity of the MM solvation model. We tested out several solvent models in MM by calculating solvation energies and comparing across them. The difference among them is quantified as well. As shown in **Figure 3.8**, the energy profiles look highly similar across different solvent models including implicit and explicit models. The solvation energy in explicit solvent was obtained by doing TI calculations which is sufficiently rigorous and accurate. The relative energy errors are shown in **Figure 3.10**. The four implicit solvent models including GB models (GB^{OBC}, GBn and GBneck2) and two PB models (PBb3 that use GBneck2 radii mbondi3, and PBtl that use Tan and Luo's radii[145]) are quite similar to each other with average REE (see **3.2.4 MM solvation calculations**) <=0.3 kcal/mol, Interestingly the difference between implicit solvent models and explicit solvent models (TIP3P and OPC) are ~0.45 kcal/mol. The system tested here is small dipeptide which is largely solvent exposed and does not form secondary structures such as helices and strands and hence insensitive

to the solvent model. The closest model to GBneck2 is PBb3 (0.15 kcal/mol). This is reasonable since GBneck2 and PBb3 calculations use same intrinsic radii (mbondi3) and GBneck2 was initially trained to partly reproduce PB parameters[92].

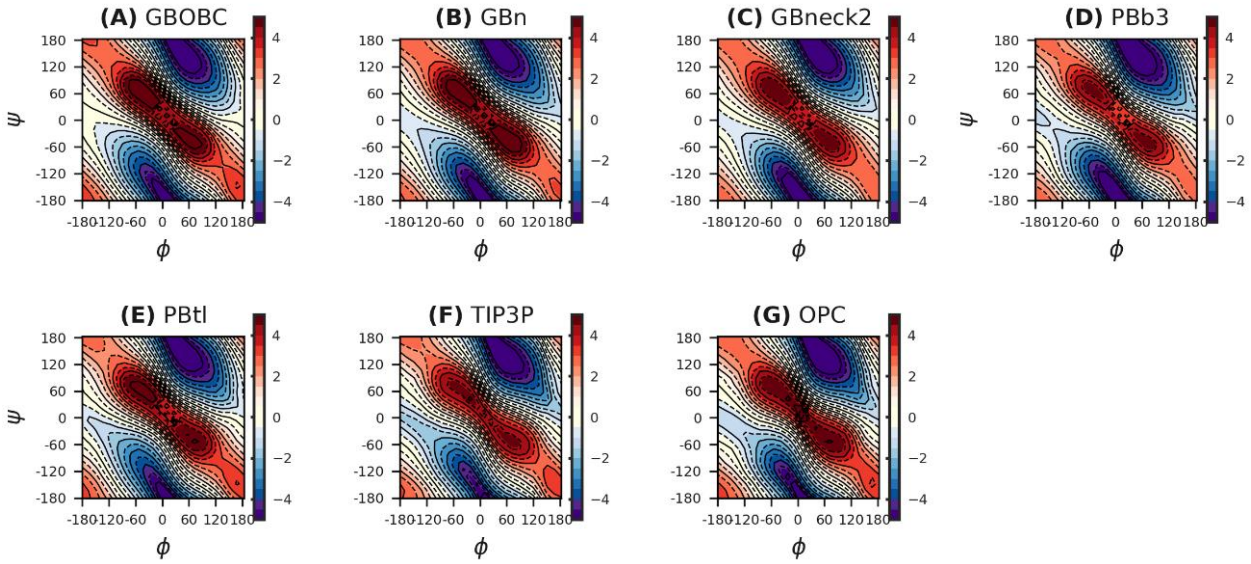


Figure 3.8 Ala dipeptide Ramachandran solvation energy (kcal/mol) surfaces calculated in (A) GB^{OBC}, (B) GBn, (C) GBneck2, (D) PBb3 (mbondi3), (E) PBtl (Tan and Luo's radii), (F) TIP3P and (G) OPC. All energies were zeroed relative to the energy in the ppII conformation (φ=-60° and ψ=150°). The values above the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

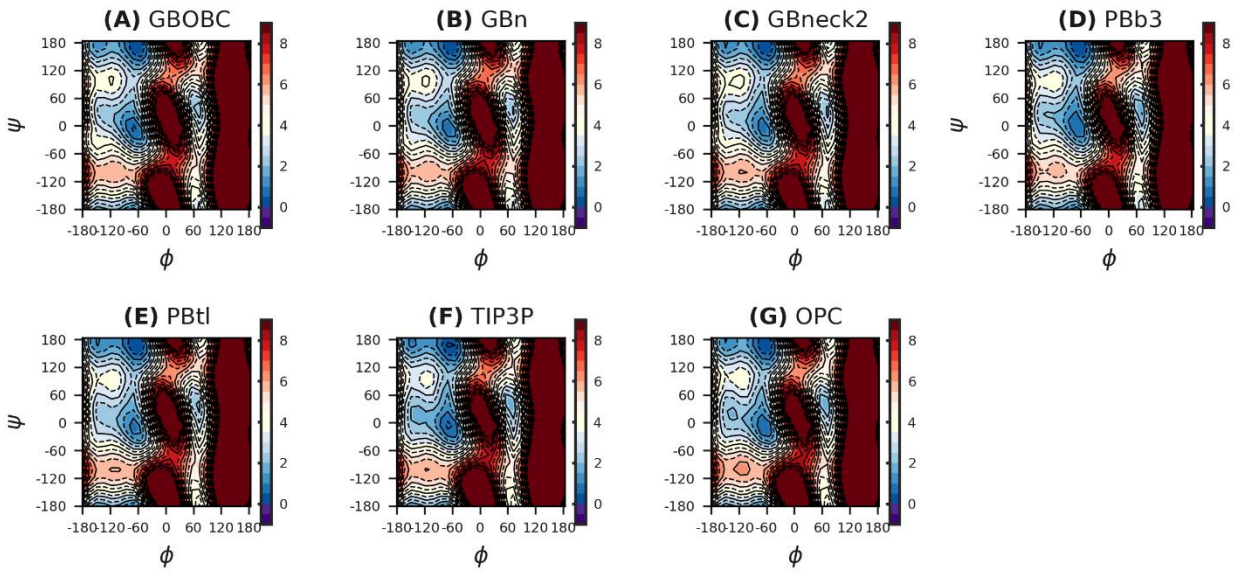


Figure 3.9 Ala dipeptide Ramachandran ff14SB energy + solvation energy (kcal/mol) surfaces calculated in (A) GBOBC, (B) GBn, (C) GBneck2, (D) PBb3 (mbondi3), (E) PBtl (Tan and Luo's radii), (F) TIP3P and (G) OPC. All energies were zeroed relative to the energy in the ppII conformation (φ=-60° and ψ=150°). The values above the color bar range are depicted in dark red. Solid, labeled contours indicate integer energy values in kcal/mol, whereas dashed contours indicate half-integer energies. The bicubic spline interpolation implemented in Python was used to calculate values between grid points.

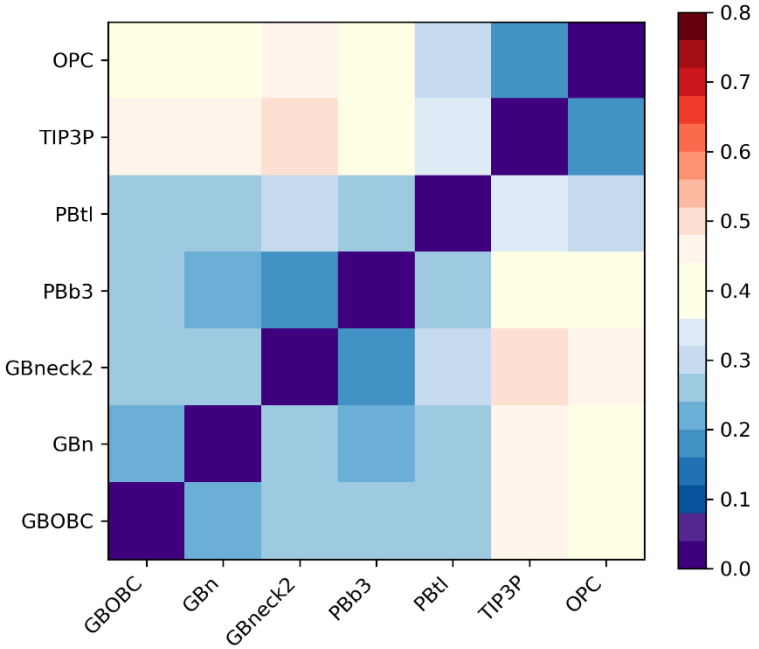


Figure 3.10 Heat map of Ala dipeptide on relative solvation energy error (REE) (kcal/mol) among different solvent models for 576 conformations in full backbone dihedral space.

Given the high similarity among GB variants and PB variants, and between implicit solvent model and explicit solvent model, we believe the assumption made in ff19SB is still reasonable. The magnitude of MM solvation to the overall ff19SB training is below **0.4 kcal/mol**, and even smaller (**0.2 kcal/mol**) when structures having ff14SB+GBneck2 energies higher than 10 kcal/mol are removed. This error is relatively small comparing to the uncertainties in the other parts of ff19SB training such as rotamer dependency and grouping of CMAP fitting.

3.4 Conclusion

In this chapter, we further investigate the physical cause of errors that are corrected by CMAP of ff19SB. Because dihedral parameters such as CMAP are fit to make up for the total QM energy profile, by improving other terms in the force field we could fundamentally improve the dihedral potential. The force field ff14SB was used as reference since ff14SB was also set as reference in ff19SB training. We explored several possibilities of source of error in ff14SB model without backbone dihedral terms, including hydrogen-bond, 1-4 scaling factor and partial charges. The backbone dihedral parameters were removed (ff14SB00, defined in **Table 2.2**) to isolate the problem. Lastly, one of the major assumptions in ff19SB training was revisited on using GBneck2 in MM solvation calculation. We quantified the energy difference among variants of MM solvation models. Based on our preliminary results, we conclude that: 1. The hydrogen bond stability is reasonably accurate in ff14SB/ff19SB. 2. A modification to the 1-4 scaling factors can improve QM/MM agreement by at most 0.2 kcal/mol. 3. There is still room of improvements on the old Amber partial charges but systematic refitting of both partial charges and dihedral parameters are required. 4. The MM solvation is insensitive to the CMAP training in ff19SB.

Chapter 4

Future directions

The biomolecular force fields have made significant progress in biophysical simulations in the past decades. A list of recent reviews on force field advances could be referred[151]. The goal of this dissertation is to significantly improve force field for biomolecular simulations. I have shown that in the updated ff19SB protein force field presented in **Chapter 2**, we have developed new backbone dihedral parameters with amino-acid specific CMAP functions. We trained the parameters to match solution phase QM data using full 2D φ/ψ scans, instead of the gas-phase minima used for training uncoupled φ and ψ cosine terms in ff99SB. Use of energies calculated from QM in solution provides better consistency with the pre-polarized partial atomic charges used by the MM model, as compared to gas-phase energies that were used previously. Fitting of dihedral corrections against QM in solution also allows the model to incorporate (to some extent) conformation-dependent polarization energy that is not present explicitly in a fixed-charge MM model such as the one used here. A total of ~6 milliseconds MD simulations in explicit solvent were performed to extensively validate ff19SB against experiments. We have observed that ff19SB, when combined with a more accurate water model such as OPC, should have better predictive power for modeling sequence-specific behavior, protein mutations, and also rational protein design. The ff19SB model provides immediate benefit to improving the overall accuracy

of energy function, however, additional adjustments on ff19SB, especially on CMAP parameters are necessary to fine tune the model and further improve the model.

In **Chapter 3**, I further investigate the physical cause of errors that are corrected by CMAP in ff19SB. The potential rooms for improvements in the rest of the force field were identified and quantified. I explored several possibilities of source of error in ff14SB00 (ff14SB model without backbone dihedral terms) including hydrogen-bond, 1-4 scaling factor and partial charges. One of the major assumptions in ff19SB training, using GB model for MM solvation, was revisited by quantifying the energy difference among variants of MM solvation models. We conclude that the force field is expected to be improved if the long-standing weaknesses in non-bonded parameters can be revised. However, systematic retraining and extensive testing are required for the terms that are dependent on non-bonded parameters such as backbone and side-chain dihedral parameters.

The other major issue is that the overly broad application of atom types lead to significant inconsistency and weakness in current models. This is caused by using atom type. For example, the partial charges were trained for each amino acid based on a few selected conformations of dipeptides. The dihedral parameters were not as diversified as partial charges. In ff19SB, every amino acid has its own backbone dihedral parameters but side chain parameters are still shared among amino acids having different side chain charges (see ff14SB paper[2c]). The partial charges and dihedral parameters were trained against different QM level of theory as well. We hope to address these inconsistency by having a consistent FF training protocol: (1) Train partial charges and dihedral parameters (side chain and backbone) against same model system dipeptide for each amino acid separately. (2) Use same QM level of theory for both ESP and energy calculation. In the meanwhile, this will overcome the issue of using inconsistent QM reference data for partial charges and dihedral parameters in all previous AMBER force fields.

Based on our preliminary results, the partial charges are highly dependent on conformation (**Figure 4.1**). Therefore, the dihedral fitting should be done with considering the sensitivity of partial charges to conformation. Two protocols are given below. In **protocol 1**, the dihedral fitting was done for each amino acid with conformation-dependent charges rather than single amino-acid specific partial charges. Multi-dimensional QM scan will be performed on the model system and two sets of reference data will be first obtained for each conformation: ESP and total QM energy. For each conformation, the partial charges will be fit following RESP scheme. The dihedral fitting will be first performed based on QM energies and MM energies with conformation-dependent

partial charges (the first fitting error, quantifying QM/MM energy disagreement). Then the resulted dihedral parameters will be evaluated with one chosen set of partial charges to check if the QM profiles can be reproduced equally well (the second fitting error, quantifying QM/MM energy disagreement). The difference between first and second fitting error is that the first is evaluated with conformation-dependent partial charges while the second is evaluated with one single set of partial charges. The selection of partial charges depends on the distribution of partial charge sets (collected from all conformations) and the second fitting error. Theoretically, the most populated partial charges should give lowest second fitting error and be used for MD. This consistent training will be performed for each amino acid separately, and the partial charges and dihedral parameters will be equally broadly applied to each amino acid. Both partial charges and dihedral parameters will be amino-acid specific and derived from consistent training in amino-acid specific manner. The equations used for fitting are as following:

$$O = \frac{2}{N(N-1)} \sum_{i}^{N-1} \sum_{j>i}^{N} \left[\left(QM_i - QM_j \right) - \left(MM(E_{ele}, E_{dih})_i - MM(E_{ele}, E_{dih})_j \right) \right] \quad (4.1),$$

where N is the number of conformations for a particular amino-acid type in dipeptide form (ACE-X-NME), MM energy is dependent on both the selected force field and also adjustable E_{ele} and E_{dih}. For each conformation i, the MM is calculated as:

$$MM(E_{ele}, E_{dih}) = MM^0 + E_{ele}(q_i) + E_{dih}(\varphi_1, \varphi_2 \dots) \quad (4.2),$$

where E_{ele} is the electrostatic energy calculated using Coulomb's law and atom-centered partial charges, E_{dih} is dihedral energy calculated using cosine functions and/or CMAP functions with adjustable dihedral parameters. MM^0 is calculated with the remaining FF terms including bonds, angles, etc. q_i is partial charge fit against the conformation i using RESP scheme. φ_1, φ_2 … are the dihedral values for conformation i. The objective function *O* will be optimized by adjusting the parameters used in E_{dih} calculation.

In **protocol 2**, the dihedral fitting will be performed with each set of partial charges (gained from fitting to each conformation) separately. The dihedral parameter + partial charges combination that gives lowest fitting error will be eventual force field parameters and used for MD.

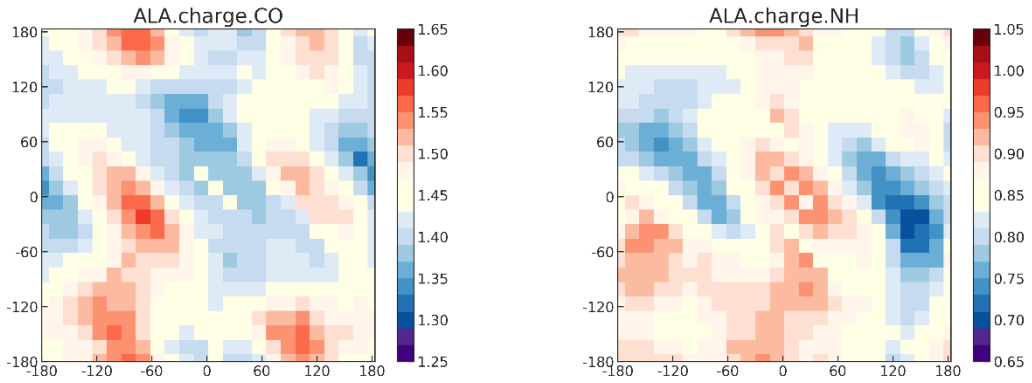


Figure 4.1 The magnitude of partial charges on amide C=O bond (left) and N-H bond (right) for each Ala dipeptide conformation. The RESP charges were fit for each conformation separately against the ESP obtained from M05-2X/6-311G**/SMD calculations.

An important component of force field development is extensive validation against data outside that used for training. In ff19SB, we performed a total of ~6 milliseconds MD simulations in explicit solvent to extensively validate force field parameters against experiments. A variety of test systems are included such as NMR scalar coupling data of short peptides, NMR S^2 parameters of folded proteins, chemical shift of polypeptides and PDB statistical data. However, additional testing results are always helpful for force field development. Negative results are often more informative than successes since they help pinpoint weaknesses and opportunities for model improvements.

One of the longstanding challenges in classical MD simulations is to predict the less well-defined intrinsic disordered peptides and proteins (IDP). Several recent force fields including Charmm36m[3d] and a99SB-disp[40d] have managed to extensively train force field parameters to better reproduce IDP data such as structural population and NMR, in the meanwhile minimizing errors on well-defined folded simulations. This makes parameter optimization a complicated process and the resulted parameters that were intentionally optimized to reduce errors for both folded and unfolded data are in fact not working well for either one. This is also difficult since the IDP is more sensitive to water model than solute FF and a rigorous FF training might need good deconvolution of the solvent model from the solute FF. We will continue test our AMBER force fields including ff19SB and its potential variants against disordered proteins. Some groups have reported comparisons of IDP simulations to experiment with a variety of force field models[3d, 40d, 87a, 87c, 152] and we could compare to these to assess improvements of our new model.

Bibliography

1. (a) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K., Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006,** *14* (3), 437-449; (b) Zhao, G. P.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J. Y.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. J., Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **2013,** *497* (7451), 643-646.

2. (a) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society* **1996,** *118* (9), 2309-2309; (b) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **2006,** *65* (3), 712-725; (c) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-713; (d) Wang, J. M.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry* **2000,** *21* (12), 1049-1074.

3. (a) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **1998,** *102* (18), 3586-3616; (b) Mackerell, A. D.; Feig, M.; Brooks, C. L., Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry* **2004,** *25* (11), 1400-1415; (c) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J Chem Theory Comput* **2012,** *8* (9), 3257-3273; (d) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2017,** *14* (1), 71-73.

4. (a) Jorgensen, W. L.; Tiradorives, J., The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *Journal of the American Chemical Society* **1988,** *110* (6), 1657-1666; (b) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.

5. Morse, P. M., Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys Rev* **1929,** *34* (1), 57-64.

6. Kao, J.; Allinger, N. L., Conformational-Analysis .122. Heats of Formation of Conjugated Hydrocarbons by Force-Field Method. *Journal of the American Chemical Society* **1977,** *99* (4), 975-986.

7. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *Journal of the American Chemical Society* **1984,** *106* (3), 765-784.

8. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J Phys Chem-Us* **1993,** *97* (40), 10269-10280.

9. Shi, Y.; Xia, Z.; Zhang, J. J.; Best, R.; Wu, C. J.; Ponder, J. W.; Ren, P. Y., Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of chemical theory and computation* **2013,** *9* (9), 4046-4063.

10. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *Journal of chemical theory and computation* **2013,** *9* (12), 5430-5449.

11. Jing, Z. F.; Liu, C. W.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J. P.; Ren, P. Y., Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics, Vol 48* **2019,** *48*, 371-394.

12. Satpati, P.; Clavaguera, C.; Ohanessian, G.; Simonson, T., Free Energy Simulations of a GTPase: GTP and GDP Binding to Archaeal Initiation Factor 2. *Journal of Physical Chemistry B* **2011,** *115* (20), 6749-6763.

13. Pang, Y. P., Use of 1-4 interaction scaling factors to control the conformational equilibrium between alpha-helix and beta-strand. *Biochemical and Biophysical Research Communications* **2015,** *457* (2), 183-186.

14. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A., An All Atom Force-Field for Simulations of Proteins and Nucleic-Acids. *Journal of computational chemistry* **1986,** *7* (2), 230-252.

15. Gao, J. L.; Pavelites, J. J., Aqueous Basicity of the Carboxylate Lone Pairs and the C-O Barrier in Acetic-Acid - a Combined Quantum and Statistical Mechanical Study. *Journal of the American Chemical Society* **1992,** *114* (5), 1912-1914.

16. Burkert, U.; Allinger, N. L., *Molecular mechanics*. American Chemical Society: Washington, D.C., 1982; p xi, 339 p.

17. Dinur, U. a. H., A. T., *Reviews in Computational Chemistry* **1991,** *Volume 2*.

18. (a) Allinger, N. L., Conformational-Analysis .130. Mm2 - Hydrocarbon Force-Field Utilizing V1 and V2 Torsional Terms. *Journal of the American Chemical Society* **1977,** *99* (25), 8127-8134; (b) Lii, J. H.; Allinger, N. L., Molecular Mechanics - the Mm3 Force-Field for Hydrocarbons .3. The Vanderwaals Potentials and Crystal Data for Aliphatic and Aromatic-Hydrocarbons. *Journal of the American Chemical Society* **1989,** *111* (23), 8576-8582; (c) Lii, J. H.; Allinger, N. L., Molecular Mechanics - the Mm3 Force-Field for Hydrocarbons .2. Vibrational Frequencies and Thermodynamics. *Journal of the American Chemical Society* **1989,** *111* (23), 8566-8575.

19. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B* **1998,** *102* (18), 3586-616.

20. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman., *AMBER 2015, University of California, San Francisco* **2015**.

21. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.

22. Momany, F. A., Determination of Partial Atomic Charges from Abinitio Molecular Electrostatic Potentials - Application to Formamide, Methanol, and Formic-Acid. *J Phys Chem-Us* **1978,** *82* (5), 592-601.

23. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A., Application of Resp Charges to Calculate Conformational Energies, Hydrogen-Bond Energies, and Free-Energies of Solvation. *Journal of the American Chemical Society* **1993,** *115* (21), 9620-9631.

24. (a) Gould, I. R.; Kollman, P. A., Abinitio Scf and Mp2 Calculations on 4 Low-Energy Conformers of N-Acetyl-N'-Methylalaninamide. *J Phys Chem-Us* **1992,** *96* (23), 9255-9258; (b) Gould, I. R.; Cornell, W. D.; Hillier, I. H., A Quantum-Mechanical Investigation of the Conformational Energetics of the Alanine and Glycine Dipeptides in the Gas-Phase and in Aqueous-Solution. *Journal of the American Chemical Society* **1994,** *116* (20), 9250-9256.

25. Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A., The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In *Computer Simulation of Biomolecular Systems*, van Gunsteren, W.; Weiner, P.; Wilkinson, A., Eds. Springer Netherlands: 1997; Vol. 3, pp 83-96.

26. (a) Wickstrom, L.; Okur, A.; Simmerling, C., Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophys J* **2009,** *97* (3), 853-856; (b) Cerutti, D. S.; Freddolino, P. L.; Duke, R. E.; Case, D. A., Simulations of a Protein Crystal with a High

Resolution X-ray Structure: Evaluation of Force Fields and Water Models. *Journal of Physical Chemistry B* **2010,** *114* (40), 12811-12824; (c) Lange, O. F.; van der Spoel, D.; de Groot, B. L., Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. *Biophys J* **2010,** *99* (2), 647-655.

27. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P., A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry* **2003,** *24* (16), 1999-2012.

28. Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H., Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *Journal of the American Chemical Society* **2007,** *129* (5), 1179-89.

29. (a) Best, R. B.; Buchete, N. V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophys J* **2008,** *95* (1), L7-L9; (b) Best, R. B.; Hummer, G., Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *Journal of Physical Chemistry B* **2009,** *113* (26), 9004-9015; (c) Seabra, G. D.; Walker, R. C.; Roitberg, A. E., Are Current Semiempirical Methods Better Than Force Fields? A Study from the Thermodynamics Perspective. *J Phys Chem A* **2009,** *113* (43), 11938-11948.

30. Li, D. W.; Bruschweiler, R., Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. *Journal of chemical theory and computation* **2011,** *7* (6), 1773-1782.

31. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* **2010,** *78* (8), 1950-1958.

32. Best, R. B.; de Sancho, D.; Mittal, J., Residue-Specific alpha-Helix Propensities from Molecular Simulation. *Biophys J* **2012,** *102* (6), 1462-1467.

33. Reiher, I. W., Theoretical studies of hydrogen bonding. *PhD Thesis at Harvard University.* **1985**.

34. (a) Hiltpold, A.; Ferrara, P.; Gsponer, J.; Caflisch, A., Free energy surface of the helical peptide Y(MEARA)(6). *Journal of Physical Chemistry B* **2000,** *104* (43), 10080-10086; (b) Feig, M.; MacKerell, A. D.; Brooks, C. L., Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations. *Journal of Physical Chemistry B* **2003,** *107* (12), 2831-2836.

35. (a) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* **2004,** *25* (11), 1400-15; (b) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society* **2004,** *126* (3), 698-9.

36. Jorgensen, W. L.; Tiradorives, J., The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *Journal of the American Chemical Society* **1988,** *110* (6), 1657-1666.

37. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B* **2001,** *105* (28), 6474-6487.

38. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of chemical theory and computation* **2015,** *11* (7), 3499-3509.

39. Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L., Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *Journal of chemical theory and computation* **2015,** *11* (7), 3499-509.

40. (a) Piana, S.; Lindorff-Larsen, K.; Shaw, David E., How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys J* **2011,** *100* (9), L47-L49; (b) Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S., Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of chemical theory and computation* **2012,** *8* (4), 1409-1414; (c) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Systematic Validation of Protein Force Fields against Experimental Data. *Plos One* **2012,** *7* (2); (d) Robustelli, P.; Piana, S.; Shaw, D. E., Developing a molecular dynamics force field for both folded and disordered protein states. *P Natl Acad Sci USA* **2018,** *115* (21), E4758-E4766; (e) Koes, D. R.; Vries, J. K., Evaluating amber force fields using computed NMR chemical shifts. *Proteins-Structure Function and Bioinformatics* **2017,** *85* (10), 1944-1956.

41. Verlet, L., Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys Rev* **1967,** *159* (1), 98-+.

42. Hockney, R. W.; Goel, S. P.; Eastwood, J. W., 10000 Particle Molecular Dynamics Model with Long-Range Forces. *Chem Phys Lett* **1973,** *21* (3), 589-591.

43. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How Fast-Folding Proteins Fold. *Science* **2011,** *334* (6055), 517-520.

44. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015,** *137* (7), 2695-2703.

45. (a) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *Journal of chemical theory and computation* **2005,** *1* (1), 153-168; (b) Lamoureux, G.; MacKerell, A. D.; Roux, B., A simple polarizable model of water based on classical Drude oscillators. *J Chem Phys* **2003,** *119* (10), 5185-5197; (c) Ponder, J. W.; Wu, C. J.; Ren, P. Y.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current Status of the AMOEBA Polarizable Force Field. *Journal of Physical Chemistry B* **2010,** *114* (8), 2549-2564; (d) Lopes, P. E. M.; Zhu, X.; Lau, A.; Roux, B.; MacKerell, A. D.,

Development of the Charmm Polarizable Force Field for Polypeptides Based on Drude Oscillators. *Biophys J* **2011,** *100* (3), 612-612.

46. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-88.

47. (a) Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C., Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *Journal of computational chemistry* **2003,** *24* (1), 21-31; (b) Garcia, A. E.; Sanbonmatsu, K. Y., alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. *P Natl Acad Sci USA* **2002,** *99* (5), 2782-2787; (c) Kamiya, N.; Higo, J.; Nakamura, H., Conformational transition states of a beta-hairpin peptide between the ordered and disordered conformations in explicit water. *Protein Science* **2002,** *11* (10), 2297-2307; (d) Higo, J.; Ito, N.; Kuroda, M.; Ono, S.; Nakajima, N.; Nakamura, H., Energy landscape of a peptide consisting of alpha-helix, 3(10)-helix, beta-turn, beta-hairpin, and other disordered conformations. *Protein Science* **2001,** *10* (6), 1160-1171; (e) Ono, S.; Nakajima, N.; Higo, J.; Nakamura, H., Peptide free-energy profile is strongly dependent on the force field: Comparison of C96 and AMBER95. *Journal of computational chemistry* **2000,** *21* (9), 748-762; (f) Wang, L.; Duan, Y.; Shortle, R.; Imperiali, B.; Kollman, P. A., Study of the stability and unfolding mechanism of BBA1 by molecular dynamics simulations at different temperatures. *Protein Science* **1999,** *8* (6), 1292-1304.

48. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A., Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *Journal of the American Chemical Society* **1997,** *119* (25), 5908-5920.

49. Janowski, P. A.; Liu, C. M.; Deckman, J.; Case, D. A., Molecular dynamics simulation of triclinic lysozyme in a crystal lattice. *Protein Science* **2016,** *25* (1), 87-102.

50. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell Jr., A. D., Optimization of the Additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012,** *8* (9), 3257-3273.

51. Best, R. B.; Zheng, W.; Mittal, J., Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of chemical theory and computation* **2014,** *10* (11), 5113-5124.

52. Georgoulia, P. S.; Glykos, N. M., Using J-Coupling Constants for Force Field Validation: Application to Hepta-alanine. *Journal of Physical Chemistry B* **2011,** *115* (51), 15221-15227.

53. Creamer, T. P.; Rose, G. D., Side-Chain Entropy Opposes Alpha-Helix Formation but Rationalizes Experimentally Determined Helix-Forming Propensities. *P Natl Acad Sci USA* **1992,** *89* (13), 5937-5941.

54. Shell, M. S.; Ritterson, R.; Dill, K. A., A test on peptide stability of AMBER force fields with implicit solvation. *Journal of Physical Chemistry B* **2008,** *112* (22), 6878-6886.

55. Chen, J.; Im, W.; Brooks, C. L., Balancing Solvation and Intramolecular Interactions: Toward a Consistent Generalized Born Force Field. *Journal of the American Chemical Society* **2006,** *128* (11), 3728-3736.

56. Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A., ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *Journal of chemical theory and computation* **2014,** *10* (10), 4515-4534.

57. Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T., Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **2004,** *120* (20), 9665-9678.

58. Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T., Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *Journal of chemical theory and computation* **2016,** *12* (8), 3926-3947.

59. Takemura, K.; Kitao, A., Water Model Tuning for Improved Reproduction of Rotational Diffusion and NMR Spectral Density. *Journal of Physical Chemistry B* **2012,** *116* (22), 6279-6287.

60. Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H., Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of chemical theory and computation* **2015,** *11* (11), 5513-5524.

61. Piana, S.; Klepeis, J. L.; Shaw, D. E., Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current opinion in structural biology* **2014,** *24*, 98-105.

62. Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A., Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. *Journal of Chemical Theory and Computation* **2015,** *11* (1), 2-7.

63. (a) Jiang, F.; Zhou, C. Y.; Wu, Y. D., Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L. *Journal of Physical Chemistry B* **2014,** *118* (25), 6983-6998; (b) Wang, W.; Ye, W.; Jiang, C.; Luo, R.; Chen, H. F., New force field on modeling intrinsically disordered proteins. *Chemical biology & drug design* **2014,** *84* (3), 253-69; (c) Song, D.; Wang, W.; Ye, W.; Ji, D.; Luo, R.; Chen, H. F., ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chemical biology & drug design* **2016**.

64. (a) Henriques, J.; Cragnell, C.; Skepo, M., Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *Journal of chemical theory and computation* **2015,** *11* (7), 3420-3431; (b) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E., Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* **2015,** *119* (16), 5113-5123.

65. Petrov, D.; Zagrovic, B., Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *PLoS Comput Biol* **2014,** *10* (5), e1003638.

66. Izadi, S.; Anandakrishnan, R.; Onufriev, A. V., Building Water Models: A Different Approach. *J Phys Chem Lett* **2014,** *5* (21), 3863-3871.

67. Shabane, P. S.; Izadi, S.; Onufriev, A. V., A general purpose water model can improve atomistic simulations of intrinsically disordered proteins. *Journal of chemical theory and computation* **2019**.

68. Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A., Advances in free-energy-based simulations of protein folding and ligand binding. *Current opinion in structural biology* **2016,** *36*, 25-31.

69. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C., Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* **2008,** *51* (7), 91-97.

70. Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C., Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *Journal of the American Chemical Society* **2014,** *136* (40), 13959-13962.

71. Perez, A.; MacCallum, J. L.; Brini, E.; Simmerling, C.; Dill, K. A., Grid-Based Backbone Correction to the ff12SB Protein Force Field for Implicit-Solvent Simulations. *Journal of chemical theory and computation* **2015,** *11* (10), 4770-4779.

72. Moreau, R. J.; Schubert, C. R.; Nasr, K. A.; Torok, M.; Miller, J. S.; Kennedy, R. J.; Kemp, D. S., Context-independent, temperature-dependent helical propensities for amino acid residues. *Journal of the American Chemical Society* **2009,** *131* (36), 13107-16.

73. Cornish, V. W.; Kaplan, M. I.; Veenstra, D. L.; Kollman, P. A.; Schultz, P. G., Stabilizing and Destabilizing Effects of Placing .beta.-Branched Amino Acids in Protein .alpha.-Helixes. *Biochemistry-Us* **1994,** *33* (40), 12022-12031.

74. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The penultimate rotamer library. *Proteins* **2000,** *40* (3), 389-408.

75. (a) Zhou, C. Y.; Jiang, F.; Wu, Y. D., Residue-Specific Force Field Based on Protein Coil Library. RSFF2: Modification of AMBER ff99SB. *Journal of Physical Chemistry B* **2015,** *119* (3), 1035-1047; (b) Kang, W.; Jiang, F.; Wu, Y. D., Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition. *Journal of chemical theory and computation* **2018,** *14* (8), 4474-4486.

76. Song, D.; Luo, R.; Chen, H. F., The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **2017,** *57* (5), 1166-1178.

77. Feller, S. E.; MacKerell, A. D., An improved empirical potential energy function for molecular simulations of phospholipids. *Journal of Physical Chemistry B* **2000,** *104* (31), 7510-7515.

78. Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D., Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme. *Biophys J* **2006,** *90* (4), L36-L38.

79. Wang, L.-P.; Martinez, T. J.; Pande, V. S., Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters* **2014,** *5* (11), 1885-1891.

80. Mackerell, A. D., Empirical force fields for biological macromolecules: Overview and issues. *Journal of computational chemistry* **2004,** *25* (13), 1584-1604.

81. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P., A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comp. Chem.* **2003,** *24* (16), 1999-2012.

82. Zgarbova, M.; Luque, F. J.; Sponer, J.; Otyepka, M.; Jurecka, P., A Novel Approach for Deriving Force Field Torsion Angle Parameters Accounting for Conformation-Dependent Solvation Effects. *Journal of chemical theory and computation* **2012,** *8* (9), 3232-3242.

83. Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A., Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *Journal of Physical Chemistry B* **2013,** *117* (8), 2328-2338.

84. Zgarbova, M.; Otyepka, M.; Banas, P.; Luque, F. J.; Cheatham, T. E.; Sponer, J.; Jurecka, P., Refinement of force field torsion parameters for nucleic acids based on inclusion of conformation-dependent solvation effects. *J Biomol Struct Dyn* **2013,** *31*, 70-70.

85. Jurecka, P.; Zgarbova, M.; Luque, F. J.; Sponer, J.; Otyepka, M., Deriving force field dihedral angle parameters that account for conformation-dependent solvation effects. *J Biomol Struct Dyn* **2013,** *31*, 67-68.

86. Wang, L. P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martinez, T. J.; Pande, V. S., Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *Journal of Physical Chemistry B* **2017,** *121* (16), 4023-4039.

87. (a) Best, R. B.; Zheng, W. W.; Mittal, J., Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of chemical theory and computation* **2014,** *10* (11), 5113-5124; (b) Petrov, D.; Zagrovic, B., Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments? *Plos Comput Biol* **2014,** *10* (5); (c) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E., Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *Journal of Physical Chemistry B* **2015,** *119* (16), 5113-5123.

88. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** *28* (1), 235-242.

89. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1983,** *79* (2), 926-935.

90. Izadi, S.; Onufriev, A. V., Accuracy limit of rigid 3-point water models. *J Chem Phys* **2016,** *145* (7).

91. Wang, L. P.; Martinez, T. J.; Pande, V. S., Building force fields: An automatic, systematic and reproducible approach. *Abstr Pap Am Chem S* **2014,** *248*.

92. Nguyen, H.; Roe, D. R.; Simmerling, C., Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of chemical theory and computation* **2013,** *9* (4), 2020-2034.

93. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A., Generalized Born model with a simple, robust molecular volume correction. *Journal of chemical theory and computation* **2007,** *3* (1), 156-169.

94. Wang, L. P.; Martinez, T. J.; Pande, V. S., Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J Phys Chem Lett* **2014,** *5* (11), 1885-1891.

95. Song, K.; Stewart, J. M.; Fesinmeyer, R. M.; Andersen, N. H.; Simmerling, C., Structural insights for designed alanine-rich helices: Comparing NMR helicity measures and conformational ensembles from molecular dynamics simulation. *Biopolymers* **2008,** *89* (9), 747-760.

96. Honda, S.; Akiba, T.; Kato, Y. S.; Sawada, Y.; Sekijima, M.; Ishimura, M.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K., Crystal Structure of a Ten-Amino Acid Protein. *Journal of the American Chemical Society* **2008,** *130* (46), 15327-15331.

97. Davis, C. M.; Xiao, S. F.; Raeigh, D. P.; Dyer, R. B., Raising the Speed Limit for beta-Hairpin Formation. *Journal of the American Chemical Society* **2012,** *134* (35), 14476-14482.

98. Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A., Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *Journal of the American Chemical Society* **2003,** *125* (30), 9179-9191.

99. Vijaykumar, S.; Bugg, C. E.; Cook, W. J., Structure of Ubiquitin Refined at 1.8 a Resolution. *J Mol Biol* **1987,** *194* (3), 531-544.

100. Young, A. C. M.; Dewan, J. C.; Nave, C.; Tilton, R. F., Comparison of Radiation-Induced Decay and Structure Refinement from X-Ray Data Collected from Lysozyme Crystals at Low and Ambient-Temperatures. *J Appl Crystallogr* **1993,** *26*, 309-319.

101. D.A. Case, R. M. B., D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke,; A.W. Goetz, N. H., S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C.; Lin, T. L., R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I.; Omelyan, A. O., D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails,; R.C. Walker, J. W., R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman, AMBER 2016. *University of California, San Francisco*.

102. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

103. Weiser, J.; Shenkin, P. S.; Still, W. C., Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. *Biopolymers* **1999,** *50* (4), 373-80.

104. Lovell, S. C.; Davis, I. W.; Adrendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C., Structure validation by C alpha geometry: phi,psi and C beta deviation. *Proteins-Structure Function and Genetics* **2003,** *50* (3), 437-450.

105. Crowley, M. F.; Williamson, M. J.; Walker, R. C., CHAMBER: Comprehensive Support for CHARMM Force Fields Within the AMBER Software. *Int J Quantum Chem* **2009,** *109* (15), 3767-3772.

106. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The journal of physical chemistry. B* **2009,** *113* (18), 6378-96.

107. Scalmani, G.; Frisch, M. J., Continuous surface charge polarizable continuum models of solvation. I. General formalism. *J Chem Phys* **2010,** *132* (11).

108. Zhao, Y.; Schultz, N. E.; Truhlar, D. G., Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *Journal of chemical theory and computation* **2006,** *2* (2), 364-382.

109. Frisch, M.; Trucks, G.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G., Gaussian 09, revision a. 02, gaussian. *Inc., Wallingford, CT* **2009,** *200*.

110. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010,** *132* (15).

111. (a) Besler, B. H.; Merz, K. M.; Kollman, P. A., Atomic Charges Derived from Semiempirical Methods. *Journal of computational chemistry* **1990,** *11* (4), 431-439; (b) Singh, U. C.; Kollman, P. A., An Approach to Computing Electrostatic Charges for Molecules. *Journal of computational chemistry* **1984,** *5* (2), 129-145.

112. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* **1977,** *23* (3), 327-341.

113. Sagui, C.; Pedersen, L. G.; Darden, T. A., Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J Chem Phys* **2004,** *120* (1), 73-87.

114. Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E., Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of chemical theory and computation* **2015,** *11* (4), 1864-1874.

115. Roe, D. R.; Cheatham, T. E., PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of chemical theory and computation* **2013,** *9* (7), 3084-3095.

116. Lifson, S., Theory of Helix-Coil Transition in Polypeptides. *J Chem Phys* **1961,** *34* (6), 1963-&.

117. Pace, C. N.; Scholtz, J. M., A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* **1998,** *75* (1), 422-427.

118. Efron, B., Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion. *Ann Stat* **1986,** *14* (4), 1301-1304.

119. Karplus, M., Contact Electron-Spin Coupling of Nuclear Magnetic Moments. *J Chem Phys* **1959,** *30* (1), 11-15.

120. Hu, J. S.; Bax, A., Determination of phi and chi(1) angles in proteins from C-13-C-13 three-bond J couplings measured by three-dimensional heteronuclear NMR. How planar is the peptide bond? *Journal of the American Chemical Society* **1997,** *119* (27), 6360-6368.

121. Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L., Intrinsic backbone preferences are fully present in blocked amino acids. *P Natl Acad Sci USA* **2006,** *103* (5), 1272-1277.

122. Onufriev, A.; Bashford, D.; Case, D. A., Modification of the generalized Born model suitable for macromolecules. *Journal of Physical Chemistry B* **2000,** *104* (15), 3712-3720.

123. Swails, J. M.; York, D. M.; Roitberg, A. E., Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *Journal of chemical theory and computation* **2014,** *10* (3), 1341-1352.

124. Lipari, G.; Szabo, A., Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .2. Analysis of Experimental Results. *Journal of the American Chemical Society* **1982,** *104* (17), 4559-4570.

125. Prompers, J. J.; Bruschweiler, R., General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. *Journal of the American Chemical Society* **2002,** *124* (16), 4522-4534.

126. Gu, Y.; Li, D. W.; Bruschweiler, R., NMR Order Parameter Determination from Long Molecular Dynamics Trajectories for Objective Comparison with Experiment. *Journal of chemical theory and computation* **2014,** *10* (6), 2599-2607.

127. Markwick, P. R. L.; Bouvignies, G.; Blackledge, M., Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *Journal of the American Chemical Society* **2007,** *129* (15), 4724-4730.

128. (a) Richardson, J. S.; Prisant, M. G.; Richardson, D. C., Crystallographic model validation: from diagnosis to healing. *Current opinion in structural biology* **2013,** *23* (5), 707-14; (b) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B., 3rd; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., MolProbity: More and better reference data for improved all-atom structure validation. *Protein science : a publication of the Protein Society* **2018,** *27* (1), 293-315.

129. Kabsch, W.; Sander, C., Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** *22* (12), 2577-2637.

130. Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009,** *25* (11), 1422-3.

131. (a) Schrauber, H.; Eisenhaber, F.; Argos, P., Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol* **1993,** *230* (2), 592-612; (b) Aurora, R.; Creamer, T. P.; Srinivasan, R.; Rose, G. D., Local interactions in protein folding: Lessons from the alpha-helix. *J Biol Chem* **1997,** *272* (3), 1413-1416; (c) De Maeyer, M.; Desmet, J.; Lasters, I., All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding & design* **1997,** *2* (1), 53-66.

132. Chothia, C.; Lesk, A. M., The Relation between the Divergence of Sequence and Structure in Proteins. *Embo J* **1986,** *5* (4), 823-826.

133. Best, R. B.; Mittal, J., Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *Journal of Physical Chemistry B* **2010,** *114* (46), 14916-14923.

134. Abascal, J. L. F.; Vega, C., A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys* **2005,** *123* (23).

135. (a) Doig, A. J.; Baldwin, R. L., N- and C-Capping Preferences for All 20 Amino-Acids in Alpha-Helical Peptides. *Protein Science* **1995,** *4* (7), 1325-1336; (b) Rohl, C. A.; Chakrabartty, A.; Baldwin, R. L., Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Protein Science* **1996,** *5* (12), 2623-2637.

136. Hall, J. B.; Fushman, D., Characterization of the overall and local dynamics of a protein with intermediate rotational anisotropy: Differentiating between conformational exchange and anisotropic diffusion in the B3 domain of protein G. *J Biomol Nmr* **2003,** *27* (3), 261-275.

137. Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A., Rotational diffusion anisotropy of human ubiquitin from N-15 NMR relaxation. *Journal of the American Chemical Society* **1995,** *117* (50), 12562-12566.

138. Buck, M.; Boyd, J.; Redfield, C.; Mackenzie, D. A.; Jeenes, D. J.; Archer, D. B.; Dobson, C. M., Structural Determinants of Protein Dynamics - Analysis of N-15 Nmr Relaxation Measurements for Main-Chain and Side-Chain Nuclei of Hen Egg-White Lysozyme. *Biochemistry-Us* **1995,** *34* (12), 4041-4055.

139. Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C., Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *Journal of Physical Chemistry B* **2007,** *111* (7), 1846-1857.

140. Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C., Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *Journal of chemical theory and computation* **2006,** *2* (2), 420-433.

141. (a) Pang, Y.-P., Use of 1–4 interaction scaling factors to control the conformational equilibrium between α-helix and β-strand. *Biochemical and Biophysical Research Communications* **2015,** *457* (2), 183-186; (b) Pang, Y.-P., FF12MC: A revised AMBER forcefield and new protein simulation protocol. *Proteins: Structure, Function, and Bioinformatics* **2016,** *84* (10), 1490-1516.

142. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.;

Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Gaussian, Inc.: Wallingford, CT, USA, 2009.

143. Onufriev, A.; Bashford, D.; Case, D. A., Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function and Bioinformatics* **2004,** *55* (2), 383-394.

144. Fogolari, F.; Brigo, A.; Molinari, H., The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* **2002,** *15* (6), 377-392.

145. Tan, C. H.; Yang, L. J.; Luo, R., How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *Journal of Physical Chemistry B* **2006,** *110* (37), 18680-18687.

146. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993,** *98* (12), 10089-10092.

147. Zou, J.; Tian, C.; Simmerling, C., Blinded prediction of protein-ligand binding affinity using Amber thermodynamic integration for the 2018 D3R grand challenge 4. *J Comput Aided Mol Des* **2019**.

148. Onsager, L., Electric moments of molecules in liquids. *Journal of the American Chemical Society* **1936,** *58*, 1486-1493.

149. Born, M., Volumes and hydration warmth of ions. *Z Phys* **1920,** *1*, 45-48.

150. (a) Lange, A. W.; Herbert, J. M., Improving Generalized Born Models by Exploiting Connections to Polarizable Continuum Models. II. Corrections for Salt Effects. *Journal of chemical theory and computation* **2012,** *8* (11), 4381-4392; (b) Lange, A. W.; Herbert, J. M., Improving Generalized Born Models by Exploiting Connections to Polarizable Continuum Models. I. An Improved Effective Coulomb Operator. *Journal of chemical theory and computation* **2012,** *8* (6), 1999-2011.

151. (a) Bottaro, S.; Lindorff-Larsen, K., Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018,** *361* (6400), 355-+; (b) Nerenberg, P. S.; Head-Gordon, T., New developments in force fields for biomolecular simulations. *Current opinion in structural biology* **2018,** *49*, 129-138; (c) Riniker, S., Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J Chem Inf Model* **2018,** *58* (3), 565-578; (d) Yoo, J.; Aksimentiev, A., New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Phys Chem Chem Phys* **2018,** *20* (13), 8432-8449; (e) Dauber-Osguthorpe, P.; Hagler, A. T., Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? *J Comput Aid Mol Des* **2019,** *33* (2), 133-203.

152. (a) Chen, W.; Shi, C. Y.; MacKerell, A. D.; Shen, J., Conformational Dynamics of Two Natively Unfolded Fragment Peptides: Comparison of the AMBER and CHARMM Force Fields. *Journal of Physical Chemistry B* **2015,** *119* (25), 7902-7910; (b) Maffucci, I.; Contini, A., An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, beta-Hairpin, and Intrinsically Disordered Peptides. *Journal of chemical theory and computation* **2016,** *12* (2), 714-727; (c) Smith, M. D.; Rao, J. S.; Segelken, E.; Cruz, L., Force-Field Induced Bias in the Structure of A beta(21-30): A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields. *J Chem Inf Model* **2015,** *55* (12), 2587-2595; (d) Huang,

J.; MacKerell, A. D., Force field development and simulations of intrinsically disordered proteins. *Current opinion in structural biology* **2018,** *48*, 40-48.