Stony Brook University

On the derivation of accurate force field parameters for molecular mechanics simulations

A Dissertation Presented

by

James Allen Maier

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Biochemistry and Structural Biology

Stony Brook University

May 2015

Stony Brook University

The Graduate School

James Allen Maier

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

David Green – Chairperson of Defense
Associate Professor, Department of Applied Mathematics and Statistics

Ken Dill
Professor, Departments of Physics and Chemistry

Fernando Raineri
Lecturer
Stony Brook University, Department of Chemistry

This dissertation is accepted by the Graduate School.

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

On the derivation of accurate force field parameters for molecular mechanics simulations

by

James Allen Maier

Doctor of Philosophy

in

Biochemistry and Structural Biology

Stony Brook University

2015

Proteins carry out many diverse but important biological tasks, the understanding of which can be greatly augmented by theoretical methods that can generate microscopic insights. A popular method for simulating proteins is called molecular mechanics. Molecular mechanics drives the dynamics of molecules according to their potential energy surface as defined by a force field. Because force fields are simple, molecular mechanics can be fast; but force fields must simultaneously be accurate enough for the conformational ensembles they generate to be useful.

One force field that has been widely adopted for its utility is AMBER force field 99 Stony Brook (ff99SB). The ff99SB protein backbone parameters were fit to quantum mechanics energies of glycine and alanine tetrapeptides, including a set of minimum energy conformations in the gas-phase. Although ff99SB rigorously reproduces many thermodynamic properties, it has shortcomings. Issues with backbone parameters may result from training against only energetic minima or from the energy calculations being in the gas phase.

Problems with side chain parameters can stem from the protocol of ff99, where the amino acid side chain parameters were trained against energies of small molecules, while transferability from small molecules to amino acids may be problematic. Small updates to the backbone potential were applied by several groups, as well as the Simmerling group as part of ff14SB.

Whereas ff99SB and ff14SB are fixed-charge, additive molecular mechanical models, there are also molecular mechanical models that include non-additive effects like charge polarization. Polarizable force fields, with their many additional degrees of freedom, promise enhanced accuracy relative to fixed charge force fields. But with so many degrees of freedom and thus parameters, polarizable force fields can be more difficult to train. Although this complexity may be overcome, it is unclear whether the utility of fixed-charge, additive force fields has been exhausted, warranting the great endeavors of developing a polarizable model.

This dissertation seeks to answer how much more fixed charge force fields can be improved. Specifically, this work addresses two questions. Firstly, can the side chain parameters of ff99SB be improved by fitting to quantum mechanics energies? We investigated different options in the calculation of energies for parameter training, finding that how the structures were minimized can significantly affect transferability of parameters trained against them. Specifically, we found that loosely restraining the side chains, which were being refined, and tightly restraining the backbone, which was not, made the errors most similar between α and β backbone contexts. This transferability can be measured by improved agreement with the quantum mechanics training set as well as experimental scalar couplings.

Secondly, can the backbone parameters of ff99SB be made more accurate, alternatively to empirical tweaks, by another, improved fitting to quantum mechanics energies? We found that better reproduction of NMR solution scalar couplings was possible, if energy calculations included solvation effects, full grids of structures were included, and, perhaps surprisingly, if parameters were extrapolated to those appropriate for a zero-length peptide.

These results show that quantum mechanics can be effectively used to improve the accuracy of molecular mechanics force fields. These improvements have implications for protein structure prediction,

aiding the successful folding of 16 of 17 proteins in GB-Neck2 implicit solvent. Beyond, the insights from the QM-based backbone training could be extended to develop residue-specific parameters that bolster the sequence-dependent structural preferences of proteins in simulation models.

영원한 나의 신부 수홍, 그리고 하나님의 선물 노아에게

Dedicated to my eternal bride Grace, and to my divine treasure Noah

Contents

List of Figures

List of Tables

List of Abbreviations

AA	Amino acid
AAE	Average absolute error
ANE	Average normalized error
BBD	Backbone dependence
ff99SB	Force field 99 Stony Brook
ff12SB	Force field 12 Stony Brook
ff14SB	Force field 14 Stony Brook
GB	Generalized Born
MM	Molecular Mechanics
OPC	Optimized point charge
PB	Poisson Boltzmann
PME	Particle-mesh Ewald
RMSD	Root mean square deviations
SCD	Side chain dependence
vdW	van der Waals

Acknowledgements

The work presented in this dissertation would not be possible without the essential contributions and assistance from many individuals. Firstly, I must thank my advisor, **Professor Carlos Simmerling**. When I had only heard of molecular dynamics from his website, Professor Simmerling allowed me to study in his lab. He gave me the independence to pursue projects in ways I saw fit, and yet did not leave me entirely alone. I am grateful for his ability to direct my endeavors and yet allow me to grow as a young scientist.

Secondly, I thank my committee, present—**Professors David Green**, **Fernando Raineri**, and **Ken Dill**—and past—**Professors Dan Raleigh**, **Miguel García-Diaz**, and **Yuefan Deng**. I appreciate their wisdom and guidance, good ideas and helping keep me on track. It amazes me that anyone can come to my project and make meaningful suggestions without intimate experience.

Carmenza Martinez, my partner on the "M&M" (Maier and Martinez) force field. Thank you for your patience with my myriad side chain revisions. Thank you also for your advice regarding Noah, and for generally being a nice coworker and friend.

Kevin Hauser, my next door neighbor through this long ride. From sitting less than one foot away from each other in *the Dungeon* to living up the two-foot distance proffered by the ever-swanky Laufer Center, your presence and partnership have been a tremendous component of my PhD. We shared so many of our problems—some perplexing, some small, some silly. At times I would help him; other times he would help me; always we would deplete the espresso. Fresh insight and collaboration are important scientific skills. And friendship is an invaluable asset.

Hai Nguyen was my first friend in the lab; thank you for the times, when you were a junior graduate student and I was between undergrad and grad school, going to the hospital every day (for lunch). For always being willing to help me out with whatever, and for being a generally nice presence during these times, I appreciate your companionship.

I thank the remaining Simmerling Lab members, those with whom I overlapped and otherwise. In chronological order of thesis defense: Salma Rafi, Lauren Wickstrom, Fangyu Ding, A.J. Campbell, Christina Bergonzo, Amber Carr, Yi (Miranda) Shang, Cheng-Tsung (Eric) Lai, Haoquan Li. Thanks to the youngin's—Kenneth Lam, He (Agnes) Huang, and Koushik Kasavajhala. And a special thanks to Dan Roe—when I attended my first Simmerling Lab meeting, I saw Dan, and thought that the lab members were cool (unaware that Dan was an alumnus that I actually would not see again for quite some time).

I must also thank my family—**Maier**s, **Kim**s, and **Hong**. I thank my parents, first and foremost, for birthing and raising me. I owe my first exposures to science and mathematics to their influence and my father's seemingly unending interests and patience. I then must thank my brother, Gregory, my *partner in crime* of many years, during which I like to think we not only taxed, but impressed our parents. I also thank my sister, Samantha, for inspiring me to be a good role model and for growing into a loving, encouraging young lady. Additionally, I thank my 부모님, for accepting me into their family and allowing me to marry their daughter. Thank you for your generosity during this time. Jessica and Jake, I appreciate all your warmth and company, as well.

And finally, I am most grateful to my lovely wife, **Grace**, and our son **Noah**. Thank you, Grace, for marrying me. Without your love and support, I don't know that I would have made it through to the PhD. Your homemade AmeriKoreItaliMexican food always gave me the "food for thought" I needed to continue the sometimes exhausting research through another day. And your brownies et al. comforted me. I can't imagine that anyone has a better partner or companion. And Noah, my little sidekick; to see you smile is a cool breeze on a summer day; to hear you laugh, an oasis in the high sun. But even when you cry, it's hard to believe a baby so sweet and adorable came from me. I am sorry for the sacrifices my family has faced, but I hope you will remember better times than these. I thank you both, for your enduring love.

On a more technical note I must acknowledge **Donald Knuth** and **Leslie Lamport** for developing TEX and LATEX, respectively, the **authors of TEXstudio**, and **Benjamin Hornberger** for a Stony Brook dissertation LATEX class file, without which this dissertation would have been much more difficult to write. I also must acknowledge **Bjarne Stroustrop** and **Guido van Rossum** for developing C++ and Python, respectively, which were used to build many optimization, analysis, and visualization tools. And I thank the **AMBER developers**, for molecular modeling tools that were used extensively in this work.

Chapter 1

Introduction

> "Nature is pleased with simplicity. And nature is no dummy"
> – Isaac Newton

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. One highlight is the time-resolved, atomic-detail folding of ubiquitin during a 1 ms simulation [Piana et al., 2013]. The conformational sampling of such extensive simulations is driven by an energy landscape defined by simple functions called *force fields*. Despite increasing accomplishments, force fields have limitations. One limitation—the focus of this dissertation, for proteins—is accuracy. Before exploring how accuracy can be improved, a brief history and overview of atomistic force field methods will be presented.

First, however, proteins and their structure—the modeling of which is the subject of this dissertation—will be described. Then, the main tenets of statistical mechanics, which allows the connection between microscopic details of molecular (in this case, protein) behavior and macroscopic observables, will be summarized. Following statistical mechanics, classical, quantum, and molecular mechanics that provide microscopic information will be presented. From

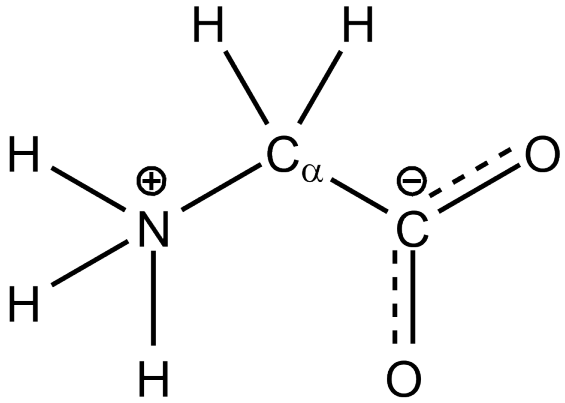

Figure 1.1: A glycine amino acid in zwitterionic form with protonated amino and deprotonated carboxylic groups.

molecular mechanics, the introduction will switch to force fields, general principles of force field development, and a recent history of AMBER protein force fields. Finally, the reader will be prepared for an outline of the dissertation that aims to improve the accuracy of protein force fields.

1.1 Proteins

Proteins are known as the *doers of the cell*. Some of their responsibilities are enzymatic, structural, or informational. The diverse tasks proteins carry out are enabled by the "alphabet" of twenty natural *amino acids* from which proteins are composed—the simplest of which, glycine, is depicted in Figure 1.1. α-amino acids possess an amino (N) base, then a carbon (Cα), and finally a carboxylic (C) acid. The Cα is often bound to a side chain that imbues the amino acid with the side chain's chemical functionality. Possessing both a base and an acid, amino acids readily form zwitterions in aqueous solution at neutral pH, possessing positively and negatively charged groups as the N-group becomes protonated while the C-group becomes deprotonated.

Sequences of amino acids, also called peptides or polypeptides, arise from polymerization, bonding the N-terminus of one amino acid to the C-terminus of another. By convention, an amino acid sequence is read from the amino acid with a free N-terminus to the amino acid with a free C-terminus—the order in

which polypeptides are synthesized in the ribosome. The order of amino acids is known as the primary structure.

Following condensation of two amino acids, the amino acids become joined by a peptide bond. The peptide bond has partial double bond character, due to the resonance shown in Figure 1.2, which arises from electron overlap between the carbonyl oxygen of one amino acid and the electron pair on the amino nitrogen of the next amino acid. Obeying this resonance, the peptide bond tends to be planar, assuming one of two stable isomers called *cis*, where the Cαs of two amino acids are on the same side of the double bond, or the preferred *trans* isomer, where the Cαs are on opposite sides of the double bond. The peptide bond therefore constrains protein structure because of its rigidity.

The flexibility in the protein main chain arises from rotation around the remaining bonds between N and Cα and between Cα and C. For a residue i (with previous residue $i-1$ and next residue $i+1$), rotation around the N_i–$C\alpha_i$ bond is canonically described by the ϕ torsion (C_{i-1}–N_i–$C\alpha_i$–C_i), while rotation around the $C\alpha_i$–C_i bond is canonically described by the ψ torsion (N_i–$C\alpha_i$–C_i–N_{i+1}) (Figure 1.3). As ϕ and ψ are the flexible dihedrals in the backbone, the conformation of a single amino acid can therefore be described by the values of the ϕ and ψ dihedrals. It is common to graph conformations of single amino acids in terms of (ϕ, ψ) coordinates. Such graphs of ψ versus ϕ are known as Ramachandran plots after Ramachandran et al. who developed the graph style [Ramachandran et al., 1963]. An example of a Ramachandran plot is depicted in Figure 1.4, using (ϕ, ψ) pairs for the second simplest amino acid, alanine, extracted from the protein data bank (PDB) [Lovell et al., 2003].

Due to favorable interactions between the polar amino and carbonyl groups of different amino acids within a peptide chain, certain conformations that facilitate N–H to O=C hydrogen bonds are preferred. The local arrangement of amino acids that leads to these hydrogen bonds is referred to as secondary structure. The prototypical secondary structure types arrange amino acids in a helix, referred to as α, or in neighboring extended strands, referred to as β (Figure 1.5).

Helices are most commonly right-handed (α_R, hereto referred to as simply

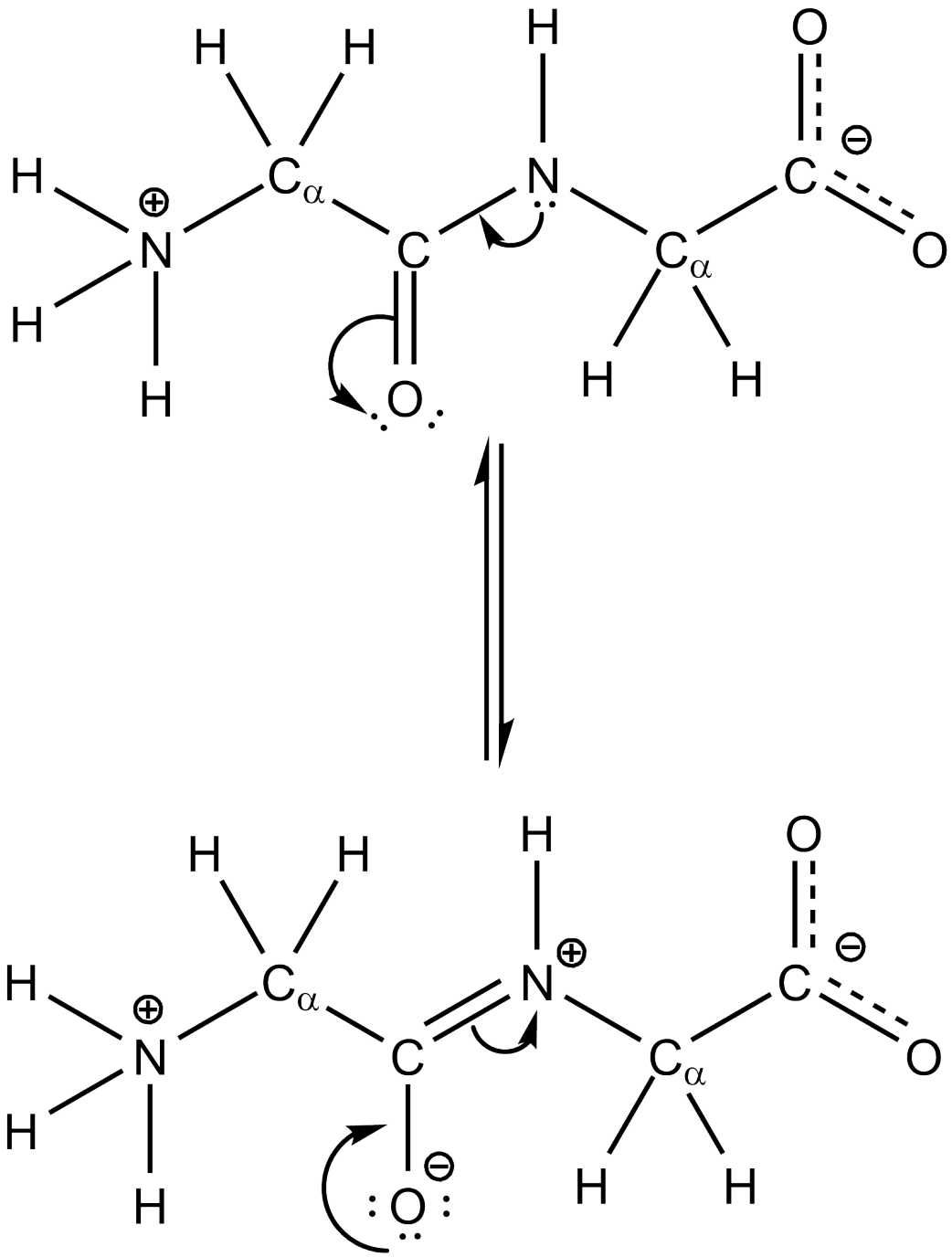

Figure 1.2: The resonance of the peptide bond that results in partial double bond character and peptide planarity.

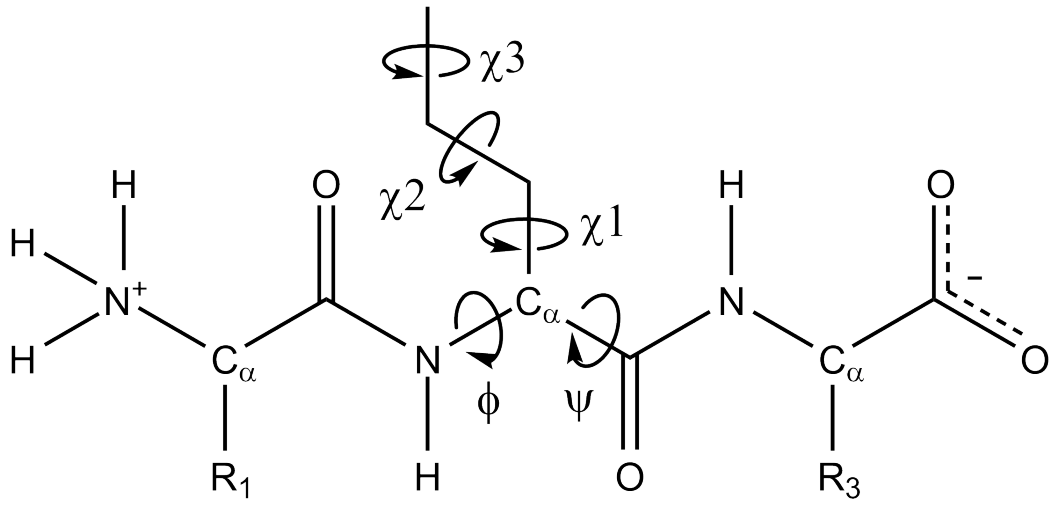

Figure 1.3: An oligomer of three amino acids. The torsions described by ϕ, ψ, and χ are indicated by curved arrows, with a label indicating the dihedral that measures that rotation.

α), but can also be left-handed (α_{L}). In both cases, helices allow hydrogen bonds between the C=O of residue i and the H–N of residue $i+4$, where i is any residue ID from the beginning of the helix to the end of the helix minus four (see helix in Figure 1.5A). Conventionally, hydrogen bonds are represented from the H to the O, thus helical hydrogen bonding can be represented as $i \leftarrow i+4$. There are also helical secondary structures with other hydrogen bonding patterns: 3_{10}-helix with $i \leftarrow i+3$ and π-helix with $i \leftarrow i+5$.

Alternatively, β-strands can be arranged parallel to each other, facilitating $j+2i \leftarrow k+2i$ and $k+2i \leftarrow j+2i+2$ interstrand hydrogen bonds, where j and k indicate the beginning of each strand and i is an integer from 0 to the number of residues with strand-strand interactions (see central two strands in Figure 1.5B). Or, β-strands can be antiparallel, facilitating $j+2i \leftarrow k-2i$ and $k-2i \leftarrow j+2i$ hydrogen bonds, where j is the beginning of one strand, k is the end of another, and i is again an integer from 0 to the number of residues with strand-strand interactions (see left two or right two strands in Figure 1.5B).

Tertiary structure describes the geometrical arrangement of strands and helices, as well as intervening turns and loops that do not form helices or

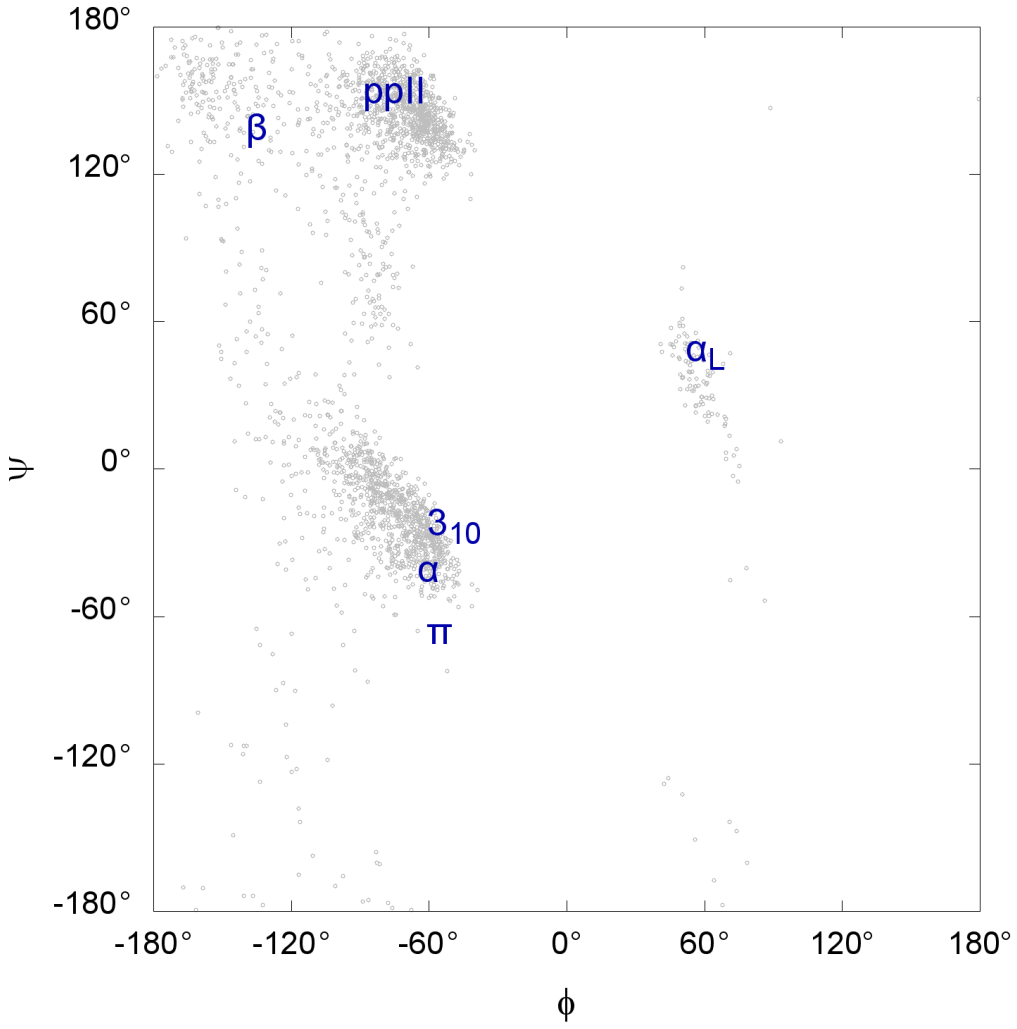

Figure 1.4: Typical ϕ/ψ coordinates of α $(-60°, -45°)$, $\alpha_L(60°, 45°)$, 3_{10} $(-49°, -26°)$, π $(-55°, -70°)$, β $(-135°, 135°)$, and ppII $(-75°, 150°)$ secondary structures on a Ramachandran plot of 500 alanine conformations [Lovell et al., 2003]. The ϕ/ψ coordinates of each secondary structure as specified are not exact, but approximately typical, average values.

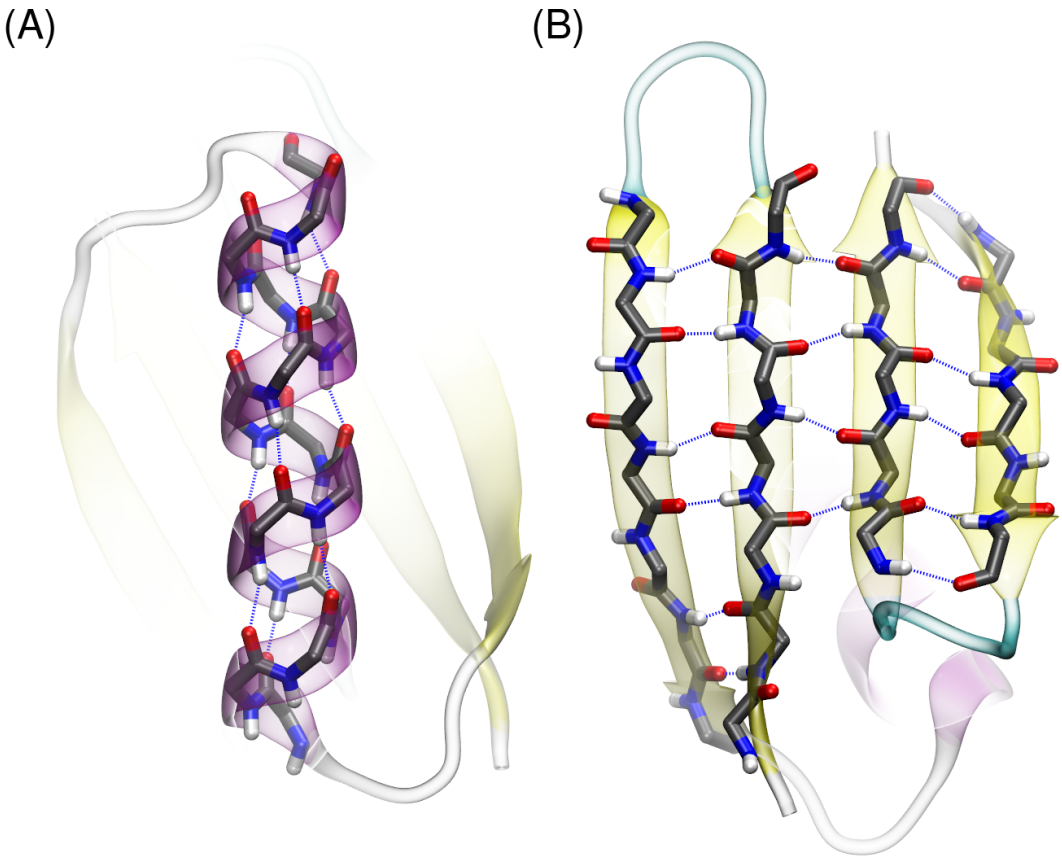

Figure 1.5: Characteristic hydrogen bonding of (A) α and (B) β secondary structures. In B, the two left and two right strands are antiparallel, whereas the two center strands are parallel. Both renderings are of an X-ray structure of GB3 refined with dipolar couplings [Ulmer et al., 2003], rendered using VMD [Humphrey et al., 1996].

strands. Tertiary structure intimately depends on key contacts between secondary structure units, as mediated by hydrophobic packing, salt bridges, and hydrogen bonding. Hydrogen bonding does not drive folding, as hydrogen bonds can form with water as easily as protein. But maintaining the total number of hydrogen bonds can be an important constraint on the folded protein structure, resulting in the motifs described above. Hydrophobic packing and salt bridge interactions between secondary structure units are usually mediated by the amino acid side chains, which thus define the preference of a sequence for a particular tertiary structure. Similar to the description of backbone conformation by ϕ and ψ dihedral angles, the conformation of the side chain can also be described by dihedral angles χ_1–χ_N (Figure 1.3), where N is the number of rotatable bonds in the side chain.

The main driving force of protein folding, however, is water. Liquid water forms a dynamic network of hydrogen bonds. When a hydrophobic solute is placed in water, water cannot form hydrogen bonds with it. To maintain the total number of hydrogen bonds, water interacts with itself in a restricted clathrate, cage-like structure. This results in a penalty as the number of accessible states, and thus the entropy, decreases. This entropic cost is lessened when hydrophobic chemicals aggregate, reducing the total surface area exposed to water. For proteins with hydrophobic and hydrophilic side chains, this "hydrophobic effect" drives hydrophobic side chains towards the core of the protein, away from water, while hydrophilic side chains remain at the protein surface.

Lastly, proteins that possess primary through tertiary structure can be further assembled, in what is called quaternary structure. Quaternary structure arises from noncovalent interactions between distinct polypeptide chains. In quaternary structure, each polypeptide is called a subunit. Homodimers, like HIV-protease, involve the association of two subunits with the same primary structure. Heterotrimers, as in G protein complexes, involve the association of three distinct subunits. These are two basic examples of numerous possible quaternary structures.

Many proteins perform their roles dynamically. For example, HIV-protease undergoes large domain motions whereby it opens and closes to gate access

of the substrate to the active site where proteolysis occurs [Ding et al., 2008]. Molecular motors change their shape, as dynein and kinesin move along microtubules or myosin moves along microfilaments. Many transcription factors have motions where two domains "open" to allow DNA binding and then close on their target sequence to form a stable complex. Some of these events are very rare from a microscopic perspective, occurring one thousand times per second or sometimes less.

To understand their dynamic nature, it is desirable to obtain a microscopic description of the behavior of proteins. For an understanding of function, this description should be accurate but highly resolved and in atomic detail. One way to obtain such information is through computational methods. These methods, however, must be both efficient and conformationally rigorous. But before moving on to how the microscopic properties can be obtained, we consider that one must be able to compare microscopic insights to the macroscopic properties observed experimentally, to ensure that the microscopic computational methods describe the same phenomena observed experimentally. We thus turn to a very superficial survey of some common experimental measurements before describing statistical mechanics that can bridge macroscopic and microscopic observations.

1.2 Experimental measurements of proteins

Protein structure determination was kickstarted with solution of the very first structure using X-ray crystallography, of myoglobin at 6 Å resolution by Kendrew et al. [1958]. Max Perutz then obtained a 5.5 Å resolution crystal structure of hemoglobin [Perutz et al., 1960]. Thus Kendrew and Perutz shared the 1962 Nobel Prize for chemistry for their pioneering work in structure determination. To date, the protein data bank lists 86 744 protein structures determined using X-ray crystallography [PDB].

But crystallizing a protein is laborious and the crystal environment may not exactly reproduce the effects on the protein provided by a solution environment. Another technique has been used to experimentally characterize proteins in solution: nuclear magnetic resonance (NMR). NMR can provide

information about the interactions of nuclei with non-zero spin. If a nucleus has an even number of protons and an even number of neutrons, its spin, and thus its magnetic moment, is zero. If, however, there is an odd number of both protons and neutrons, the nucleus has integer spin. Otherwise, the nucleus has half-integer spin. Whereas any nonzero spin nucleus will resonate in response to a magnetic field, spin-$\frac{1}{2}$ nuclei—including ^{1}H, ^{13}C, ^{19}F, and ^{31}P, but particularly ^{1}H—are most commonly studied. Spin-$\frac{1}{2}$ nuclei can assume two different spin states that, in the absence of a magnetic field, are degenerate, or energetically equal, and thus neither is preferred.

In a magnetic field, the two nuclear spin states separate in energy, assuming a Boltzmann distribution where the preferred state is called α and the other state β. Although the difference in the number of nuclei with each spin is small, it causes net magnetization of the system. With electromagnetic radiation of the appropriate frequency, resonant absorption will occur, whereby a nucleus with α spin can transition to the less favorable β spin. It is possible to add energy to the system such that α and β are evenly distributed, at which point the system will absorb no more radiation, a phenomenon called saturation.

But not all nuclei of the same element and spin number will require the same amount of energy for resonant absorption. Electrons also have spin, and due to their magnetic properties can shield nuclei from an external magnetic field. A reduced effective magnetic field at a spin-$\frac{1}{2}$ nucleus decreases the energy separation of its two spin states; thus a lower frequency of radiation is needed for resonant absorption. This difference of the apparent magnetic field felt by a nucleus from that applied externally is referred to as a chemical shift.

Nuclei may also shield each other. The strength of the effect of one magnetic nucleus on the resonant frequencies of another can be described by a scalar coupling constant. Scalar couplings, especially those that occur across three bonds, are particularly relevant for local protein structure. Conformational changes concerning secondary or tertiary structure commence by dihedral motion that can be described by atoms connected by three bonds, such as atoms connected to the N and Cα defining the ϕ dihedral torsion.

The dependence of three-bond scalar couplings on dihedral torsions can be estimated using Karplus relations [Karplus, 1959, 1963]. Karplus relations

establish ^3J scalar couplings as a cosine series of a single dihedral ϕ with parameters A, B, and C, as in Equation 1.1. Karplus parameter sets have been derived from experiment by comparing scalar coupling values from NMR spectroscopy to dihedral angles in X-ray structures that include the nuclei they describe, as has been done for the H$-$H$\boldsymbol{\alpha}$ scalar coupling [Hu and Bax, 1997]. Karplus parameter sets have also been derived from theoretical chemistry [Case et al., 2000], by computing the magnetic scalar coupling interactions. Three H$-$H$\boldsymbol{\alpha}$ Karplus curves are depicted in figure 1.6.

$$^3J = A\cos(\phi)^2 + B\cos(\phi) + C \tag{1.1}$$

Scalar couplings, however, arise from interactions within a complex arrangement of magnetic nuclei, as well as electrons. Therefore, a single dihedral may not sufficiently capture all phenomena that result in an experimental measurement. One theoretical study has proposed a grid-based method of calculating protein backbone scalar couplings as a function of both $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ [Salvador et al., 2011]. However, the calculated scalar couplings depend on the basis set used in the calculations, as well as how a set of pre-calculated scalar couplings is interpolated to the scalar coupling for a particular conformation. Additionally, the calculations Salvador et al. performed were for Ala3, and thus do not account for other spin systems that may be present in real proteins.

Spin relaxation is of great interest in NMR, as well. The net magnetization of a sample aligned in a magnetic field can be rotated by applying a pulse of circularly polarized electromagnetic radiation. If one visualizes the direction of the magnetic field as along the z-axis, and the direction of the electromagnetic pulse as along the y-axis, then after a pulse of electromagnetic radiation that rotates the system by 90°, the net magnetization of the system can be visualized as along the x-axis. In such an alignment, the system has an equal number of α and β spins. The magnetization will then recover to its equilibrium state, a process called spin relaxation. Spin can relax in two ways: longitudinally, along the z-axis, in spin-lattice relaxation, where spins transition from β back to α while exchanging energy with the surroundings, or lattice; or transversely, in

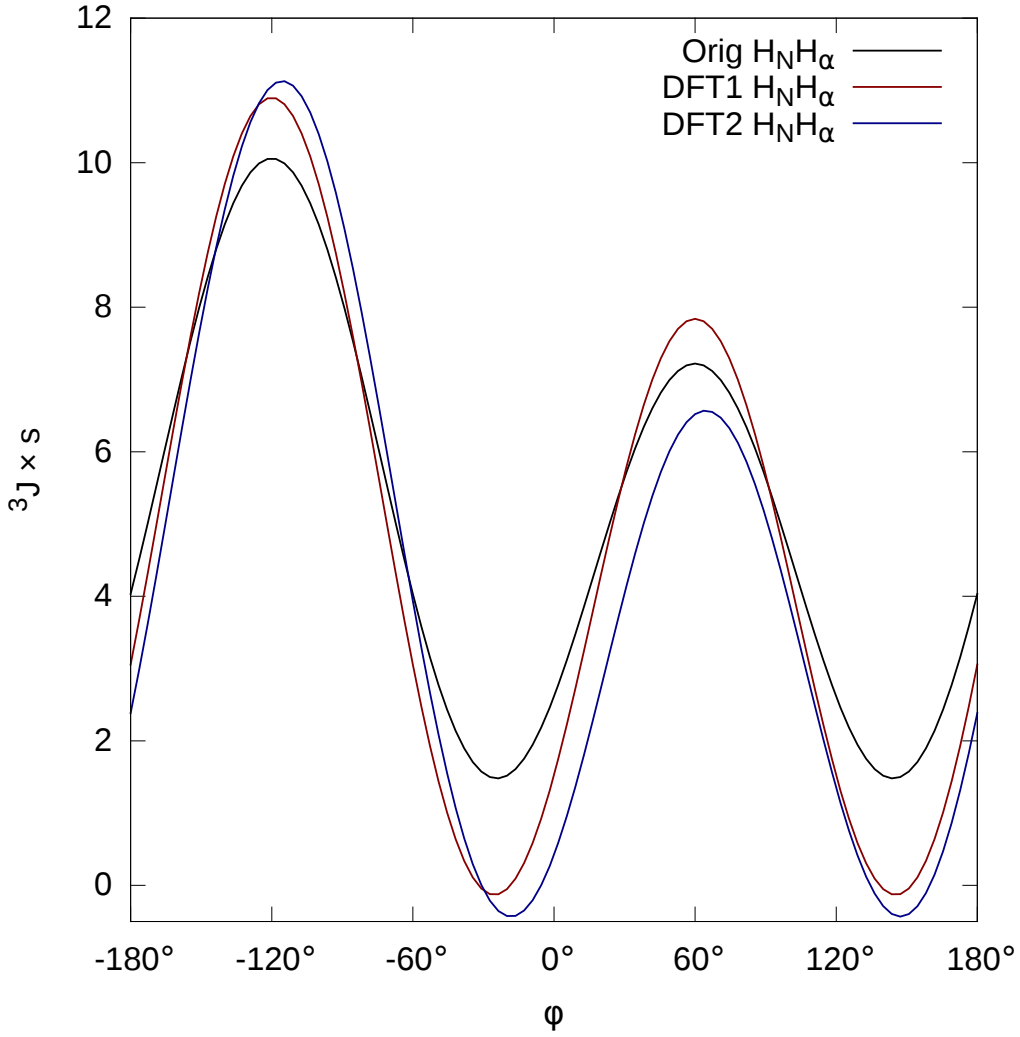

Figure 1.6: Orig [Hu and Bax, 1997], DFT1, and DFT2 [Case et al., 2000] H−Hα Karplus curves. Names from Best et al. [2008].

the xy-plane, in spin-spin relaxation, where the orientations of spins randomize relative to other spins. Spin relaxation depends on how molecular motions orient spins relative to local and global magnetic fields.

Lipari and Szabo worked out a "model-free approach" for interpreting NMR (spin) relaxation spectra [Lipari and Szabo, 1982]. Their approach produces "generalized order parameters" ($\mathcal{S}^2$) that provide an intuitive measure of how much the motion of a bond vector is restricted. NH order parameters are thus used to probe the flexibility of the protein backbone, whereas CH order parameters can be used to measure the flexibility of protein side chains.

Nuclear Overhauser Effects (NOEs) can be used to determine the distances between protons. Spin can be transferred via cross-relaxation when protons are near each other. This effect depends upon the dipolar field generated by each proton, as dipolar interaction facilitates spin-lattice relaxation. This dipolar field dies off through space inversely with the distance cubed, i.e. as r^{-3}, where r represents the distance from a proton. As the effect on relaxation is proportional to the square of the dipolar field, NOE intensity can be related to r^{-6}. It is not unusual for hundreds of interproton NOEs to be recorded during NMR structure characterization. NOEs have become important in establishing tertiary contacts in a solution structure as they are not confined to nuclei connected by chemical bonds.

Besides structural measurements, thermodynamic measurements can also be applied to proteins. For example, free energy of interaction between molecules, e.g. the binding between a protein and a drug, can be measured using calorimetry methods like isothermal titration calorimetry. All the experimental methods described are quite effective at obtaining macroscopic properties of molecular ensembles.

1.3 Statistical mechanics

When studying chemical systems such as proteins one is often interested in the relationship between microscopic behavior and the *macroscopic* properties of an ensemble including many (on the order of Avogadro's number, $N_A = 6.022 \times 10^{23}$) particles, which is the regime traditionally measured by

experiments (though single-molecule experiments constitute a growing field). Such macroscopic quantities include thermodynamic variables like pressure, temperature, volume, number of particles, chemical potential, energy, and entropy, as well as structural measurements like the average distance between two protons. In contrast to these macroscopic properties, simulations produce the microscopic interactions and motions of molecules. *With enough computational power*, one could hypothetically simulate the $\sim N_A$ particles needed to match the macroscopic scale. Such calculations are not only prohibitively expensive, however, they are also unnecessary.

Statistical mechanics connects the microscopic to the macroscopic. Through statistical mechanics, information about molecular behavior obtained from simulations of a single molecule could be used to predict thermodynamic observables like energy. The connection between the microscopic and macroscopic worlds is achieved through a *partition function*. The nature of this partition function depends on the boundary conditions of the microscopic system, which are particular to different *ensembles*. The most common ensembles are the canonical, microcanonical, and isobaric-isothermal:

- The canonical ensemble fixes the number of particles (N), the volume (V), and the temperature (T); thus, this ensemble is also called the NVT ensemble.
- The microcanonical ensemble also fixes N and V, but instead of temperature, fixes the energy (E); this ensemble is therefore also called the NVE ensemble.
- The isobaric-isothermal ensemble differs from the canonical in that pressure (p) is held constant rather than volume; this ensemble is hence called the NpT ensemble.

The canonical partition function is Equation (1.2). The partition function is a sum over the Boltzmann factors, $e^{-\beta E_j}$, of all t states of a system, where j is an identifier for each state. The result of this sum, Q, can be used as a normalizing factor when calculating macroscopic observables. The probability of a state j, p_j, can be calculated by dividing its Boltzmann factor by Q

(Equation (1.3)). Additionally, one can calculate the average energy by summing over the Boltzmann factor-weighted energies of each state of the system, and dividing by Q (Equation (1.4)).

$$Q = \sum_{j=1}^{t} e^{-E_j/kT} \tag{1.2}$$

$$p_j = \frac{e^{-E_j/kT}}{Q} \tag{1.3}$$

$$\langle E \rangle = \frac{\sum_{j=1}^{t} e^{-E_j/kT} E_j}{Q} \tag{1.4}$$

Likewise, any average property, $\langle P \rangle$, can be calculated analogously to Equation (1.4), as

$$\langle P \rangle = \frac{\sum_{j=1}^{t} e^{-E_j/kT} P_j}{Q} \tag{1.5}$$

One can also use Equation (1.4) to calculate the relative energies between two states i and j, $\Delta E_{ij} = E_i - E_j$.

$$\frac{p_j}{p_i} = \frac{e^{-E_j/kT}/Q}{e^{-E_i/kT}/Q} \tag{1.6}$$

$$= e^{E_i/kT - E_j/kT} \tag{1.7}$$

$$= e^{\Delta E/kT} \tag{1.8}$$

$$\Delta E = kT \ln \frac{p_j}{p_i} \tag{1.9}$$

Upon first inspection, one aspect of the partition function is daunting. To calculate Q, one needs information about every microscopic state of the system. Practically, however, this requirement is not rigid; the vast majority of states contribute negligibly to the partition function, with only the most populated states contributing significantly. Thus, with knowledge of the major states, one can reasonably approximate Q.

Earlier, the notion that observation of enough states of a single molecule will produce the same distribution as that of many molecules was raised. This

is called the Ergodic hypothesis. Given infinite time, the fraction of each state sampled is expected to be equivalent to the fraction observed over an ensemble of infinite molecules. Ergodicity is central to approaches that use simulations of single molecules to calculate ensemble properties using statistical mechanics.

Still needed is an accurate but computationally feasible route to obtaining the microscopic behavior of a protein system. Two methods that are not effective—one that cannot describe molecular behavior and one that is too slow for large, dynamic systems like proteins—will be introduced, followed by a third method, borrowing from the first two, that will be the subject of this dissertation.

1.4 Classical mechanics describes the motions of bodies

With the 1687 volume *Philosophiæ Naturalis Principia Mathematica* (or *Principia*), Sir Isaac Newton revolutionized physics. For the first time, the interactions and motions of projectile and celestial bodies, their *mechanics*, could be predicted using a single set of equations. Newton prescribed three laws:

1. Objects maintain their velocity unless an external force causes acceleration.
2. The forces acting on an object equal the object's mass times acceleration.
3. The action of a body on another always accompanies an equal but opposite reaction of the second body on the first.

In other words:

1. In the absence of an external force, $v(t) = v$, where $v(t)$ and v represent velocity as a function of time or as a constant, respectively.
2. $\vec{F} = m\vec{a}$, where $\vec{F}$ represents force, m is mass, and $\vec{a}$ is acceleration.
3. $\vec{F}_{ij} = -\vec{F}_{ji}$, where $\vec{F}_{ij}$ is the force of body j acting on body i, and $\vec{F}_{ji}$ is the converse.

Moreover, Newton had laid out calculus—a system for interrelating variables, including physical properties. From calculus, one can relate energy, acceleration, velocity, and position. Force, $\vec{F}$, is the derivative of potential energy, $U(\vec{x})$, with respect to the coordinates, $\vec{x}$ (Equation (1.10)). Acceleration—obtained from force by Newton's second law—is the derivative of velocity, $\vec{v}$, with respect to time, t (Equation (1.11)). In turn, velocity, $\vec{v}$, is the derivative of position, $\vec{x}$, with respect to time, t (Equation (1.12)). Thus, if one knows the form of the differentiable function $U(\vec{x})$ and the masses $\vec{m}$, one can use equations 1.10 to 1.12 to go from coordinates to forces (and accelerations), to velocities, and back to coordinates.

$$\vec{F} = -\frac{\partial U(\vec{x})}{\partial \vec{x}} \tag{1.10}$$

$$\vec{a} = \frac{\partial \vec{v}}{\partial t} \tag{1.11}$$

$$\vec{v} = \frac{\partial \vec{x}}{\partial t} \tag{1.12}$$

It is often desired to apply these equations circularly to predict the time-trajectory of coordinates $\vec{x}(t)$. Principally, the goal is to obtain $\vec{x}(t)$ at some time t, τ in the future from a reference time t_0. One can accomplish this using Equation (1.13) to determine position $\vec{x}_i(t)$ of each atom i from previous position and velocity $\vec{x}_i(t_0)$ and $\vec{v}_i(t_0)$, respectively. A key assumption in this equation is that the force (Equation (1.10)) is continuous over the time interval $[t_0 : t)$ or, more generally, that higher order time derivatives are negligible; thus τ must be sufficiently small for constant force to be a reasonable approximation.

$$\vec{x}_i(t = t_0 + \tau) = \vec{x}_i(t_0) + \vec{v}_i(t_0)\tau - \frac{1}{2m_i}\frac{\partial U(\vec{x}(t_0))}{\partial \vec{x}_i(t_0)}\tau^2 \tag{1.13}$$

Equation (1.13) is equivalent to the Euler method or a second order Taylor series about $\vec{x}(t_0)$. Classically, gravitational potential has been used for $U(\vec{x})$ to predict the motions of arrows or planets. Within this regime, Newton's classical

mechanics safely assumes that energy and matter can be treated continuously.

1.5 The failure of classical mechanics

Despite its predictive power for celestial bodies and projectiles, shortcomings were discovered in the 19th century that render classical mechanics unsuitable for the atomic-scale.

For example, the Rayleigh-Jeans formula for predicting black-body radiation resulted in the *ultraviolet catastrophe*—the prediction that black-bodies can radiate with infinite power at high frequencies. This is clearly unphysical and violates the law of conservation of energy. Max Planck observed that this prediction resulted from the assumption that energy is continuous, whereby any quantity of energy can be absorbed or emitted.

Other gaps in classical mechanics were discovered around the turn of the 20th century. For example, the photoelectric effect describes the emission of electrons by metals in response to light. Classically, the transfer of energy from light to electrons in the metal could occur gradually, and therefore changes to the amplitude or wavelength of the incident light could modulate electron emission. It was assumed that light of sufficient intensity was required for the photoelectric effect. But it was observed that low intensity light could evoke the photoelectric effect; meanwhile, high intensity light was insufficient for the photoelectric effect if the light was above a certain wavelength. Something was wrong.

1.6 Quantum mechanics

Addressing these issues, Planck postulated that energy can only be exchanged in integer multiples of a quantity $h\nu$, where h is Planck's constant and ν is the radiation frequency. Though it has been debated whether Planck unwittingly invented quantum mechanics [Galison, 1981], Planck's law, $\Delta E = nh\nu$, meant that the highest frequencies of radiation could not be emitted without a sufficient quantity, or *quantum*, of energy. Black-body radiation predictions

incorporating this consideration agreed with experimental emission spectra. In 1905, Einstein hypothesized that light was composed of particles called photons, and that these photons were Planck's quanta, with energy proportional to frequency. Einstein went a step further to explain that the photoelectric effect required photons of certain energy or frequency, independent of intensity. Thus began a revolution in the understanding of physics on the scale of atomic particles.

Heisenberg discovered that, for particles, position and momentum cannot be simultaneously known—barring the Newtonian equations of motion. Instead, matter can be described by wavefunctions that indicate the probability of a particle occupying a given region in space. A Hamiltonian operating on the wavefunction can yield the energy of the system, as detailed by the Schrödinger equation (1.14), where $\imath$ is the imaginary unit, $\hbar$ is the reduced Planck constant, $\frac{h}{2\pi}$, t is time, Ψ is the wavefunction, and $\mathscr{H}$ is the Hamiltonian.

$$\imath\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) = \mathscr{H}\Psi(\vec{r},t) \tag{1.14}$$

For a non-relativistic particle, the Hamiltonian may be defined as:

$$\mathscr{H} = -\frac{\hbar^2}{2\mu}\nabla^2 + U(\vec{r},t) \tag{1.15}$$

where μ is the reduced mass, ∇^2 is the Laplacian, and $U(\vec{r},t)$ is the potential at coordinates $\vec{r}$ and time t. The first term in this equation, $-\frac{\hbar^2}{2\mu}\nabla^2$, is the kinetic operator—when operating on the wavefunction, it yields kinetic energy. It is analogous to the classical definition of kinetic energy, $\frac{1}{2}mv^2$, when one considers that the momentum $p = mv$, hence the kinetic energy can be expressed, $\frac{p^2}{2m}$. Substituting μ for m and $\imath\hbar\nabla$ for p, one obtains the quantum mechanical version in Equation (1.15).

If the potential is assumed to be independent of time, i.e., $U(\vec{r},t) = U(\vec{r})$, then one may use the time-independent Schrödinger equation (1.16) to predict stationary states.

$$\mathscr{H}\psi(\vec{r}) = E\psi(\vec{r}) \tag{1.16}$$

Whereas the time-dependent Schrödinger equation requires an initial wavefunction to project the wavefunction some time in the future, the time-independent Schrödinger equation allows the solution of stationary states that are of interest for physicists and chemists. Unfortunately, the Schrödinger equation is not able to be directly applied to systems with multiple electrons. Thus approximations are necessary.

1.7 Theoretical chemistry

> "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."
>
> – Paul Dirac

To discuss the approximations typically made in the field of theoretical chemistry, we first consider the Hamiltonian for N electrons and M nuclei:

$$\mathscr{H} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{r_{AB}} \tag{1.17}$$

Above, M_A is the ratio between the mass of nucleus A and the mass of an electron, Z_A is the charge of nucleus A in units of the electron charge (and of nucleus B, *mutatis mutandis*), r_{iA} is the distance between electron i and nucleus A, r_{ij} is the distance between electrons i and j, and r_{AB} is the distance between nuclei A and B. From left to right, the terms include the electronic kinetic operator, the nuclear kinetic operator, the Coulomb attraction between nuclei and electrons, the interelectronic repulsion, and the internuclear repulsion.

The most fundamental approximation to quantum mechanics is the Born-Oppenheimer Approximation (BOA). The mass of the electron is dwarfed by the mass of a proton-containing nucleus. Therefore, with commensurate momenta, an electron will move much more quickly than a proton. It can then be assumed that electrons will adapt their configuration to an arrangement of nuclei before the nuclei can move appreciably. This separates nuclear and electronic degrees of freedom.

In terms of Equation (1.17), the BOA implies that the second term can be neglected, as the nuclei can be assumed to have no kinetic energy, and that the final term is constant, as the internuclear distances are invariant. Being constant, the final term will not contribute to the wavefunction, but can be applied to the wavefunction obtained using the electronic Hamiltonian:

$$\mathscr{H}_{\text{elec}} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{1.18}$$

An additional number of approximations together with the BOA comprise Hartree-Fock (HF)—the foundation of nearly all theoretical chemistry. HF is unique in that it allows an approximate solution of the time-independent Schrödinger equation from first principles—ab initio. HF assumes:

- BOA
- Relativistic effects can be neglected
- The wave function is a linear combination of basis functions
- A finite number of basis functions can approximate a complete set
- A single Slater determinant will describe each energy eigenfunction
- Electrons only interact with each other in an average way in the mean field approximation

The final two assumptions are perhaps the weakest, though they enable the unsurpassed power of HF to even approximately solve the Schrödinger equation. To compensate for inaccuracies due to the HF approximations, various

post-HF methods have been developed employing different corrections to the HF energy for the last two approximations.

A description of electron correlation accounting for deviations from the mean field approximation is of particular interest; several methods with a range of computational expense have been introduced. Correlation arises because electrons have specific interactions that are not perfectly captured by a mean-field approximation. To accurately model long-range chemical effects like London dispersion interactions, correlation must be included.

One of the most widely used classes of post-HF methods, Møller-Plesset perturbation theory (MP), employs Rayleigh-Schrödinger perturbation theory to account for electron correlation. This perturbation, when truncated after the second order, is referred to as MP2. Commonly, MP2 is applied to the study of small- to medium-sized biomolecules, up to hundreds of atoms.

There are other methods that are much more accurate than MP2, including coupled-cluster theory and, the golden standard, full configuration interaction (CI). CI can recover the true BOA energy surface, compensating for all non-BOA approximations in HF. Yet, CI is extremely expensive, and is thus typically applied to small molecules with on the order of 10 atoms.

Although theoretical chemistry based on HF can garner many insights, it is still generally too slow for problems of biological interest like evaluating millions of conformations of molecular systems with thousands of atoms. Even evaluating several conformations of a solvated protein would be challenging. Thus there is special interest in methods that may compromise some accuracy or generality to attain greater speed.

There are semi-empirical methods with functional forms that can reproduce HF-like behavior with appropriate parameterization. The necessity of parameters means that these methods are no longer ab initio and, commonly, different semi-empirical methods will be better suited for different problems. Unfortunately, despite being less accurate than ab initio methods, semi-empirical methods are still generally too slow for studying solvated biomolecules like proteins.

1.8 Molecular mechanics

Instead of making further approximations to quantum mechanics, one can forgo the Schrödinger equation and wavefunctions altogether. The BOA assumes that the motions of the nuclei are slow enough that their positions can be frozen while the time-independent Schrödinger equation is solved. Conversely, one can assume that as nuclei move, they experience averaged effects from the much faster moving electrons. Thus one can approximate the ground state of the Born-Oppenheimer surface by modeling averaged electronic effects as a function of nuclear coordinates. This approximation fuels molecular mechanics (MM) models that treat a system in terms of prescribed nuclear interactions, a key component in multi-scale molecular simulations for which the 2013 Nobel Prize in Chemistry was awarded.

MM methods employ force fields, sets of empirical formulae that collectively describe the potential energy of a class of molecules. Thus MM describes a system's chemistry not by a wavefunction that must be self-consistently solved for each nuclear arrangement, but by simple algebraic functions of nuclear coordinates like bond lengths or angles that can leverage the prior derivation of physically meaningful parameters. Unlike the approximations of theoretical chemistry, which are deductive, MM is an inductive approximation. MM methods have become popular for calculation of molecular properties because of their computational efficiency over quantum mechanics—they can be applied to one trillion evaluations of ubiquitin in explicit solvent [Piana et al., 2013].

MM uses variations of Equation (1.13), typically Velocity Verlet [Swope et al., 1982] or Leapfrog integration, to propagate the dynamics of a system. The timestep typically employed is on the order of 1 fs, as required to stably model the vibration of bonds including hydrogen. This requirement arises from the assumption that higher-order derivatives of potential energy with respect to time are negligible, as first described in Section 1.4. $U(\vec{x})$ is provided by a sum of simple functions of molecular geometries called a force field.

1.9 Force fields

In 1968, Shneior Lifson and Arieh Warshel developed the *consistent force field* (CFF) [Lifson and Warshel, 1968]. Prior to the CFF, functions were derived independently for specific problems one was interested in, such as bond vibrations or van der Waals interactions. Lifson and Warshel, however, developed a framework whereby a single function could have parameters that were solved together—consistently—to reproduce enthalpies, vibrations, and geometries of alkanes. Using this framework, a number of functional forms were evaluated and the most useful incorporated into the CFF, upon which most modern force fields [Jorgensen and Tirado-Rives, 1988, Cornell et al., 1995, Wang et al., 2000, Hornak et al., 2006, MacKerell et al., 1998, 2004a, Best et al., 2012] are based. One essential set of terms preserves the most basic geometric results of chemical connectivity—bond lengths and angles.

The bond lengths and angles are approximated as harmonic oscillators, with each unique class, or *type*, having a characteristic equilibrium length r_0 or equilibrium angle θ_0 with harmonic vibrations according to bond and angle force constants k_r and k_θ, respectively (Equation 1.19).

$$V_{\text{bonded}} = \sum_{r}^{\text{bonds}} k_r(r - r_0)^2 + \sum_{\theta}^{\text{angles}} k_\theta(\theta - \theta_0)^2 \tag{1.19}$$

Equation 1.19 has strengths and weaknesses. The separation of each bond and angle makes a force field more intuitive and the harmonic approximation is very computationally efficient. On the other hand, as bond lengths and angles change, so do the densities of electrons and, therefore, there should be some coupling between different bonded interactions. Additionally, approximations beyond harmonicity may aid prediction of strained geometries, where overtones from higher order vibrational levels may become significant.

Addressing the first issue of coupling, the extensively parameterized organic force fields of Norman Allinger [Allinger, 1977, Allinger et al., 1989] include coupling between bonded terms. A number of options have also been employed to better model bond anharmonicity. A Morse potential (Equation 1.20), for

example, can allow the system to adapt to longer bond lengths more naturally, while predicting greater strain at close range than a harmonic approximation. More simply, adding third or higher-order terms, rather than only the second order harmonic contribution, can improve the accuracy of bond and angle energy functions. This has been employed, for example, by Norman Allinger [Allinger, 1977, Allinger et al., 1989] or by Ren and Ponder in AMOEBA [Ren and Ponder, 2003]. The CHARMM line of force fields employs not only 1–3 interactions parameterized by the covalent angle between the two atoms, but Urey-Bradley potentials (Equation 1.21) parameterized by the 1–3 distance. Other force fields [Weiner et al., 1984, 1986, Cornell et al., 1995, Wang et al., 2000, Duan et al., 2003, Hornak et al., 2006, Best and Hummer, 2009], however, have maintained the simple, additive form of Equation 1.19 for computational efficiency and have been successfully applied to dynamics of polypeptides, nucleic acids, carbohydrates, and lipids.

$$V_{\text{Morse}} = D_e(1 - e^{-a(r-r_0)})^2 \tag{1.20}$$

$$V_{\text{Urey-Bradley}} = \sum_u V_u(u - u_0)^2 \tag{1.21}$$

Besides the bonded interactions that restrain chemical connectivity, atoms can also have non-bonded interactions. Non-bonded interactions are often not considered between atoms in a bond or angle; instead, 1–4 interactions and beyond are evaluated. In so-called Class I force fields, as in the CFF, these are summarized as Coulombic potentials and van der Waals (in this case, Lennard-Jones [Jones, 1924]) potentials (Equation 1.22). The latter principally model repulsion due to the Pauli exclusion principle and long-range attraction due to London dispersion.

$$V_{\text{non-bonded}} = \sum_{i,j}^{\text{non-bonds}} \frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{1.22}$$

Whereas Coulomb's law is used also in quantum mechanics, the derivation of nuclear point charges q_i and q_j is not straightforward as electron

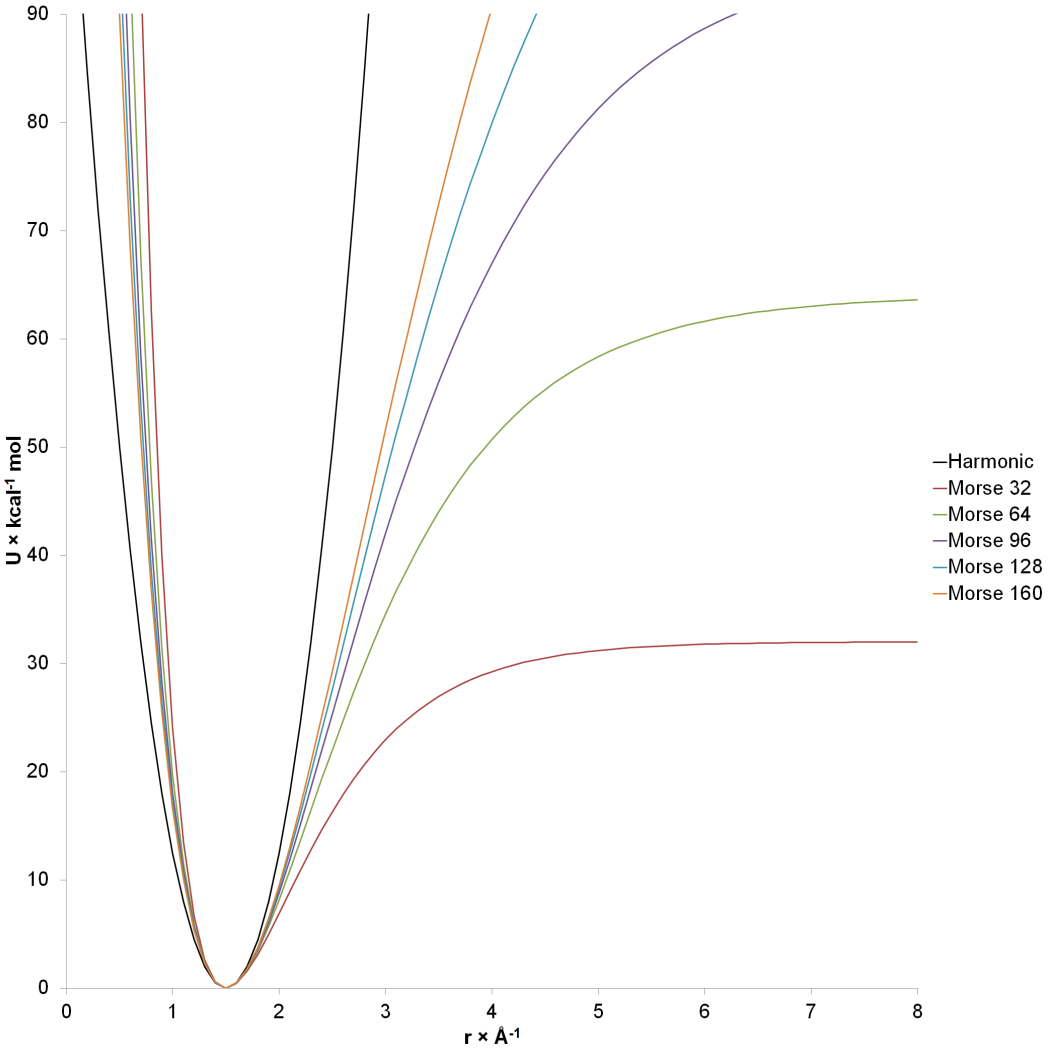

Figure 1.7: The Morse potential describes the anharmonicity characteristic of real bonds, as illustrated for this toy system with 1.5 Å equilibrium bond length, 100 kcal/mol/Å^2 force constant at the minimum, and harmonic energy (black) or Morse well depth of 32 (red), 64 (green), 96 (purple), 128 (blue), or 160 (orange) kcal/mol.

density is not localized to the nucleus but comprises dynamic molecular orbitals. Some fixed-charge force fields take the simple approach of calculating the electrostatic potential at several points outside a molecular surface, and fitting atomic point charges that best reproduce the electrostatic potentials. Contemporary AMBER charge models [Cornell et al., 1995, Duan et al., 2003, Cerutti et al., 2013] have used restrained electrostatic potential (RESP) fits to ensure that underdetermined charges within molecules do not become arbitrarily large [Bayly et al., 1993]. Meanwhile, other force fields, such as CHARMM [MacKerell et al., 1998], have used an iterative refinement of nonbonded and bonded parameters together, to reproduce interaction energies, geometries, dipole moments, heats of vaporization, molecular volumes, and heats of solvation.

Some more detailed models attempt to account for off-nuclear electrostatic components like dipoles and quadrupoles, as in AMOEBA [Ren and Ponder, 2003]. Also like AMOEBA, a number of charge models also allow polarization—the redistribution of electron density in response to an electric field. In theory, these types of models promise fundamental improvements over simple fixed nuclear point-charge models. Due to great investments in the latter and the few degrees of freedom that need to be parameterized, however, fixed-charge models can often be nearly as accurate as polarizable models, with less computational cost [Fried et al., 2013].

The Lennard-Jones van der Waals contributions in Equation 1.22 consist of dispersion, modeled by r^{-6}, and repulsion, modeled by r^{-12}. Whereas r^{-6} has physical basis, r^{-12} is used because it is the square of r^{-6}, and thus is a very inexpensive treatment of repulsion requiring only multiplication of r^{-6} (already calculated for dispersion) times itself, times the Lennard-Jones parameter A. The approximation may overestimate internuclear repulsion at close distances, however. Realistically, repulsion may be more accurately modeled by an exponential form, as in the Buckingham (Equation (1.23)) or Morse (Equation (1.20)) potentials. Some models, like AMOEBA, have used "buffered" van der Waals functions, instead [Ren and Ponder, 2003]. Yet for cases where molecules are not very strained or experience little repulsion, the 12-6 form may be reasonable [Weiner et al., 1984].

$$V_{\text{Buckingham}} = Ae^{-Br} - \frac{C}{r^6} \tag{1.23}$$

Whereas bond lengths and angles are relatively rigid, many conformational changes of interest consist of the dihedral torsion of groups on opposite ends of single bonds, such as rotation of ϕ and ψ dihedral torsions in the protein backbone. Dihedral torsional motion depends on 1–4 interactions between A and D in consecutively bonded atoms A−B−C−D. As one of the limitations of the Lennard-Jones form is exaggerated repulsion at close range, AMBER force fields scale 1–4 van der Waals interactions by $\frac{1}{2}$. Electrostatic interactions between 1–4 atoms are scaled by $\frac{1}{1.2}$ in AMBER, as this facilitated agreement with ab initio calculations [Cornell et al., 1993]. CHARMM22 [MacKerell et al., 1998, 2004a] and GLYCAM [Kirschner et al., 2008], however, have not required 1–4 scaling factors. Meanwhile, because of the importance of torsion, dihedral corrections coordinate with the 1–4 non-bonded interactions to adequately model this soft degree of freedom. AMBER and CHARMM force fields include Fourier series dihedral corrections as presented in Equation 1.24.

$$V_{\text{dihedral}} = \sum_{\phi}^{\text{dihedrals}} \sum_{n}^{\text{periodicities}} V_{\phi n}(1 + \cos(n\phi - \gamma_{\phi n})) \tag{1.24}$$

This form descends from a study of alkanes showing that cosines map to torsional vibrations better than harmonic functions [Pitzer, 1951]. Pitzer hypothesized that the relevant effects that result in cosine-shaped torsional energy contributions included quadrupole/quadrupole and van der Waals interactions. In force fields that include van der Waals interactions explicitly, a cosine correction can still serve to mask errors in 1–4 non-bonded interactions, besides accounting for all missing multipole, inductive, and bonding and antibonding orbital effects. Importantly, dihedral corrections are the only term in class I force fields that can maintain planarity across double bonds.

As all other functions presented, the sinusoidal dihedral correction has advantages and disadvantages. Its elegance is that by simply adding more cosines to the series, one can map nearly any correction for a single dihedral.

Additionally, its locality to a single torsion allows some physical rationale based on the effects of quadrupole-quadrupole or van der Waals interactions. In practice, the physical basis is limited by the challenge of partitioning errors relative to quantum mechanics or experiments, that may be of any nature, into corrections for single dihedral torsions. The greatest limitation is that the cosine terms in Equation 1.24 must use individual dihedral angles to produce corrections that apply at all relevant combinations of remaining dihedrals. This problem is nontrivial as there may be coupling between dihedrals that is not fully captured by other, e.g. non-bonded, force field terms.

In recent CHARMM force fields [MacKerell et al., 2004a, Best et al., 2012], an additional correction that is simultaneously a function of two dihedrals is added. This correction, called a coupled correction map (CMAP) [MacKerell et al., 2004b], consists of a two-dimensional grid, where each gridpoint represents an energy correction for that combination of dihedral values. From this grid, bicubic interpolation can map a correction for any point on the two-dimensional surface. Thus the CMAP allows the near-quantitative reproduction of any two-dimensional surface. Their one disadvantage is lack of clear physical motivation (although physical motivation isn't necessarily characteristic of cosine-based corrections in practice, either). There is a risk that CMAP corrections may therefore mask a deficiency in nonbonded parameters, for instance, that could either not transfer to other amino acids with different side chains or that could potentially create an imbalance between local and global interactions. Effective use of CMAP, however, enables much finer force field tuning than simple cosine-based corrections.

In addition to the limitations of the force field functionals just discussed, there are fundamental weaknesses in molecular mechanics. For example, molecular mechanics is not quantized, but allows arbitrary energies. Additionally, molecular mechanics lacks entanglement, tunneling, and other potentially relevant quantum effects. Furthermore, real chemistry is dynamic and coupled in a way that is explicitly absent from class I fixed-charge force fields, and that polarizable models and cross-terms may not sufficiently capture. A simple illustration is that charge polarization seeks to capture changes in electron density, but these same changes should simultaneously affect van der Waals

and all other interactions. The partitioning of chemistry is an arduous approximation.

But the approximations are not unredeemed. Balancing MM's limitations are its speed and the many force field parameters that can be added to more precisely describe the amino acid energy landscape. Molecular mechanics on the millisecond timescale has revealed a potential folding path for 76-residue ubiquitin, a feat that would not be possible with even semi-empirical quantum methods. Meanwhile, "effective" force field training (if such a term can be defined, but see below) can compress a great deal of information into the parameters describing individual molecular interactions, with a trend in recent years toward more parameters.

1.10 Force field development principles

Due to the inductive, empirical nature of molecular mechanics, opinions differ on exactly how to perform "effective" force field training. But there are sets of principles that recur in force field development. A classic theme, originating with the early liquid simulations of OPLS [Jorgensen and Tirado-Rives, 1988], is that there should be a balance between intramolecular and intermolecular interactions. Especially, protein-solvent interactions should be of an appropriate magnitude relative to protein-protein and solvent-solvent interactions. This has led AMBER force fields to develop amino and nucleic acid charges that polarize the solute to a level comparable to that of the commonly employed TIP3P [Jorgensen et al., 1983] water model [Bayly et al., 1993, Cornell et al., 1993]. Additionally, CHARMM optimization of non-bonded parameters explicitly included interactions with TIP3P water in the training [MacKerell et al., 1998].

Another critical issue to consider is how many parameters to use to describe model chemistry. The variety and constitution of force field parameters is controlled by the segregation of atoms into unique types. These atom types are often divided based on element and bonding partners. For example, there may be sets of atom types for hydrogen and carbon. Hydrogen types may be divided into those bound to nitrogen, oxygen, aliphatic carbon, or aromatic carbon.

Carbon types may be divided into *sp*3-hybridized carbon, *sp*2-hybridized carbonyl carbon, and aromatic carbon.

The complexity of atom types is largely a function of two competing ideals. First, one wants to ensure ample specificity in describing chemically unique atoms. As in ff94, for example, one may wish to describe carbonyl oxygen as atom type O, but carboxylic oxygen as atom type O2 [Cornell et al., 1995]. Second, one must ensure there are adequate data for the number of parameters being defined. Therefore, without physical motivation for distinction, one may leave chemically similar atoms with the same type—as ff94 uses the CT atom type for the β-carbon of every amino acid (except glycine, which has no β-carbon)—to maximize the ratio of data to parameters. Thus force field design must account for the competing goals of specificity—the number of parameters—and robustness—the amount of data used to derive each parameter.

This information requirement restricts how complicated a force field can be. Many force fields, for example, have only two or three sets of backbone dihedral corrections, with one or a combination of corrections being applied to each amino acid. AMBER, for example, has traditionally had one set of ϕ/ψ backbone dihedral corrections that applies to all amino acids, and is the only set that glycine has, with a second set of ϕ'/ψ' corrections that adds to the first set to correct all amino acids with a β-carbon—every amino acid except glycine. CHARMM, as of CHARMM22/CMAP [MacKerell et al., 2004a], has had one set of dihedral parameters for glycine, one set for proline, and then another alanine-based set for everything else.

Intuitively, force fields that include explicit coupling of different parameters would require an enormous amount of fitting data to ensure all coupled degrees of freedom are adequately sampled. What is less obvious is that force fields with uncoupled parameters still include coupling, albeit implicitly. For example, the correction needed for $\chi 1$ may depend on whether the backbone conformation is α, β, or something else. Thus if trained for only one case, say the case of a β backbone conformation, then the $\chi 1$ parameters would likely be appropriate in the context of β backbone conformations, but with no

guarantee of transferability to other contexts such as helical backbone conformations. Alternatively, this implicit coupling could be averaged over multiple backbone conformations. In that case, the $\chi 1$ correction may not be as ideal for β backbone conformations, but will likely be more reasonable across multiple contexts.

What information sources to use in training is a very important question, as well. Simulations must be able to explain experimental observations to aid our understanding of protein properties and behavior, for example. But because of ensemble averaging in many experiments, it can be difficult to obtain information that is uniquely identifiable to a single conformation. An NMR measurement like an NOE might reveal that two atoms are, on average, in close proximity. But exactly how close they are, the dynamic distribution of that proximity, and more importantly why the two atoms associate and what this means for the rest of the molecule, may not always be obvious from a single experimental observation. Thus a major goal of force field development is to produce simulation models that can be consistent with and support experiments.

Unfortunately, it is not generally the case that experimental observations simultaneously offer the microscopic detail and thermodynamic rigor needed to train a reliable force field. For example, a high-resolution crystal structure will tell much about a protein's average structure, but much less about that structure's dynamics and energetics that would be needed for a force field defining potential energy as a function of coordinates that can change. Calorimetry may reveal thermodynamic properties—at what temperature a phase transition occurs and how much heat is associated with the transition. Three-bond scalar couplings may reveal dynamical information about bond rotamer preferences. But these experiments typically only report on one property at a time. Calorimetry does not provide structure. A scalar coupling may only suggest conformations for a single ϕ or $\chi 1$ dihedral. Even with knowledge of the ϕ distribution of an amino acid, one still will generally not know whether this distribution arises because of energetics internal to the ϕ dihedral, or for some other cause. Some conformations, like those along a transition between two stable structures, may be nearly impossible to determine experimentally because

of the overwhelming signal from the stable structures. To obtain additional, highly specific information, most force field training efforts incorporate, along with experimental observables, information from quantum mechanics.

Quantum mechanics (QM), though too slow to simulate solvated proteins over biologically relevant timescales, can be used to calculate a reference potential energy for any set of nuclear coordinates of a computationally accessible number of atoms, as well as to calculate atomic forces, geometry, electrostatic potential, or any chemical property. QM is often utilized to calculate the potential energy surface for a set of conformations, which can be compared to the same energy surface according to MM. A QM reference potential energy surface on a two-dimensional grid spanning backbone dihedrals ϕ and ψ, for example, can allow the calibration of the MM potential energy surface by adjusting relevant backbone parameters [Cornell et al., 1995, Kollman et al., 1997, Wang et al., 2000, Hornak et al., 2006, MacKerell et al., 1998, 2004a, Best et al., 2012]. Or, returning to the example of side chain parameters appropriate for multiple backbone conformations, with QM one can obtain side chain potential energy scans multiple times with the backbone fixed to different conformations, such as α and β. Or, by calculating the interaction between two molecules at varying orientations, one can train force field parameters to describe the interaction energy, ideal interaction geometry, and the penalty for deviation from that ideal geometry [MacKerell et al., 1998]. All would be important for dynamical simulations where two interacting molecules can translate or rotate relative to each other, constrained only by the force field describing their interactions.

Alas, as with all things, there are limitations in fitting against QM data. One limitation is that the exact answer prescribed by QM may depend on the level of theory used, which is necessarily approximate. Perhaps a more important limitation in fitting to QM data is that MM and QM are fundamentally different. QM is able to couple chemistry in a way that MM is unable to. For example, whereas MM models may have one optimal length or angle for each bond with harmonic penalties for deviations, QM may reasonably have different optimal bond lengths and angles as the electronic wavefunction optimizes for different nuclear coordinates. Additionally, MM includes some unphysical

approximations. The r^{-12} MM repulsions means that energy errors for conformations with steric interactions could be different than they would be without the steric interactions. If dihedral parameters were to account for errors with the steric interaction, then the effect of the steric interaction would persist, inappropriately, when the encroaching atoms separate, as may happen during a simulation. In that case, the dihedral parameters would be erroneous. Thus, inclusion of many diverse conformations in training should enable some implicit coupling to different contexts, while minimizing the impact of artifacts that depend on conformation.

How force field development efforts juggle the issues of number of parameters, the balance of experimental and quantum mechanical targets, and the development of training data has varied over the years. Decades ago, extensive quantum mechanics calculations were largely intractable, so quantum calculations were limited to the conformations of small model systems most salient for force field training. This meant that a small number of parameters could be supported in training, to ensure reasonable sampling of each parameter. Recently, additional unique parameters have been adopted to more precisely describe the chemistry of different fragments [Pérez et al., 2007, Lindorff-Larsen et al., 2010].

1.11 Recent AMBER protein force field history

Owing to the success of AMBER protein force field 99SB (ff99SB) at agreeing with experimental properties [Hornak et al., 2006, Showalter and Bruschweiler, 2007, Li and Brüschweiler, 2009, Lange et al., 2010, Cerutti et al., 2010], this dissertation focuses on the AMBER line of force fields descending from ff94 [Cornell et al., 1995], on which ff99SB was based. The ff94/ff99SB protein atom types are illustrated in Figure 1.8. The backbone of every amino acid includes the N atom type for the amide nitrogen, H for the amide hydrogen (except in proline), CT for the α-carbon, H1 for the α-hydrogen, C for the carbonyl carbon, and O for the carbonyl oxygen. The atom type landscape is intentionally simple, differentiating chemistries only with clear physical justification. The amino acid with the most complicated atom typing is histidine,

where some nitrogens are protonated, having the NA atom type, and, in uncharged histidine, some aren't, having the NB atom type. The adjacent carbon atom types also change, with CW bound to the protonated NA nitrogen, and CV bound to the deprotonated NB nitrogen. Otherwise, ff94 utilizes a minimal number of atom types.

Table 1.1: The ff94/ff99/ff99SB atom types for natural amino acids defined by Cornell et al. [1995]

Atom type	Element	Hybridization	Bonding environment
CT	C	$sp3$	Aliphatic
CA	C	$sp2$	Aromatic
CB	C	$sp2$	Aromatic, 5- & 6-member ring junction
CC	C	$sp2$	Aromatic, 5-member ring ring
CR	C	$sp2$	Aromatic, 5-member ring, between two nitrogens
CV	C	$sp2$	Aromatic, 5-member ring with 1 N and 1 H (His)
CW	C	$sp2$	Aromatic, 5-member ring with 1 NH and 1 H (His)
C*	C	$sp2$	Aromatic, 5-member ring with 1 substituent
C	C	$sp2$	Carbonyl
H	H	s	Nitrogen
HC	H	s	C without electron withdrawing group
H1	H	s	C with 1 electron withdrawing group
H2	H	s	C with 2 electron withdrawing groups
H3	H	s	C with 3 electron withdrawing groups

Table 1.1: Continued

Atom type	Element	Hybridization	Bonding environment
H4	H	s	Aromatic C with 1 electron withdrawing group
H5	H	s	Aromatic C with 2 electron withdrawing group
HA	H	s	Aromatic C without electron withdrawing group
HO	H	s	Hydroxyl
HP	H	s	C next to positively charged group
HS	H	s	Sulfhydryl
N	N	$sp2$	Amide
N3	N	$sp3$	Charged groups
NA	N	$sp2$	5-member ring with H atom (His)
NB	N	$sp2$	5-member ring with lone pair (His)
O	O	$sp2$	Carbonyl
O2	O	$sp2/sp3$	Carboxyl
OH	O	$sp2$	Hydroxyl
SH	S	$sp3$	Sulfhydryl
S	S	$sp3$	Disulfide

In ff94, bond vibrational parameters were determined from infrared spectroscopy and crystal structure geometries, partial charges from restrained fits to HF/6-31G* electrostatic potentials (ESPs) [Bayly et al., 1993, Cornell et al., 1993], and van der Waals parameters from liquid simulations or from OPLS [Jorgensen and Tirado-Rives, 1988]. The HF/6-31G* charge model was used as this level of theory and basis set tends to approximate the observed $\sim 20\%$ increase in TIP3P water charges relative to the gas phase charge distribution [Bayly et al., 1993, Cornell et al., 1993, 1995]. Therefore, HF/6-31G* allows the rapid assignment of atomic charges for solution based on gas phase quantum calculations.

Dihedral parameters are also used, in which the energy profile for rotation

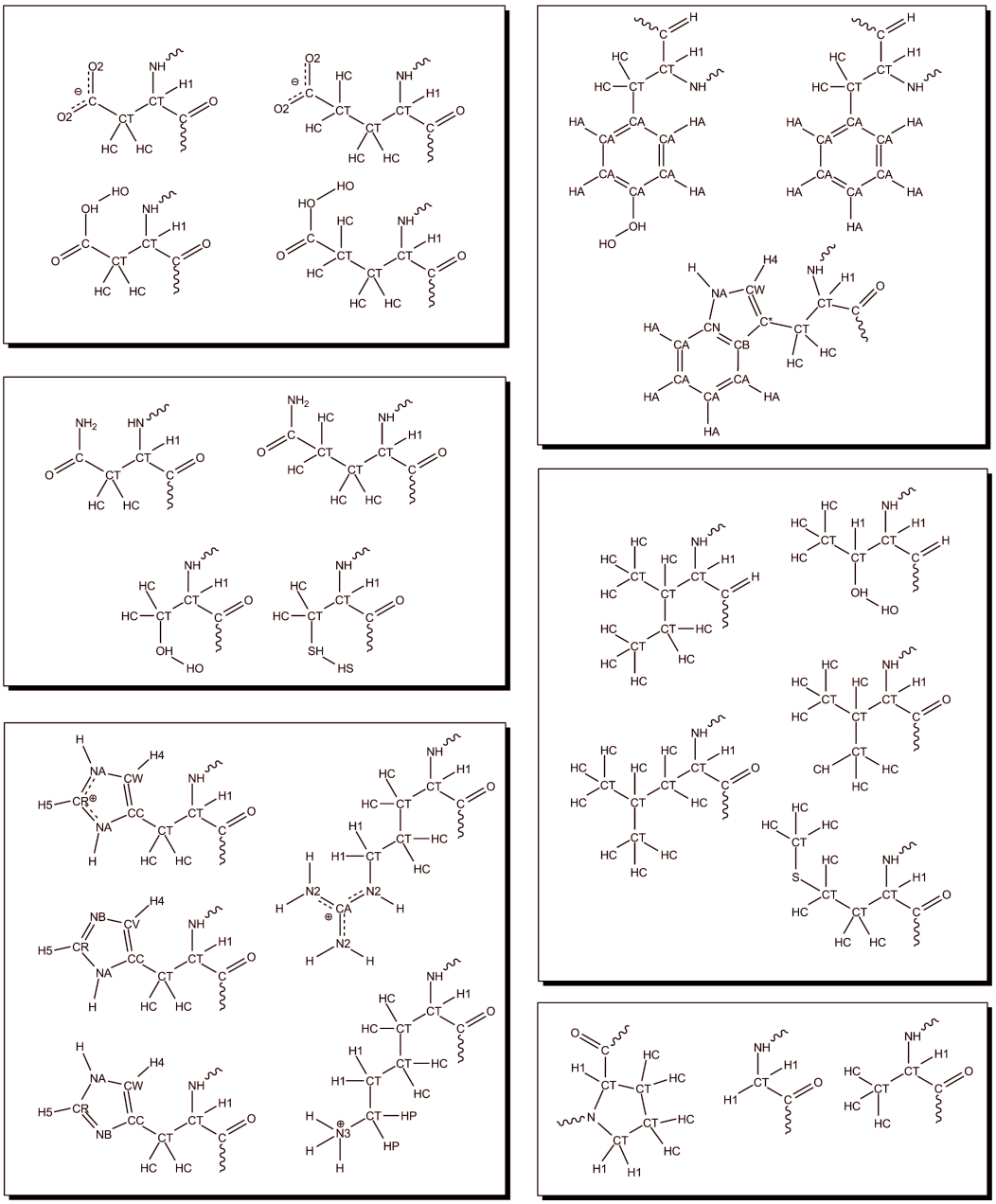

Figure 1.8: The ff94 atom types of each amino acid, separated by group. In the first column are acids, polar residues, and bases. In the second column are aromatics, nonpolar residues, and amino acids with less (Pro) or more (Gly, Ala) flexibility.

around bonds has a Fourier series adjustment applied to the profile arising from the other terms (such as sterics). A key assumption in these force fields is that the dihedral corrections are uncoupled, and thus the correction has no explicit dependence on values of neighboring dihedrals. Optimization of dihedral corrections is typically the last step in fitting force field parameters. "Generic" torsions applying to all sets of four atoms around a bond between two atom types (using a wildcard for the outer 2 atoms) were fit to a set of experimental small molecule barrier heights. In ff99 [Wang et al., 2000], multiple-periodicity specific torsional parameters applicable to protein side chains were fit to a larger set of small molecules.

An important component of protein force fields is the "backbone" dihedral parameters that can alter secondary structure preferences. Since the backbone ϕ and ψ rotations around 2-atom bonds each have contributions from multiple sets of 4 atom dihedral terms (due to multiple atoms bonded to the central 2 atoms), there has been some confusion over the years as to which of the dihedral terms should be optimized and applied to various amino acids, such as glycine (which lacks Cβ), proline (which lacks HN) and all others. In ff94, the baseline backbone dihedral profile for ϕ (C–N–Cα–C) and ψ (N–Cα–C–N) dihedral corrections were fit to glycine dipeptide conformation energies from QM. Then, the influence of the side chain was added by fitting parameters for the so-called ϕ' (C–N–Cα–Cβ) and ψ' (Cβ–Cα–C–N) based on alanine dipeptide QM conformational energies. Importantly, the ϕ' and ψ' were fit as a correction on top of the ϕ and ψ parameters that had already been fit to glycine. Thus all amino acids except glycine had 2 full sets of "backbone" dihedral contributions, one for the backbone and a second correction set using the Cβ atom. The ubiquity of ff94-based force fields shows its overall effectiveness, despite specific weaknesses in performance for proteins, such as exaggerated helical propensity [Hornak et al., 2006].

Several attempts to improve secondary structure balance were reported. In ff96 and ff99, the backbone ϕ/ψ dihedral corrections were adjusted to better reproduce the QM energies of a set of alanine dipeptide (blocked Ala1) and tetrapeptide (blocked Ala3) conformation energies [Beachy et al., 1997, Wang et al., 2000]. This decision was significant as the tetrapeptide can form a single

α-helical hydrogen bond, and thus have a local minimum in the helical conformation, which is absent for the dipeptide in the gas phase. This is especially important since only minima were used in fitting energy profiles. However, the approach in ff99 had two significant weaknesses: first, the optimization used alanine energies, but the parameter fitting was done for the C-N-CA-C and N-CA-C-N 4-atom dihedrals, which had been fit to glycine in ff94. The ff94 ϕ' and ψ' corrections were left in place during the ff99 fitting. As a result, simulations of glycine with ff99 employed ϕ and ψ parameters that were fit to alanine, rather than to glycine as originally intended [Hornak et al., 2006]. The second problem in ff99 was that the fitting was done to reproduce the energies of multiple backbone conformations, but each energy was defined relative to the energy of the helical conformation (since the zero of energy is arbitrary in MM) [Wang et al., 2000]. However, this procedure resulted in an overly strong influence of the helical conformation energy in ff99 [Hornak et al., 2006], and as a result ff99 favored helices even more strongly than the ff94 model that it was intended to improve. Both of these problems were recognized and largely corrected in ff99SB [Hornak et al., 2006].

With ff99SB, protein backbone dihedrals were refit by expanding upon the methods used in ff94 and ff99. A larger set of alanine tetrapeptide conformations were used in fitting ϕ' and ψ', as well as introducing glycine tetrapeptide conformations for fitting ϕ and ψ. The problem with using the helical structure as a reference was resolved by using as a fitting target the relative energies of all conformation pairs. These relative energies are what control populations and barriers in an MM model, and thus they were direct targets in the parameter optimization. The conformations were limited to local minima because of computational expense, but the fitting struck a balance of secondary structure suitable for a range of systems [Hornak et al., 2006, Showalter and Bruschweiler, 2007, Li and Brüschweiler, 2009, Lange et al., 2010, Cerutti et al., 2010]. Resultantly, ff99SB became widely adopted in the simulation community.

Limitations in models often only become apparent after extensive use and testing. One advantage of the wide adoption of ff99SB is that trends in the weaknesses were noted, as compared to single anecdotal failures for which the

cause may be difficult to determine, or unknown weaknesses in force fields that are not widely distributed. Most notably, rotamer preferences for several side chains were observed to be less accurate than others [Lindorff-Larsen et al., 2010]. This likely arose since ff99SB inherited amino acid side chain dihedral parameters from ff99, which were derived against a limited set of relative energies for small organic compounds [Wang et al., 2000]. The transferability of energy correction parameters for small molecules with relatively simple energy landscapes to amino acids may be an issue. The atoms in the amino acids typically have different partial atomic charges than the reference compounds, as well as more complex coupling to neighboring fragments. Due to recent increases in computational power, more extensive calculations (including full rotational energy profiles rather than selected stable conformations) can now be used to train side chain parameters directly against QM data on complete amino acids.

Although the secondary structure preferences in ff99SB were a dramatic improvement from previous Amber force fields, several studies noted room for improvement. After ff99SB was published, solution scalar coupling data for short peptides [Graf et al., 2007] became available, against which ff99SB and other force fields were compared [Best et al., 2008, Wickstrom et al., 2009], and the potential for improvement was discussed [Wickstrom et al., 2009]. Maier et al. hypothesized that two potential weaknesses in the ff99SB backbone parameter fitting strategy may be the dominant factors limiting accuracy: (1) the lack of backbone fitting data outside gas-phase minima and (2) using prepolarized MM partial charges intended for aqueous solution simulations while fitting dihedral parameters against gas-phase QM data [Maier et al., 2015]. Limiting the backbone parameter training to gas-phase local minima left potentially arbitrary energies for transition barriers or, importantly, in regions that become minima in solution or in the complex landscape of the protein interior. Additionally, the additive ff99SB model employs HF/6-31G* RESP partial atomic charges [Bayly et al., 1993, Cornell et al., 1993, 1995] that overestimate gas-phase dipoles by a similar amount as obtained in water models such as TIP3P [Jorgensen et al., 1983], thus approximating the polarization expected in aqueous solution [Bayly et al., 1993, Cornell et al., 1993]. However,

subsequent refitting of dihedral energy profiles to more accurate gas-phase energies calculated at the MP2 level results in dihedral parameters that may partially counteract the contribution of implicit polarization effects on the rotational energy profiles. Thus empirical corrections may provide additional benefit in reproducing experiments in water. While an alternative strategy to account for solvation effects in a more consistent way might be to develop an entirely new charge model [Duan et al., 2003, Cerutti et al., 2013], the original ff94 RESP charge model [Bayly et al., 1993, Cornell et al., 1993, 1995] developed by Peter Kollman has been extensively tested, and retaining it also maintains compatibility with many other parameter sets such as those modeling nucleic acids and carbohydrates [Kirschner et al., 2008]. Likewise, refitting the entire backbone dihedral profile rather than just minima would potentially lose the advantage of extensive studies [Showalter and Bruschweiler, 2007, Li and Brüschweiler, 2009, Wickstrom et al., 2009, Best et al., 2008, Thompson et al., 2010] evaluating ff99SB's strengths and weaknesses. Maier et al. [2015] investigated the simpler strategy of developing a small empirical adjustment to the ff99SB backbone parameters to improve reproduction of the experimental data in solution.

As described in more detail in Carmenza Martinez's dissertation [Martinez, 2014], Maier et al. created an array of small empirical perturbations to ϕ' and ψ (or ψ') torsions that were designed to overcome the shortcomings of the ff99SB training set described above [Maier et al., 2015]. Several combinations of ϕ' corrections with ψ or ψ' corrections were tested, pruning the number of variations as the test systems became more complex. This differs from other recent work that has focused on modifying a single torsional term to reproduce solution measurements directly [Best and Hummer, 2009, Nerenberg and Head-Gordon, 2011] or deriving ϕ/ψ coupled corrections against protein chemical shifts [Li and Brüschweiler, 2011]. The goal was parameters that were transferable between short disordered peptides (such as Ala5) and larger peptides with propensity to adopt stable secondary structure. Thus the ff14SB backbone approach was between that of Nerenberg and Head-Gordon [Nerenberg and Head-Gordon, 2011] and those of Best and Hummer [Best and Hummer, 2009], and Li and Brüschweiler [Li and Brüschweiler, 2011]. The best performer

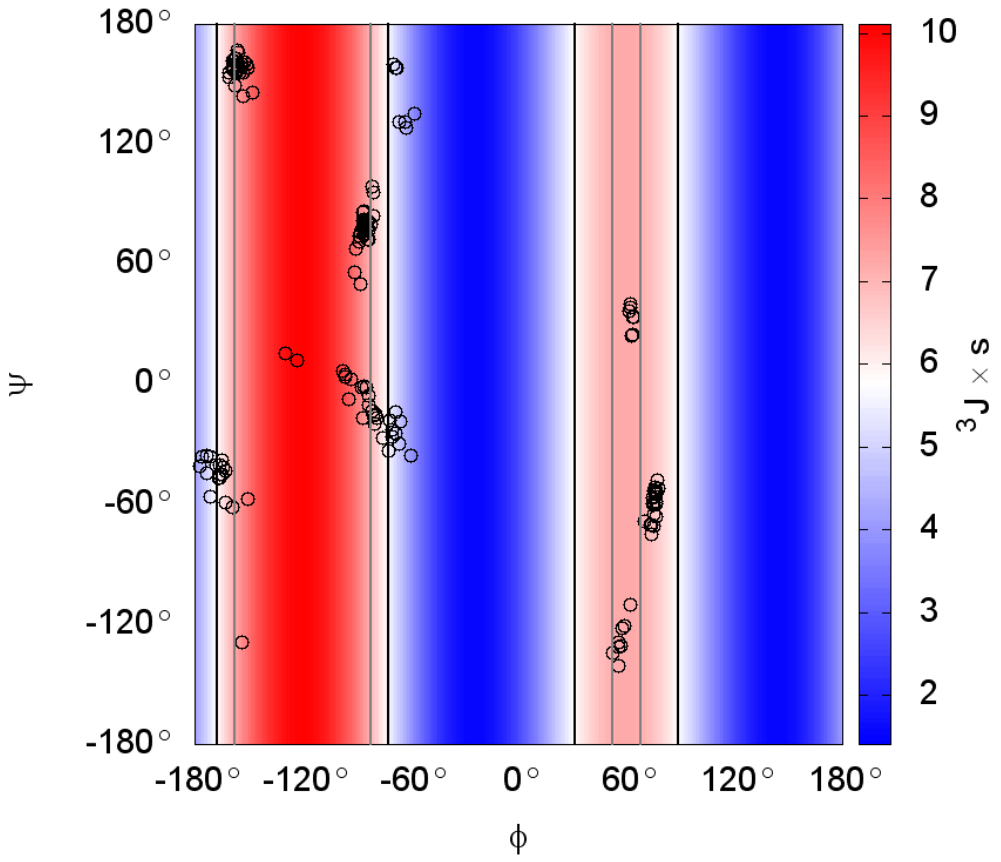

Figure 1.9: The ff99SB Ala_3 training set as circles, with the Hu and Bax [Hu and Bax, 1997] Karplus curve behind. Vertical lines indicate where the Karplus curve matches NMR (black) or ff99SB (gray). The maximum of the Karplus curve ($\phi = -120°$) is undersampled.

in small peptides was chosen as the backbone component to ff14SB [Maier et al., 2015].

A few studies have arisen, comparing force fields from AMBER and CHARMM to experimental data [Beauchamp et al., 2012, Lindorff-Larsen et al., 2012]. One particular goal of these studies was to evaluate the performance of force fields over time. Early force fields, such as CHARMM22 [MacKerell et al., 1998, Lindorff-Larsen et al., 2012] or ff99 [Wang et al., 2000, Beauchamp et al., 2012] had the greatest deviations from experiment. Results

with later force fields, particularly CHARMM22* [Piana et al., 2011] and variants of ff99SB [Hornak et al., 2006, Best and Hummer, 2009, Lindorff-Larsen et al., 2010, Nerenberg and Head-Gordon, 2011, Li and Brüschweiler, 2010], deviated least from experiment according to both studies [Beauchamp et al., 2012, Lindorff-Larsen et al., 2012]. Although both studies sought to assess the performance of force fields over time, ff99SB, upon which many later force fields were based, was inexplicably excluded. Yet, the trend remains that force fields are continuing to improve. According to Lindorff-Larsen et al. [2012], ff99SB*-ILDN and CHARMM22* both have an error score of 0. Although, as Lindorff-Larsen et al. [2012] concede that their score depends on "subjective choices," improving beyond the latest force fields will clearly require a thorough effort. In particular, fine details such as the possible necessity of amino acid-dependent parameters or precise details in energy calculations for training must be evaluated carefully. Hence, as described in Chapters 2 and 3, investigation of these options was pursued.

1.12 Outline

Chapter 2 describes development of the side chain component of ff14SB, whereby different approaches to refine side chain parameters were examined and the best applied to training parameters against QM data of dipeptides of each amino acid. Specifically, we investigated the effects of different levels of restraints on the side chain and backbone during side chain conformational sampling and the effect of re-optimizing structures molecular mechanically versus using quantum mechanically optimized structures to calculate molecular mechanics energies before training. Parameters derived against dipeptides with multiple backbone conformations resulted in better reproduction of experimental data, such as NMR $\chi 1$ scalar couplings.

Beyond the small backbone tweaks employed in ff14SB [Maier et al., 2015], Chapter 3 investigates whether the insights gained in Chapter 2 for QM-based fitting of side chain parameters can also be applied to more rigorously retrain the entire backbone energy surface. The chapter evaluates various choices in backbone parameter development. One choice is the level of sampling—energy

minima, as done for ff99SB [Hornak et al., 2006], structures from high temperature simulation, and grid-based scans, as done for the ff14SB side chain parameters [Maier et al., 2015] described in Chapter 2. Chapter 3 also probes structure optimization (QM versus MM structures), inclusion of implicit solvation effects in training, size of the training target (tetrapeptides, dipeptides, or "monopeptides"), and the sequence dependence of the necessary backbone correction. A method was derived that proved effective in reproducing Ala5 and Val3 backbone scalar couplings, reproducing the former as well as ff14SB, which was optimized against Ala5 scalar couplings, and the latter as well as ff99SB. A plan for the future force field enabled by the method developed is laid out.

Chapter 4 concerns the implications of Chapter 2 and Chapter 3 for development of force field models. The impact is discussed in terms of simulation results for the new side chain parameters. Additionally, the implications for the continuing role of QM-based force field training is discussed.

Chapter 2

Examine and improve accuracy of amino acid side chain sampling using molecular mechanics force field ff99SB

2.1 Acknowledgments

This chapter is reproduced in part with permission from the Journal of Chemical Theory and Computation, submitted for publication.

2.2 Introduction

There are several possible routes to attempting to improve the quality of the protein side chain model in ff99SB, such as using a more complex functional form for each dihedral, or including explicit coupling between various terms in the force field, or fitting to a higher level of QM theory. While each could potentially improve ff99SB, we focus here on improving the aspects that are

most likely to be the greatest weaknesses in the current model. Several recent reports of force field training have focused on application of more accurate quantum theory [Hornak et al., 2006, Pérez et al., 2007, Lindorff-Larsen et al., 2010, Best et al., 2012, Zgarbova et al., 2011, 2013]. While the level of theory is certainly important, we feel that improving the conformational diversity in the training data set is more likely to improve the model. In principle, dihedral parameters account for orbital effects missing in a classical model for bond rotation, but in practice also serve as empirical corrections for all differences between the QM and MM models, including lack of charge polarization changes during rotation, as well as dihedral-dependent errors in other terms in the force field (such as bond, angle and nonbonded interactions). As a result, the appropriate correction needed to match the MM torsion rotation energy profile to that obtained using QM may differ depending on chemical or conformational context, such as backbone conformation or other side chain torsions. In most biomolecular MM force fields, however, each rotatable bond is described by parameters that are independent of the conformation of the rest of the molecule (a notable exception is the CHARMM correction map (CMAP) of ϕ and ψ [MacKerell et al., 2004b]). As a result, while the net energy profile for rotating about a given bond will likely depend on other dihedrals (through steric effects, for example), the lack of explicit coupling in dihedral parameters limits the parameters to an implicit account for any coupling missing in the classical model. Therefore, it is important that the structures used for dihedral fitting include neighboring regions of the molecule where the parameters will be used. It is paramount to include conformational variety in those regions to avoid implicit coupling to a limited subset of their phase space, for example, a single rotamer or backbone conformation. In the present case, this led us to use complete amino acids in the QM calculations for the training data, as opposed to the small organic compounds used in ff99 [Wang et al., 2000]. Furthermore, implicit coupling was incorporated by fitting a single set of dihedral parameters using a large set of conformations that included multidimensional scans of all side chain χ rotatable bonds, with both α and β backbone conformations for the dipeptide. Thus, while the model lacks explicit coupling, the correction parameters for each dihedral are optimized in a mean field of conformational

variability for the remainder of the molecule.

The culmination of this work, ff14SB improves upon ff99SB and the backbone tweaks thereto in protein side chains (χ1 scalar couplings) and peptide helicity (CD, CSDs), while maintaining the quality of ff99SB local dynamics (Lipari-Szabo S^2 order parameters) and hairpin structure (CLN025). We recommend ff14SB.

2.3 Fitting Strategy and Goals

As discussed above, our general strategies for improving the protein force field have focused on fitting to QM calculations for systems more directly relevant to proteins (tetrapeptide for the backbone, dipeptide for side chains), and generating training sets of conformations with simultaneous variations in multiple dihedral angles so that the final parameters can include implicit (averaged) coupling between dihedral corrections. The goal is maximum transferability of the dihedral parameters for alternate chemical or conformational diversity in the remainder of the system.

It is important, however, to note that lack of explicit coupling in the dihedral correction parameters does not imply that the overall energy profiles are not coupled, since much of the observed coupling likely comes from steric and electrostatic through-space effects that are not directly related to the dihedral parameters. For this reason, a model with side chain χ parameters that do not depend on backbone conformation could still be capable of reproducing backbone-dependent side chain rotamer preferences [Lovell et al., 2000]. In this work, we extend the approach taken in ff99SB development, and the side chain χ dihedral parameter fitting is performed using multiple backbone conformations to implicitly average the corrections over possible backbone/side chain coupling needed to best reproduce the QM data and maximize transferability.

In ff94 and ff99, side chain torsions were decoupled by fitting data to small organic compounds representative of functionality seen in side chains, such as ethane, butane, or methanol, and deriving parameters based on these model compounds [Cornell et al., 1995, Wang et al., 2000]. Here, we used a more

realistic chemical context of the dipeptide, which also provides the opportunity to include implicit coupling to the presence of the backbone polypeptide chain, as well as include conformational diversity in the backbone in the side chain training set. In a recent revision of a small number of ff99SB side chain torsions that were identified by comparison of rotamer preferences in a helical context against the PDB, training against quantum mechanics energies for conformations with extended backbone improved $\chi 1$ rotamer preferences in β-rich proteins. However, while two of four amino acids showed considerable improvement in the helical test case, the other two showed more modest reduction in errors [Lindorff-Larsen et al., 2010]. As discussed above, our goal is to derive parameters that are transferable across chemical and conformational diversity, thus we explicitly included dipeptide conformations with both α and β backbone when sampling side chain rotational profiles. As with backbone parameters, we did not fit to side chain scalar couplings, but used them only to evaluate results of parameter changes. This differentiates our approach from CHARMM36 [Best et al., 2012], for example, where side chain parameters were finally adjusted to better reproduce $\chi 1$ scalar couplings.

Another choice concerns the diversity of conformations to use in the side chain rotamer training set. One option is to scan each dihedral rotamer separately, but as discussed above, this approach can fail to incorporate coupling needed in the correction terms, and may provide parameters that work well for some rotamer combinations, but fail for others that were absent in the training data. As including coupling via multidimensional scans can generate large numbers of conformations, one option to reduce the size of the training set is to include only minima. However, the exact locations of side chain χ minima can depend on backbone conformation, solvent, and packing with nearby residues. We thus made the decision to sample side chain conformations via full two-dimensional scans for the thirteen amino acids (counting different protonation states for Asp and His) with two side chain dihedrals (Table S1). For larger amino acids, conformational diversity was generated via symmetry considerations or dynamics simulations (described below). Because positions of minima may change in the context of a more complex system, and because energies for transitions may be relevant, each point in these scans

was considered equally important as compared to weighting data points by their energy. As side chain preferences are coupled to the backbone [Dunbrack and Karplus, 1993, Lovell et al., 2000], we performed these scans at both α ($\phi, \psi = -60°, -45°$) and β ($-135°, 135°$) backbone conformations. While additional backbone conformations could be employed, we considered only the archetypal α and β secondary structures due to computational cost. A separate ppII conformation ($-75°, 150°$) was not included, as the interaction of the side chain with the N-terminal peptide group is comparable in ppII and α conformations, while the interaction of the side chain with the C-terminal peptide is comparable in ppII and β conformations, thus these interactions are represented in the two backbone conformations already included in our set.

Our fitting targets for the side chains were gas-phase ab initio energies, as in ff99SB. To accommodate the 15 082 dipeptide conformations in our training set, we employed a relatively modest level of theory, with geometries calculated at HF/6-31G* and single point energies calculated at MP2/6-31+G**. Given the fundamental approximations, such as additivity, fixed partial charges, r^{-12} repulsion, and harmonic bond and angle vibrations, we do not expect the quantum theory to be the limiting factor in improving our model and focused on increasing the conformational diversity in the training set.

Additional choices that must be made relate to the generation of the QM and MM energies for conformations in the training set. First, we investigated what restraints to use in potential energy surface scans. Restraining the 4-atom set defining each χ dihedral, as well as those for ϕ and ψ, is natural given the goal of scanning combinations. Less obvious is whether other dihedral restraints should be included for the rotatable bond being scanned, such as those sharing the same 2 atoms defining the central bond, but varying the outer atoms. For example, the restraint for χ1 in Val is defined using N–Cα–Cβ–Cγ1, but the dihedral N–Cα–Cβ–Cγ2 could either be restrained or allowed to freely optimize in the presence of the χ1 restraint. Another choice is whether (and how strongly) to restrain other parts of the molecule, such as methyl rotations, or the peptide ω rotation. Next, given the fundamental

differences in QM and MM models, and weakness in MM description of energetics beyond dihedral profiles, we investigated whether to optimize geometries once—calculating molecular mechanical energies of quantum mechanical structures—or to re-optimize the QM structures with the MM model prior to comparing energies. The energies could be calculated for identical structures, for example the quantum mechanical structures. An advantage of this approach is that all coordinates and non-bonded distances would be identical. Alternatively, energies could be compared for structures optimized with the corresponding method (i.e. MM energies for MM-optimized structures). The MM model may not reproduce small changes in bond and angle geometries for different rotamers in the QM model, and the stiffness of the MM quadratic function could result in these differences making large contributions to energy errors that could be relaxed with MM structural re-optimization (as would also occur during molecular dynamics), thus focusing the resulting energy profile on the rotamer changes rather than MM covalent structure approximations.

Like several other MM force fields, the Amber-related models have traditionally used atom types to apply a small number of bond, angle and dihedral parameters to similar fragments in different amino acids. Ideally, the parameters would be highly transferable, and show accuracy for a variety of contexts. This approach reduces the number of parameters needed, but also limits the accuracy of the model since the implicit coupling we seek is worsened when the parameters are averaged over too great a variety of neighboring functional groups that can influence charge distribution. Since many sets of four atoms in the amino acid backbone and side chains shared the same atom types (and therefore the same dihedral corrections) with each other and also with nucleic acids in ff99SB, new atom types were created when needed to improve specificity. For example, asparagine $\chi 2$ (Cα–Cβ–Cγ–Nδ), glutamine $\chi 3$ (Cβ–Cγ–Cδ–Nϵ), and ψ' (Cβ–Cα–C–N) all shared atom types `CT-CT-C -N` , and therefore the same dihedral corrections applied to all three bonds. Here, additional atom types were created to allow independent adjustment of backbone parameters and different side chain parameters, with all atom types depicted in Figure 2.1. A new atom type for the α-carbon (CX) was created to separate main chain, $\chi 1$, and $\chi 2$ parameters, enabling the independent adjustment of

the side chain parameters. Where cross-referencing simulation data and errors fitting quantum energies suggested that solving corrections for particular amino acids together led to inaccuracies that solving separately would alleviate, additional atom types were also introduced to segregate them. Within the side chains, atom types 2C and 3C were developed for carbons bound to two or three heavy atoms, respectively, more thoroughly describing branched amino acids while isolating the revisions to amino acids (and preventing application to nucleic acids, which was possible in previous models). The CO atom type was introduced to distinguish carboxylate carbon from other carbonyl carbons. The C8 atom type was added for arginine and lysine, to distinguish them from glutamate, glutamine, and methionine. Each side chain atom type was added only if it allowed better reproduction of both quantum mechanics fitting targets and dynamic properties (described more fully below), to verify that additional parameters are appropriate.

Potential limitations in our approach We retain many of the approximations present in ff99SB, such as weaknesses in the harmonic description of covalent bonds and angles (Equation (1.19)), as well as the 6-12 Lennard-Jones function (Equation (1.22)). We also retained the same 1-4 nonbonded scaling factors employed with ff99SB. We refit all side chain dihedral parameters except the "generic" terms applied to nonpolar hydrogens, which were left at the values from ff99. We continue to assume that explicit coupling between dihedral pairs can be neglected. We additionally assume that gas-phase comparison against MP2/6-31+G**//HF/6-31G* quantum energies is sufficient to improve the side chain parameters. Ultimately, improving the level of theory or adding solvent in QM may alleviate some errors or inconsistencies in this approach. These assumptions, however, allowed us to overhaul the side chain dihedral parameters that had been carried over from ff94, in the context of the RESP charge model used in many force fields.

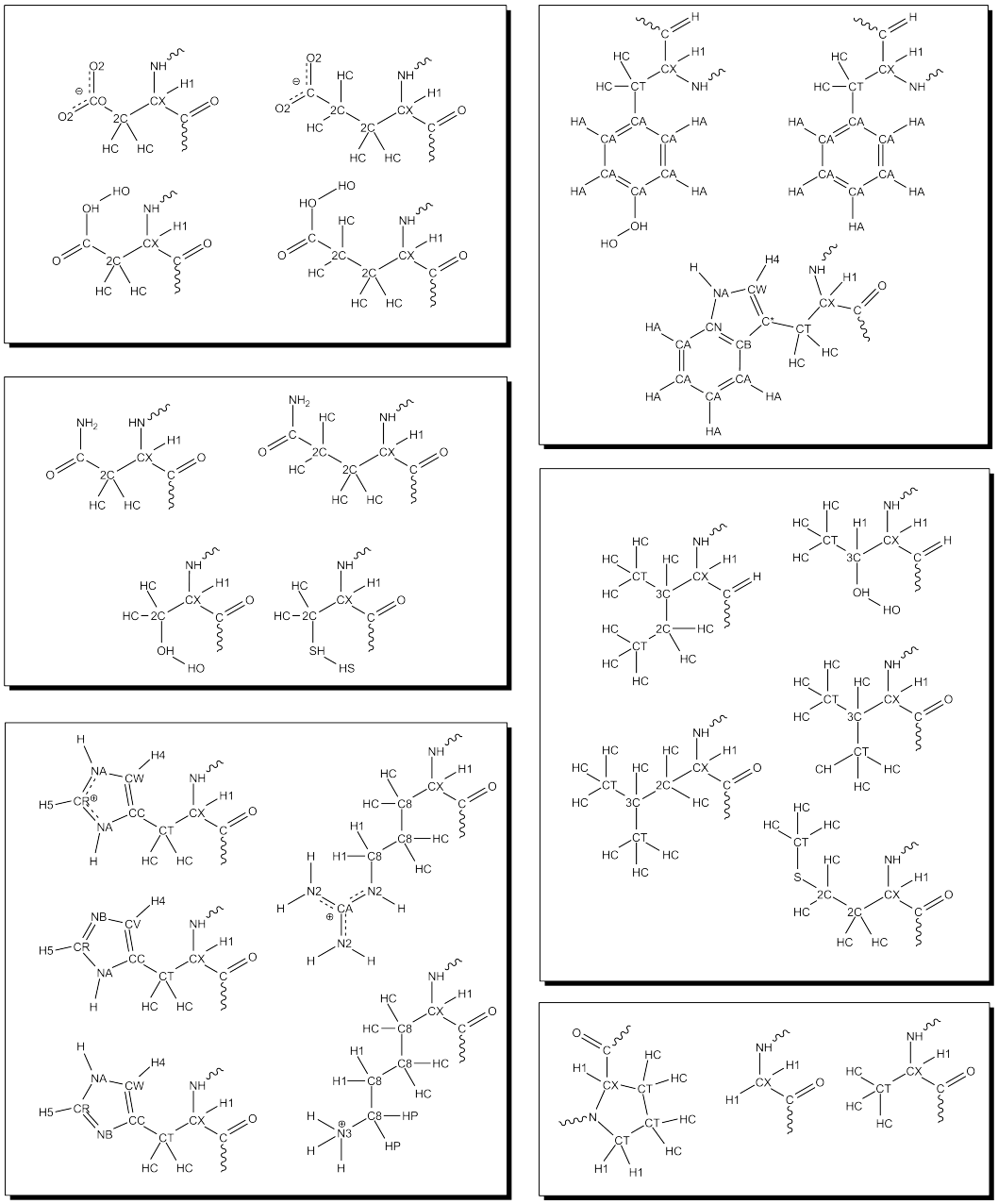

Figure 2.1: The amino acids drawn with their AMBER ff14SB [Maier et al., 2015] atom types.

2.4 Methods

2.4.1 Side chain dihedral training

Structure generation

Acetyl and N-methyl capped dipeptides of the natural amino acids, except proline, alanine, and glycine, were built using LEaP [Zhang et al., 2010] at α $(-60°, -45°)$ and β $(-135°, 135°)$ backbone conformations. We explored $\vec{\chi}$ by rotating in 10° increments, re-optimizing at each step, or by high temperature simulation (described in Results).

Molecular mechanics optimizations were performed with ff99SB using the sander module of AMBER11 [Case et al., 2005, 2012] for a maximum of 1.0×10^7 cycles or until the RMS gradient was less than 1.0×10^{-4} kcal mol^{-1} Å^{-1}, with a non-bonded cutoff of 99.0 Å and initial step size of 10^{-4}. All backbone torsions and one torsion per side chain rotatable bond were restrained with weights of 2×10^5 kcal mol^{-1} rad^{-2}. Minimization proceeded by 10 steps of steepest descent followed by conjugate gradient. Molecular mechanics energies were calculated from the last step of ff99SB minimization.

Quantum mechanics optimizations were performed with RHF/6-31G*. Scanned residues were optimized using GAMESS (US) [Schmidt et al., 1993, Gordon and Schmidt, 2005], version 1 MAY 2012 (R1), with default options. Optimization continued until the RMS gradient was less than 1.0×10^{-4} Hartree/Bohr, with an initial trust radius of 0.1 Bohr that could then adjust between 0.05 and 0.5 Bohr. Minimization proceeded by the quadratic approximation. Residues sampled by high temperature simulations were optimized using Gaussian98, Revision A.7 [Frisch et al., 1998] with VTight convergence criteria. Quantum mechanics energies were calculated with MP2/6-31+G**.

Generating conformational diversity in the training set

Scans To maximize transferability of the parameters, multidimensional structure scans were employed to generate conformational diversity and incorporate implicit coupling, despite lack of explicit coupling in the energy function. For

smaller side chains, we used grid scans in dihedral space to generate side chain variety, including both α and β backbone conformations for each side chain rotamer. Grid scans were generated for Val in one dimension, as it only has $\chi 1$, at an interval of 10°. Grids were generated for Asp^-, Asn, Cys, Phe, His (δ-, ϵ-, and doubly-protonated), Ile, Leu, Ser, Thr, and Trp in two dimensions, as they have $\chi 1$ and $\chi 2$, at intervals of 20°, yielding 324 structures per each combination of amino acid and backbone conformation.

We were unable to exhaustively explore side chain conformational space side chains with more than two rotatable bonds. Tyrosine has 3 rotatable χ bonds, but dihedral space is reduced since 180° rotation of either the phenol ($\chi 2$) or of the hydroxyl produce the same effect when accounting for symmetry of the ring. We therefore fully scanned each tyrosine dihedral when the other two were at a stable rotamer defined as any instance of that value in the rotamer library for this amino acid, rounded to the nearest 10° and limiting $\chi 2$ to $(-90°, 90°]$ to account for symmetry. Stable rotamers were 180° or ±60° for $\chi 1$, ±30° or 90° for $\chi 2$, and 0° or 180° for the hydroxyl. Conformations were generated using a full scan for each dihedral (at 20° increments), repeated for every combination of stable rotamer values for the other two dihedrals, yielding 288 conformations per backbone conformation. As protonated aspartate has nearly the same dihedrals as Tyr ($\chi 1$, $\chi 2$, and hydroxyl), we scanned it in the same manner, but without $\chi 2$ restriction because aspartate does not have the same symmetry properties, yielding 576 conformations per backbone conformation.

Cysteine presents a special case, as it can form disulfide bonds that cross a single amino acid. In addition to developing parameters for reduced Cys (no disulfide), we employed a pair of Cys dipeptides with a disulfide bond to scan the S–S energy profile. However, a disulfide between Cys_A and Cys_B has a total of five dihedrals: $\chi 1_A$, $\chi 2_A$, χ_{SS}, $\chi 2_B$, and $\chi 1_B$. Since full sampling across five dihedrals is clearly intractable, we reduced the conformation space by applying the same $\chi 1/\chi 2$ values to both dipeptides. Using this symmetry, we performed a two-dimensional scan for all $\chi 1/\chi 2$ combinations using 20° spacing; this scan was repeated with χ_{SS} restrained to 180°, ±60°, or ±90° (five 2D scans). Separately, the χ_{SS} profile was scanned with 20° spacing using

$\chi 1$ of 180° or ±60° and $\chi 2$ of 180° or ±60° (nine 1D scans total). As with the other amino acids, the entire procedure was repeated with the backbone in α and β conformations; here, both dipeptides adopted the same backbone conformation.

Simulations The remaining side chains, Arg^+, Gln, Glu (protonated), Glu^-, Lys^+, and Met, have at least three side chain dihedrals (Table S1). Rather than performing a grid search, we used MD simulations to generate diverse side chain conformations for these side chains. We simulated each dipeptide twice, with α or β backbone restraints, for 100 ns each. To overcome kinetic traps, these simulations were performed at 500 K and the dielectric was set to 4 r. Next, we generated a diverse subset by mapping each conformation to a multidimensional grid spaced 10° in each χ. We saved the five lowest energy conformations at each grid point. From each simulation grid, five hundred structures were randomly selected (comparable to the number generated by the grid procedure described above for Tyr). Since the longer, more flexible side chains of these amino acids can adopt conformations with strong interactions between backbone and side chain, we excluded conformations where we suspected the in vacuo MM description may produce fitting artifacts, using electrostatic and distance cutoffs defined below.

Simulations details Simulations were maintained at 500 K by a Langevin thermostat with $\gamma_{\ln} = 1.0$. Restraints of 2×10^3 kcal mol^{-1} rad^{-1} on ϕ and ψ maintained the backbone throughout. The simulations were integrated with a 1 fs timestep. Structures were saved every 2 fs to capture short-lived transitions.

To generate sets structures that varied principally in side chain conformation, minor differences in backbone conformation had to be reconciled. For each set of structures, the average value of every backbone dihedral was determined. Each structure was minimized for 100 000 cycles using ff99SB, with incremental restraints on the dihedrals describing all four-atom torsions within the backbone to their average values at weights of 100 kcal mol^{-1} rad^{-2}, 500 kcal mol^{-1} rad^{-2}, 1×10^3 kcal mol^{-1} rad^{-2}, 5×10^3

$\mathrm{kcal\,mol^{-1}\,rad^{-2}}$, and then $1 \times 10^4\,\mathrm{kcal\,mol^{-1}\,rad^{-2}}$, for 500 steps each, and finally $1.5 \times 10^4\,\mathrm{kcal\,mol^{-1}\,rad^{-2}}$ for the remainder of minimization. The methyl groups at the N- and C-termini, due to their C3 symmetry, were not restrained to their average values but to $-5°$ for one of the H−C−C−N and $60°$ for one of the C−N−C−H dihedrals.The simulations generated some structures with challenging sterics or electrostatics, where atoms were within close proximity. One concern is that the molecular mechanical model employs charges that are fixed for a dielectric. A second concern is that the r^{-12} Lennard-Jones approximation of repulsion is an unphysical mathematical convenience (the dispersive r^{-6} squared) that may be too hard at close range. We did not want to fit the errors of fixed charges in a vacuum, nor possible MM repulsion artifacts. Structures where the distances between atoms not in a bond or angle breached the sum of their van der Waals radii [Bondi, 1964] divided by 1.3 were eliminated. The scaling factor 1.3 was chosen empirically as a value that left a reasonable number of structures in the training set but targeted those with the greatest degree of contact.

A second non-bonded filter specifically targeted electrostatic interactions. When the Coulombic energy between a side chain particle exceeded $42\,\mathrm{kcal\,mol^{-1}}$ in magnitude, the interaction was evaluated "very strong" and the structure was discarded from the training set. Although any number between 40 and $43\,\mathrm{kcal\,mol^{-1}}$ could have been suitable, 42 received preference based on previous work [Adams, 1979]. Both ff99SB and MP2/6-31+G** Mulliken charges were considered in this evaluation.

Structures with relative energies greater than $30\,\mathrm{kcal\,mol^{-1}}$ were considered overly strained and were also therefore removed from the training.

Summary Thus, valine has 72 structures, tyrosine has 576, protonated aspartate has 1152, all 2d-scanned residues have 648, and the remaining residues have the numbers in Table 2.1.

Table 2.1: Distribution of REEs in terms of mean, standard deviation (stdev), minimum (min), and maximum (max), using ff99SB or ff14SB side chain parameters, against the conformations of each amino acid (AA) and backbone conformation (BB). '# confs' is the number of conformations in each set.

			ff99SB				ff14SB			
AA	BB	# confs	mean	stdev	min	max	mean	stdev	min	max
Arg	α	396	1.41	1.07	0.00	6.97	1.06	0.86	0.00	7.87
Arg	β	355	1.50	1.23	0.00	8.71	1.06	0.96	0.00	8.10
Ash	α	576	1.87	1.44	0.00	10.51	1.22	0.96	0.00	7.76
Ash	β	576	2.06	1.55	0.00	10.71	0.96	0.78	0.00	7.21
Asn	α	324	1.99	1.50	0.00	8.70	1.06	0.79	0.00	4.86
Asn	β	324	1.97	1.44	0.00	9.02	0.97	0.73	0.00	4.62
Asp	α	324	2.63	1.95	0.00	10.65	1.11	0.85	0.00	4.63
Asp	β	324	2.46	1.76	0.00	8.76	0.73	0.68	0.00	4.43
Cys	α	324	1.42	1.17	0.00	7.11	1.14	1.02	0.00	5.43
Cys	β	324	1.15	0.84	0.00	5.14	0.91	0.68	0.00	3.95
Glh	α	436	1.63	1.25	0.00	9.29	1.23	1.01	0.00	7.53
Glh	β	420	1.48	1.11	0.00	7.14	0.92	0.73	0.00	4.88
Gln	α	313	1.69	1.30	0.00	8.18	1.23	0.92	0.00	6.49
Gln	β	282	1.29	1.03	0.00	7.01	0.91	0.70	0.00	4.52
Glu	α	284	1.96	1.48	0.00	7.62	1.26	0.97	0.00	5.95
Glu	β	271	2.03	1.52	0.00	8.40	1.29	1.00	0.00	6.31
Hid	α	324	1.09	0.83	0.00	5.52	0.76	0.56	0.00	3.71
Hid	β	324	1.24	0.99	0.00	6.66	0.79	0.65	0.00	4.14
Hie	α	324	1.24	0.99	0.00	7.51	0.92	0.69	0.00	4.52
Hie	β	324	1.37	1.00	0.00	6.11	0.99	0.72	0.00	3.98
Hip	α	324	1.55	1.18	0.00	6.24	1.46	1.08	0.00	6.24
Hip	β	324	1.36	1.03	0.00	7.02	1.04	0.82	0.00	6.45
Ile	α	324	1.64	1.20	0.00	6.60	0.97	0.73	0.00	4.62
Ile	β	324	1.17	0.87	0.00	5.40	0.77	0.58	0.00	3.42
Leu	α	324	1.19	0.87	0.00	4.86	1.01	0.78	0.00	4.45
Leu	β	324	1.28	0.93	0.00	4.87	0.76	0.56	0.00	3.30

Table 2.1: Continued

AA	BB	# confs	ff99SB mean	ff99SB stdev	ff99SB min	ff99SB max	ff14SB mean	ff14SB stdev	ff14SB min	ff14SB max
Lys	α	466	1.66	1.34	0.00	8.07	1.30	1.08	0.00	7.80
Lys	β	439	1.38	1.09	0.00	8.07	0.96	0.79	0.00	6.17
Met	α	483	1.33	1.10	0.00	8.01	1.16	0.96	0.00	7.30
Met	β	497	1.22	1.01	0.00	7.96	1.02	0.85	0.00	6.82
Phe	α	324	0.86	0.70	0.00	4.47	0.88	0.67	0.00	4.24
Phe	β	324	0.98	0.75	0.00	4.08	0.77	0.60	0.00	3.56
Ser	α	324	1.68	1.26	0.00	7.67	1.01	0.75	0.00	4.50
Ser	β	324	1.17	0.88	0.00	5.00	0.71	0.53	0.00	3.04
Thr	α	324	1.58	1.16	0.00	6.37	0.92	0.73	0.00	4.14
Thr	β	324	1.22	0.89	0.00	5.00	0.81	0.63	0.00	3.51
Trp	α	324	1.37	1.15	0.00	9.76	1.13	1.12	0.00	11.54
Trp	β	324	1.25	1.02	0.00	8.26	0.91	0.95	0.00	8.81
Tyr	α	288	2.44	1.92	0.00	8.63	0.83	0.62	0.00	3.82
Tyr	β	288	2.45	1.92	0.00	8.85	0.73	0.55	0.00	3.43
Val	α	36	1.17	0.86	0.00	3.37	0.77	0.58	0.00	2.77
Val	β	36	0.53	0.37	0.00	1.44	0.36	0.32	0.00	1.08

Objective function for parameter optimization

As in ff99SB, we evaluated the errors of relative energies between all pairs of conformations to alleviate the choice of a single, potentially arbitrary reference conformation. We first defined the relative energy error (REE) between a single pair of conformations i and j

$$\mathrm{REE}(i,j) = (E_{\mathrm{QM},i} - E_{\mathrm{QM},j}) - (E_{\mathrm{MM},i} - E_{\mathrm{MM},j}) \tag{2.1}$$

where $E_{\mathrm{QM},i}$ and $E_{\mathrm{MM},i}$ are the quantum and molecular mechanics energies of conformation i. E_{MM} is calculated as either ff99SB or, during parameter search, as the ff99SB energy with the dihedral energy, $E_{\chi}^{\mathrm{ff99SB}}$, replaced using

candidate dihedral parameters, yielding $E_{\text{ff\_new}}$:

$$E_{\text{ff\_new}} = E_{\text{ff99SB}} + \sum_{\chi} E_{\chi}^{\text{ff\_new}} - E_{\chi}^{\text{ff99SB}} \tag{2.2}$$

where the sum is taken over all side chain rotatable bonds χ. For each force field ff, E_{χ}^{ff} is the sum of dihedral contributions of N_{χ} sets of four atoms around each rotatable bond, excluding those containing nonpolar hydrogens (Tables 2.2 to 2.13). For each dihedral, we summed over periodicity $n \in [1, 4]$ Fourier series contributions (c) with amplitudes $V_{\chi[c],n}^{ff}$ and phases $\gamma_{\chi[c],n}^{ff}$,

$$E_{\chi}^{ff} = \sum_{c}^{N_{\chi}} \sum_{n}^{4} V_{\chi[c],n}^{ff} \left(1 + \cos(n\phi_{\chi[c]} - \gamma_{\chi[c],n}^{ff})\right) \tag{2.3}$$

We note that this equation is consistent with the AMBER standard and lacks a factor of $\frac{1}{2}$; hence the true amplitude of each cosine is actually $2V_{\chi[c],n}^{ff}$. The fitting was limited to the fourth order term in each correction. Test fits using more terms resulted in noisier corrections without significantly altering fit quality. Aspartate, for example, converged with an error of $1.832\,\text{kcal}\,\text{mol}^{-1}$ with a max periodicity of 6, rather than $1.839\,\text{kcal}\,\text{mol}^{-1}$ with a max periodicity of 4. The simpler correction was preferred to such small ($< 0.4\%$) improvement.

To focus the energy differences on side chain rotamer profiles, we excluded comparisons between structures with different backbone conformations, or of different amino acids. Alternate protonation states for ionizable amino acids were summed separately. For each amino acid, in either α or β backbone conformation, we summed the magnitude of REE over all pairs of N side chain conformations, dividing by the number of pairs to obtain the average absolute error (AAE, as defined by Hornak et al. [2006]). The AAE for each amino acid, in a given protonation state, in a specific backbone conformation is

$$\text{AAE} = \frac{2}{N(N-1)} \sum_{i} \sum_{j<i} |\text{REE}(i,j)| \tag{2.4}$$

We then minimized the sum of the AAEs for each amino acid and backbone conformation by adjusting the amplitude and phase parameters for all terms

in Equation (2.3). Formally, we minimized the objective function

$$\mathscr{O} = \frac{1}{N_{\text{profiles}}} \sum_{r=1}^{\text{aminoacids}} \sum_{bb=\alpha,\beta} \text{AAE}_{r,bb} \tag{2.5}$$

where N_{profiles} is the number of profiles, resulting in a normalized $\mathscr{O}$ value that represents the error in energy differences for conformation pairs, averaged over all backbone contexts, amino acids and protonation states.

We optimized the parameters for all non-hydrogen dihedrals in the protein side chains describing rotation around single bonds, as well as hydroxyl or sulfhydryl torsions (Tables 2.2 to 2.14). As discussed above, our structure training set is designed to include amino acid conformation pairs with simultaneous changes to more than one rotatable bond, thus necessitating concurrent optimization of parameters for multiple dihedrals (rather than the simpler approach of scanning parameter space one rotatable bond at a time [Wang and Kollman, 2001]). This enables the optimized energy corrections for each rotatable bond to incorporate implicit coupling to nearby conformational diversity. Furthermore, the presence of similar local structure (as described by atom types) in multiple amino acids often led to the requirement for fitting parameters using data from all amino acids where that functionality exists. This provides parameters that implicitly account for nearby chemical diversity, as opposed to training in a single amino acid and use in others. As a result of these two design factors, the parameter space for optimization is considerable.

Table 2.2: Group 1 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-2C	Glh χ1 Gln χ1 Glu χ1 Met χ1	N-C_α-C_β-C_γ
C-CX-2C-2C	Glh χ1 Gln χ1 Glu χ1 Met χ1	C-C_α-C_β-C_γ
CX-2C-2C-C	Glh χ2 Gln χ2	C_α-C_β-C_γ-C_δ
2C-2C-C-OH	Glh χ3	C_β-C_γ-C_δ-$O_{\epsilon 2}$
2C-2C-C-N	Gln χ3	C_β-C_γ-C_δ-$N_{\epsilon 2}$
CX-2C-2C-CO	Glu χ2	C_α-C_β-C_γ-C_δ
2C-2C-C-O2	Glu χ3	C_β-C_γ-C_δ-$O_{\epsilon 1}$ C_β-C_γ-C_δ-$O_{\epsilon 2}$
CX-2C-2C-S	Met χ2	C_α-C_β-C_γ-S_δ
2C-2C-S-CT	Met χ3	C_β-C_γ-S_δ-C_ϵ

Table 2.3: Group 2 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-3C-CT	Ile $\chi 1$	N-C_α-C_β-$C_{\gamma 2}$
	Thr $\chi 1$	
	Val $\chi 1$	N-C_α-C_β-$C_{\gamma 1}$
		N-C_α-C_β-$C_{\gamma 2}$
C-CX-3C-CT	Ile $\chi 1$	C-C_α-C_β-$C_{\gamma 2}$
	Thr $\chi 1$	
	Val $\chi 1$	C-C_α-C_β-$C_{\gamma 1}$
		C-C_α-C_β-$C_{\gamma 2}$
N-CX-3C-2C	Ile $\chi 1$	N-C_α-C_β-$C_{\gamma 1}$
C-CX-3C-2C	Ile $\chi 1$	C-C_α-C_β-$C_{\gamma 1}$
N-CX-3C-OH	Thr $\chi 1$	N-C_α-C_β-$O_{\gamma 1}$
C-CX-3C-OH	Thr $\chi 1$	C-C_α-C_β-$O_{\gamma 1}$
CX-3C-2C-CT	Ile $\chi 2$	C_α-C_β-$C_{\gamma 1}$-$C_{\delta 1}$
CT-3C-2C-CT	Ile $\chi 2$	$C_{\gamma 2}$-C_β-$C_{\gamma 1}$-$C_{\delta 1}$
CX-3C-OH-HO	Thr oh	C_α-C_β-$O_{\gamma 1}$-$H_{\gamma 1}$

Table 2.4: Group 3 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-C	Ash $\chi 1$	N-C_α-C_β-C_γ
	Asn $\chi 1$	
C-CX-2C-C	Ash $\chi 1$	C-C_α-C_β-C_γ
	Asn $\chi 1$	
CX-2C-C-O	Ash $\chi 2$	C_α-C_β-C_γ-$O_{\delta 1}$
	Asn $\chi 2$	
CX-2C-C-OH	Ash $\chi 2$	C_α-C_β-C_γ-$O_{\delta 2}$
CX-2C-C-N	Asn $\chi 2$	C_α-C_β-C_γ-$N_{\delta 2}$

Table 2.5: Group 4 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-CT-CC	Hid χ1	N-C_α-C_β-C_γ
	Hie χ1	
	Hip χ1	
C-CX-CT-CC	Hid χ1	C-C_α-C_β-C_γ
	Hie χ1	
	Hip χ1	
CX-CT-CC-NA	Hid χ2	C_α-C_β-C_γ-$N_{\delta1}$
	Hip χ2	
CX-CT-CC-NB	Hie χ2	C_α-C_β-C_γ-$N_{\delta1}$
CX-CT-CC-CV	Hid χ2	C_α-C_β-C_γ-$C_{\delta2}$
CX-CT-CC-CW	Hie χ2	C_α-C_β-C_γ-$C_{\delta2}$
	Hip χ2	

Table 2.6: Group 5 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-CA	Phe χ1	N-C_α-C_β-C_γ
	Tyr χ1	
C-CX-2C-CA	Phe χ1	C-C_α-C_β-C_γ
	Tyr χ1	
CX-2C-CA-CA	Phe χ2	C_α-C_β-C_γ-$C_{\delta1}$
	Tyr χ2	C_α-C_β-C_γ-$C_{\delta2}$
CA-C-OH-HO	Tyr oh	$C_{\epsilon1}$-C_ζ-O_η-H_η
		$C_{\epsilon2}$-C_ζ-O_η-H_η

Table 2.7: Group 6 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-C8-C8	Arg χ1 Lys χ1	N-C_α-C_β-C_γ
C-CX-C8-C8	Arg χ1 Lys χ1	C-C_α-C_β-C_γ
CX-C8-C8-C8	Arg χ2 Lys χ2	C_α-C_β-C_γ-C_δ
C8-C8-C8-C8	Lys χ3	C_β-C_γ-C_δ-C_ϵ
C8-C8-C8-N3	Lys χ4	C_γ-C_δ-C_ϵ-N_ζ
C8-C8-C8-N2	Arg χ3	C_β-C_γ-C_δ-N_ϵ
C8-C8-N2-CA	Arg χ4	C_γ-C_δ-N_ϵ-C_ζ

Table 2.8: Group 7 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-SH	Cys χ1	N-C_α-C_β-S_γ
C-CX-2C-SH	Cys χ1	C-C_α-C_β-S_γ
CX-2C-SH-HS	Cys χ2	C_α-C_β-S_γ-H_γ

Table 2.9: Group 8 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-S	Cyx χ1	N-C_α-C_β-S_γ
C-CX-2C-S	Cyx χ1	C-C_α-C_β-S_γ
CX-2C-S-S	Cyx χ2	C_α-C_β-S_γ-$S_{\gamma'}$
2C-S-S-2C	Cyx χ_{SS}	C_β-S_γ-$S_{\gamma'}$-$C_{\beta'}$

Table 2.10: Group 9 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-CT-C*	Trp χ1	N-C_α-C_β-C_γ
C-CX-CT-C*	Trp χ1	C-C_α-C_β-C_γ
CX-CT-C*-CW	Trp χ2	C_α-C_β-C_γ-$C_{\delta 1}$
CX-CT-C*-CB	Trp χ2	C_α-C_β-C_γ-$C_{\delta 2}$

Table 2.11: Group 10 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-CO	Asp χ1	N-C_α-C_β-C_γ
C-CX-2C-CO	Asp χ1	C-C_α-C_β-C_γ
CX-CS-CO-O2	Asp χ2	C_α-C_β-C_γ-$O_{\delta 1}$ C_α-C_β-C_γ-$O_{\delta 2}$

Table 2.12: Group 11 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-3C	Leu χ1	N-C_α-C_β-C_γ
C-CX-2C-3C	Leu χ1	C-C_α-C_β-C_γ
CX-CS-3C-CT	Leu χ2	C_α-C_β-C_γ-$C_{\delta 1}$ C_α-C_β-C_γ-$C_{\delta 2}$

Table 2.13: Group 12 atom types of each correction modified, the residues, bonds, and atom names affected

Dihedral atom types	Bonds affected	Dihedral atom names
N-CX-2C-OH	Ser χ1	N-C_α-C_β-O_γ
C-CX-2C-OH	Ser χ1	C-C_α-C_β-O_γ
CX-2C-OH-HO	Ser oh	C_α-C_β-O_γ-H_γ

Table 2.14: The amino acids and the bonds that have been corrected, the four atom combinations, and the atom types of each correction that has been modified. This table has the same contents as Tables 2.2 to 2.13, but sorted by amino acid rather than solving group.

Amino acid	Rotatable bond	Atom names	Atom types	Solving group
Arg	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-C8-C8	6
		C-C_α-C_β-C_γ	C-CX-C8-C8	
	$\chi 2$	C_α-C_β-C_γ-C_δ*	CX-C8-C8-C8	
	$\chi 3$	C_β-C_γ-C_δ-N_ϵ*	C8-C8-C8-N2	
	$\chi 4$	C_γ-C_δ-N_ϵ-C_ζ*	C8-C8-N2-CA	
Ash	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-C	3
		C-C_α-C_β-C_γ	C-CX-2C-C	
	$\chi 2$	C_α-C_β-C_γ-$O_{\delta 1}$*	CX-2C-C-O	
		C_α-C_β-C_γ-$O_{\delta 2}$	CX-2C-C-OH	
	oh	C_β-C_γ-$O_{\delta 2}$-$H_{\delta 2}$*	2C-C-OH-HO	
Asn	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-C	3
		C-C_α-C_β-C_γ	C-CX-2C-C	
	$\chi 2$	C_α-C_β-C_γ-$O_{\delta 1}$*	CX-2C-C-O	
		C_α-C_β-C_γ-$N_{\delta 2}$	CX-2C-C-N	
Asp	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-CO	9
		C-C_α-C_β-C_γ	C-CX-2C-CO	
	$\chi 2$	C_α-C_β-C_γ-$O_{\delta 1}$*	CX-CS-CO-O2	
		C_α-C_β-C_γ-$O_{\delta 2}$	CX-CS-CO-O2	
Cys	$\chi 1$	N-C_α-C_β-S_γ*	N-CX-2C-SH	7
		C-C_α-C_β-S_γ	C-CX-2C-SH	
	$\chi 2$	C_α-C_β-S_γ-H_γ*	CX-2C-SH-HS	
Cyx	$\chi 1$	N-C_α-C_β-S_γ*	N-CX-2C-S	0
		C-C_α-C_β-S_γ	C-CX-2C-S	
	$\chi 2$	C_α-C_β-S_γ-$S_{\gamma'}$*	CX-2C-S-S	
	χ_{SS}	C_β-S_γ-$S_{\gamma'}$-$C_{\beta'}$*	2C-S-S-2C	
Glh	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-2C	1
		C-C_α-C_β-C_γ	C-CX-2C-2C	

Table 2.14: Continued

Amino acid	Rotatable bond	Atom names	Atom types	Solving group
	$\chi 2$	C_α-C_β-C_γ-C_δ*	CX-2C-2C-C	
	$\chi 3$	C_β-C_γ-C_δ-$O_{\epsilon 1}$*	2C-2C-C-OH	
		C_β-C_γ-C_δ-$O_{\epsilon 2}$	2C-2C-C-OH	
Gln	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-2C	1
		C-C_α-C_β-C_γ	C-CX-2C-2C	
	$\chi 2$	C_α-C_β-C_γ-C_δ*	CX-2C-2C-C	
	$\chi 3$	C_β-C_γ-C_δ-$O_{\epsilon 1}$*	2C-2C-C-O	
		C_β-C_γ-C_δ-$N_{\epsilon 2}$	2C-2C-C-N	
Glu	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-2C	1
		C-C_α-C_β-C_γ	C-CX-2C-2C	
	$\chi 2$	C_α-C_β-C_γ-C_δ*	CX-2C-2C-CO	
	$\chi 3$	C_β-C_γ-C_δ-$O_{\epsilon 1}$*	2C-2C-C-O2	
		C_β-C_γ-C_δ-$O_{\epsilon 2}$	2C-2C-C-O2	
Hid	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-CT-CC	4
		C-C_α-C_β-C_γ	C-CX-CT-CC	
	$\chi 2$	C_α-C_β-C_γ-$N_{\delta 1}$*	CX-CT-CC-NA	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-CT-CC-CV	
Hie	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-CT-CC	4
		C-C_α-C_β-C_γ	C-CX-CT-CC	
	$\chi 2$	C_α-C_β-C_γ-$N_{\delta 1}$*	CX-CT-CC-NB	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-CT-CC-CW	
Hip	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-CT-CC	4
		C-C_α-C_β-C_γ	C-CX-CT-CC	
	$\chi 2$	C_α-C_β-C_γ-$N_{\delta 1}$*	CX-CT-CC-NA	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-CT-CC-CW	
Ile	$\chi 1$	N-C_α-C_β-$C_{\gamma 1}$*	N-CX-3C-2C	
		C-C_α-C_β-$C_{\gamma 1}$	C-CX-3C-2C	
		N-C_α-C_β-$C_{\gamma 2}$	N-CX-3C-CT	2
		C-C_α-C_β-$C_{\gamma 2}$	C-CX-3C-CT	

Table 2.14: Continued

Amino acid	Rotatable bond	Atom names	Atom types	Solving group
	$\chi 2$	C_α-C_β-$C_{\gamma 1}$-$C_{\delta 1}$*	CX-3C-2C-CT	
		$C_{\gamma 2}$-C_β-$C_{\gamma 1}$-$C_{\delta 1}$	CT-3C-2C-CT	
Leu	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-3C	10
		C-C_α-C_β-C_γ	C-CX-2C-3C	
	$\chi 2$	C_α-C_β-C_γ-$C_{\delta 1}$*	CX-CS-3C-CT	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-CS-3C-CT	
Lys	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-C8-C8	6
		C-C_α-C_β-C_γ	C-CX-C8-C8	
	$\chi 2$	C_α-C_β-C_γ-C_δ*	CX-C8-C8-C8	
	$\chi 3$	C_β-C_γ-C_δ-C_ϵ*	C8-C8-C8-C8	
	$\chi 4$	C_γ-C_δ-C_ϵ-N_ζ*	C8-C8-C8-N3	
Met	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-2C	1
		C-C_α-C_β-C_γ	C-CX-2C-2C	
	$\chi 2$	C_α-C_β-C_γ-S_δ*	CX-2C-2C-S	
	$\chi 3$	C_β-C_γ-S_δ-C_ϵ*	2C-2C-S-CT	
Phe	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-CA	5
		C-C_α-C_β-C_γ	C-CX-2C-CA	
	$\chi 2$	C_α-C_β-C_γ-$C_{\delta 1}$*	CX-2C-CA-CA	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-2C-CA-CA	
Ser	$\chi 1$	N-C_α-C_β-O_γ*	N-CX-2C-OH	11
		C-C_α-C_β-O_γ	C-CX-2C-OH	
	oh	C_α-C_β-O_γ-H_γ*	CX-2C-OH-HO	
Thr	$\chi 1$	N-C_α-C_β-$O_{\gamma 1}$*	N-CX-3C-OH	
		C-C_α-C_β-$O_{\gamma 1}$	C-CX-3C-OH	
		N-C_α-C_β-$C_{\gamma 2}$	N-CX-3C-CT	2
		C-C_α-C_β-$C_{\gamma 2}$	C-CX-3C-CT	
	oh	C_α-C_β-$O_{\gamma 1}$-$H_{\gamma 1}$*	CX-3C-OH-HO	
		$C_{\gamma 2}$-C_β-$O_{\gamma 1}$-$H_{\gamma 1}$	CT-3C-OH-HO	
Trp	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-CT-C*	8

Table 2.14: Continued

Amino acid	Rotatable bond	Atom names	Atom types	Solving group
		C-C_α-C_β-C_γ	C-CX-CT-C*	
	$\chi 2$	C_α-C_β-C_γ-$C_{\delta 1}$*	CX-CT-C*-CW	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-CT-C*-CB	
Tyr	$\chi 1$	N-C_α-C_β-C_γ*	N-CX-2C-CA	5
		C-C_α-C_β-C_γ	C-CX-2C-CA	
	$\chi 2$	C_α-C_β-C_γ-$C_{\delta 1}$*	CX-2C-CA-CA	
		C_α-C_β-C_γ-$C_{\delta 2}$	CX-2C-CA-CA	
	oh	$C_{\epsilon 1}$-C_ζ-O_η-H_η*	CA-C-OH-HO	
		$C_{\epsilon 2}$-C_ζ-O_η-H_η	CA-C-OH-HO	
Val	$\chi 1$	N-C_α-C_β-$C_{\gamma 1}$*	N-CX-3C-CT	2
		N-C_α-C_β-$C_{\gamma 2}$	N-CX-3C-CT	
		C-C_α-C_β-$C_{\gamma 1}$	C-CX-3C-CT	
		C-C_α-C_β-$C_{\gamma 2}$	C-CX-3C-CT	

To reduce problem size and accelerate convergence, amino acids were separated into the solving groups listed in Table 2.15 based on shared dihedral atom types, and a separate objective function (Equation (2.5)) was constructed for each of the solving groups. Specific tables of dihedrals in each solving group are provided in Tables 2.2 to 2.14. Since each group shares no four-atom dihedrals with other groups, the full parameter space could be partitioned, with each solving group providing all conformations and energies necessary for separate optimization of each parameter subset. Optimized values of the objective function for each solving group are provided in Table 2.16.

Fitting details

Six populations of 63 individuals each were created: two with ff99SB parameters, two with zero parameters, and two with random parameters created with different random seeds. Each set of populations was then subjected to a series of evolutions using random seeds 314 159 and 271 828, carried out

Table 2.15: Dihedral solving groups noting the amino acids that were included in each combined fit due to shared dihedral parameters. The numbering for each group is arbitrary.

Group	Amino acids
1	Glutamate Protonated glutamic acid Glutamine Methionine
2	Isoleucine Threonine Valine
3	Protonated aspartic acid Asparagine
4	δ-protonated Histidine ϵ-protonated Histidine Doubly-protonated Histidine
5	Phenylalanine Tyrosine
6	Arginine Lysine
7	Cysteine
8	Cysteine dimer (disulfide)
9	Tryptophan
10	Aspartate
11	Leucine
12	Serine

using GAlib [Wall, 1996]. An elitist regime maintained the fittest tenth of the population from one generation to the next. Initially, each population evolved for 200 000 generations at a mutation rate of 0.01 and crossover rate of 0.8. Mutation frequency defined the probability with which a given parameter pair—amplitude and phase shift—will mutate. Upon mutation, the lowest bit of the random number was used to determine whether the amplitude or the phase shift will change. Perturbation to amplitude (*mutateBy*) depended upon mutation rate (*mutRate*) and the random number ($random \in [0, 1)$), by the relation:

$$mutateBy = \begin{cases} random/mutRate - 0.5, & random < mutRate \\ 0, & random \geq mutRate \end{cases} \tag{2.6}$$

This scheme alternated with a second, where all changes to amplitude were $0.001\,\mathrm{kcal\,mol^{-1}}$ in magnitude. If perturbing the phase shift, the lowest bit of the random number determined whether the phase shift would be 0 or 180 degrees.

To narrowly locate sets of parameters that minimize error after the first 200 000 generations, each population was continued with a mutation rate of 0.005 and crossover rate of 0.8 with the second scheme, then 0.002 and 0.8 with the first and second scheme, and finally 0.001 and 0.8 with the first and second scheme until convergence.

Convergence was evaluated as a run starting from ff99SB finding the same (as defined in the next sentence) steady (less than 0.001 kcal/mol/pair improvement in 10 000 generations) solution as a run starting from zero or random parameters. Solutions were considered the same if the correction energy profile scanned every 10° was identical within $0.01\,\mathrm{kcal\,mol^{-1}}$. In most cases, however, different runs achieved the same dihedral corrections, with no differences in amplitude to three decimal places.

Fitting χ1 N-Cα-Xβ-Xγ and C-Cα-Xβ-Xγ parameters required consideration of the sp3 hybridization of the α-carbon. Multiple three-fold dihedral

corrections cannot be partitioned between the N and the C, as they are rotated 120° relative to the Cα-Xβ bond; the same three-fold correction must be applied to both. In the AMBER12-bundled ff12SB, we did not account for this, and so amplitudes like the three-fold around χ1 were of arbitrary magnitude, with potentially undesirable effects on small peptides like Val_3 and loop regions. Other periodicity terms, however, were trained separately, such as one-fold corrections around χ1 describing whether the side chain γ substituent should be placed gauche to the N, the C, or both.

Where amino acids had planar moieties, however, there were multiple sets of 4-atom combinations with dihedrals offset by 180°. In these cases, it was necessary to choose a single set of atom types to apply corrections to, as a 180° offset means odd-periodicity terms will be out-of-phase, and have exactly opposite effects, while even-periodicity terms will be in-phase, and thus cannot be distinguished. In the case of amide and carboxylic groups sharing atom types, we only fit the terms correcting C-C-C-OH and C-C-C-N, as these are distinct between the two.

In the case of histidine, the various atom types of the different protonation states required us to fit multiple corrections. Histidine has two χ2 non-hydrogen dihedrals, to the Nδ and to the Cδ. Histidine Nδ1 and Cδ2 are atom types NA and CV in δ-protonated histidine, NB and CW in ϵ-protonated histidine, and NA and CW in ionic histidine. Since only δ-protonated histidine possessed a χ2 CV and only ϵ-protonated histidine possessed a χ2 NB, it seemed logical to fit one of these χ2 corrections independently, while fitting the remaining two protonation states with two sets of corrections. We tried both and chose fitting ϵ-protonated histidine separately as the combination that best fit the quantum data.

Backbone dependence analysis

Our goal is to use the AAEs to optimize a single set of parameters that minimizes the REE for multiple backbone conformations. However, the AAE for α and β are averages over all side chain pairs, while the ability of the optimization procedure to maximize transferability hinges on the backbone dependence

of the QM-MM energy error for pairs of side chain conformations. Greater similarity would indicate a better likelihood of being able to optimize a single set of parameters that is transferable among different backbone conformations. To quantify this, we subtracted the β REE from the α REE for each pair i, j of side chain conformations, averaging the magnitudes of these differences:

$$\mathrm{BBD} = \frac{2}{N(N-1)} \sum_{i} \sum_{j<i} |\mathrm{REE}(i,j)_{\alpha} - \mathrm{REE}(i,j)_{\beta}| \tag{2.7}$$

where the same notation is used as defined for Equation (2.4). We note that the BBD does not report on how well the QM and MM energies match, only on whether the differences between QM and MM energies are consistent as the backbone conformation changes. Thus BBD for each amino acid is a measure of the ultimate ability of side chain dihedral parameters to match QM data in the absence of explicit coupling between backbone and side chain parameters; the difference cannot be corrected with side chain dihedral parameters.

2.4.2 Test dynamics simulations

Initial structures

Helical conformations were defined as all $(\phi, \psi) = (-60°, -40°)$. Linear conformations were defined as all $(\phi, \psi) = (180°, 180°)$. Native conformations, as appropriate, were defined for each system as below. Explicit solvation was achieved with truncated octahedra of TIP3P water [Jorgensen et al., 1983] with a minimum 8 Å buffer between solute and the water box boundary. All structures were built via the LEaP module [Zhang et al., 2010] of Ambertools.

General details

Except where otherwise indicated, equilibration was performed with a weak-coupling (Berendsen) thermostat and barostat [Berendsen et al., 1984] targeting 1 bar pressure with isotropic position scaling as follows. With positional restraints on protein heavy atoms of 100 kcal mol^{-1} Å^{-2}, structures were minimized for up to 10 000 cycles and then heated at constant volume from

100 K to 300 K over 100 ps, followed by another 100 ps at 300 K. The pressure was equilibrated for 100 ps and then 250 ps with time constants of 100 fs and then 500 fs on coupling of pressure and temperature to 1 bar and 300 K, and 100 kcal mol^{-1} Å^{-2} and then 10 kcal mol^{-1} Å^{-2} positional restraints on protein heavy atoms. The system was again minimized, restraining only the protein main chain N, C_α, and C positionally with 10 kcal mol^{-1} Å^{-2} for up to 10 000 cycles. Three 100 ps simulations with temperature and pressure time constants of 500 fs were performed, with backbone restraints of 10 kcal mol^{-1} Å^{-2}, 1 kcal mol^{-1} Å^{-2}, and then 0.1 kcal mol^{-1} Å^{-2}. Finally, the system was simulated unrestrained with pressure and temperature time constants of 1 ps for 500 ps with a 2 fs time step, removing center-of-mass translation every ps.

SHAKE [Ryckaert et al., 1977] was performed on all bonds including hydrogen with the AMBER default tolerance of 10^{-5} Å for NpT/NVT and 10^{-6} Å for NVE. Non-bonded interactions were calculated directly up to 8 Å with cubic spline switching and the particle-mesh Ewald approximation [Darden et al., 1993] in explicit solvent, with direct sum tolerances of 10^{-5} for NpT/NVT or 10^{-6} for NVE. The timesteps for NpT/NVT and NVE simulations were 2 fs and 1 fs, respectively. Tighter convergence criteria and a shorter timestep facilitated the energy conservation required for NVE.

System-specific details

HBSP The HBSP sequence denoted 3a by Wang et al. [2006] (Ac-GQVA RQLAEIY-NH_2) was chosen, as it had the greatest measured helical content. HBSP has a covalently pre-organized α-turn, with the O of the first CO and the H of the NH of residue 5 substituted by carbons, with a covalent single bond between the substituted carbons. Modeling of this covalent modification was approximated by a harmonic distance restraint between the CO of the acetyl cap and the NH of A5 with force constant 100 kcal mol^{-1} Å^{-2}. This restraint was chosen as it well reproduced the distribution of hydrogen bond distances present in a crystal structure of aquaporin (PDB ID: 3ZOJ [Kosinska Eriksson et al., 2013]). HBSP was solvated with 2643 TIP3P water molecules and simulated for 1.6 µs in the NVT ensemble. Two independent runs were conducted

with all helical or semi-extended conformations. The semi-extended conformation was built with the first five residues helical in obeisance to the covalent bond in the experiment, with the remaining residues extended.

CLN025 As a model system to carry out initial tests of secondary structure balance, we turned to CLN025, an engineered fast-folding hairpin that is a thermally optimized variant of chignolin, with N- and C-terminal glycine-to-tyrosine substitutions. Thus the CLN025 sequence was YYDPETGTWY. The native conformation was chosen as the fifth conformation in the NMR ensemble [Honda et al., 2008], as that conformation was closest to the average of the NMR ensemble.

Proteins We simulated four folded proteins for comparison of dynamic properties against NMR. First was the third Igg-binding domain of protein G (GB3, sequence: MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVD GVWTYDDATKTFTVTE). The native structure was defined as a liquid crystal NMR structure (PDB ID: 1P7E [Ulmer et al., 2003]). Second was the bovine pancreatic trypsin inhibitor (BPTI, sequence: RPDFCLEPPYTGPCKARIIR YFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA). The native structure was defined as a joint neutron/X-ray diffraction structure (PDB ID: 5PTI [Wlodawer et al., 1984]). Third was ubiquitin (Ubq, sequence: MQIF VKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDG RTLSDYNIQKESTLHLVLRLRGG), with the native structure defined as a crystal structure (PDB ID: 1UBQ [Vijay-Kumar et al., 1987]). Fourth was hen egg white lysozyme (Lys, sequence: KVFGRCELAAAMKRHGLDNYRGYS LGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTP GSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKG TDVQAWIRGCRL), with the native structure defined as a crystal structure (PDB ID: 6LYT [Young et al., 1993]). Owing to their larger size, the proteins were equilibrated as above, but with the unrestrained step extended to a full nanosecond, rather than 500 ps. NVE was employed for the production simulations, so that artifacts from barostats or thermostats would not influence the dynamics.

Urea-denatured GB1 calculations

Experimental $\chi 1$ scalar couplings have been determined for urea-denatured GB1 were at pH 2.0 [Vajpai et al., 2010], at which glutamate is protonated. GB1 has a C-terminal glutamate, but protonated C-terminal glutamate was not in the standard AMBER libraries, it was built as follows. The charge differences from Glu to C-Glu and from Glu to Glh were summed and then added to the original Glu charges. The -0.014 charge needed to bring the charge to -1 was distributed across all atoms evenly. After rounding to 6 decimal places, there was a net charge of -1.000008; the 0.000004 was added to the charges on the backbone amide hydrogen and the side chain carboxyl hydrogen, bringing the charge to -1.

Initial structures were built fully extended with LEaP, without adding solvent. To allow LEaP fully extended structures to relax, GB1 was simulated with GB-Neck2 implicit solvent [Nguyen et al., 2013] and minimized for 100 000 cycles, heated from 100 to 300 K over 500 ps, and then simulated at 300 K, writing snapshots every 5 ps. This protocol was repeated for both ff99SB and ff14SB.

Then, the first 10 ns from simulations with each force field were clustered using the hierarchical-agglomerative algorithm in cpptraj [Roe and Cheatham, 2013] with a target of 10 clusters. The centroids from each cluster were then solvated with 8 molar urea, using protein parameters in conjunction with RESP charges and N-C-N bond angle vibrational parameters determined by Alexey Onufriev [Case et al., 2005]. First, they were each solvated with an 8 Å buffer to determine how many solvent molecules would be needed to cover each conformation. Then, the buffer size was increased as necessary so that all conformations had, to within 5 molecules, the same quantity of solvent. The solvated centroids of the four most populated clusters were then subject to the same protocol as the folded proteins, following the general equilibration and NVE production settings detailed above.

2.4.3 Analysis

Calculation of NMR observables

Scalar couplings were calculated from simulations using Karplus relations [Karplus, 1959, 1963]. Side chain scalar couplings were calculated using Ile, Thr, and Val C/N-Cγ Karplus parameters from Chou et al. [2003], and Perez et al. Karplus parameters [Perez et al., 2001] for all other χ1 scalar couplings.

Backbone NH Lipari-Szabo S^2 order parameters were calculated using the iRED method [Prompers and Brüschweiler, 2002] via cpptraj [Roe and Cheatham, 2013].

NOE reproduction in CLN025 was evaluated by computing r^{-6} for all interproton vectors for every trajectory frame, and comparing $\langle r^{-6} \rangle^{-\frac{1}{6}}$ for each vector with the NOE-based restraints published by Honda et al. [2008], downloaded via the BMRB [Ulrich et al., 2008]. For ambiguous restraints, contributions from each proton pair to the NOE were summed [Nilges, 1995]. For each force field we generated two ensembles, one combining structures from the 4 initially folded simulations and the other combining the 4 initially linear simulations. These were used to calculate NOE deviations, with the difference between ensembles from different initial structures used to quantify precision.

Calculation of helical content

Helical content was defined as the average fraction of residues, excluding the first two and last, in α-helix (H) or 3-10 helix (G) as defined by DSSP [Kabsch and Sander, 1983], as implemented in cpptraj [Roe and Cheatham, 2013].

Significance analysis

One way to incorporate statistical significance into analyzing force field differences is to plot the average difference in errors for all scalar couplings whose comparison satisfies any of a range of p-values. Those that satisfy lower p-values may be considered more informative than those that only satisfy higher p-values. The probability that distributions are not different was approximated

by Welch's t-test [Welch, 1947]. The p-value was calculated using the survival function of the SciPy [Travis, 2007, Jones et al., 2001-] stats module, based on t as in Equation (2.8) with μ_{FF} and SE_{FF} corresponding to the average normalized error and standard error of the mean normalized error, respectively, for each force field FF, and degrees of freedom approximated by the Welch-Satterthwaite equation [Welch, 1947, Satterthwaite, 1946], (Equation (2.9); $n\breve{}1$, being 3 for all force fields, enters the numerator).

$$t = \left(\mu_{\text{ff14SB}} - \mu_{\text{ff99SB-ILDN}}\right)\left(SE^2_{\text{ff14SB}} - SE^2_{\text{ff99SB-ILDN}}\right)^{-\frac{1}{2}} \tag{2.8}$$

$$d.f. = \frac{3\left(SE^2_{\text{ff14SB}} + SE^2_{\text{ff99SB-ILDN}}\right)^2)}{SE^4_{\text{ff14SB}} + SE^4_{\text{ff99SB-ILDN}}} \tag{2.9}$$

2.5 Fitting Results

2.5.1 Side chain rotamer energies improved match to QM data and better transferability between backbone conformations

An important question is how to define $E_{\text{QM},i}$ and $E_{\text{MM},i}$ used for calculating REE (Equation (2.1)). As discussed above, restraints could be applied to dihedrals other than the specific 4-atom set defining the ϕ, ψ, and χ rotatable bonds. We tested several choices, including restraining only the 4-atom sets defining ϕ, ψ, and χ, as well as restraining all possible 4-atom dihedrals, or restraining all dihedrals in the backbone but only the defining dihedrals in the side chains (see Table 2.14 for dihedral classifications). We also tested the impact of MM re-optimization of QM geometries. As discussed above, these choices in the generation and comparison of structures can introduce artifacts in the energy profiles that hamper parameter optimization and weaken transferability. We evaluated the impact of these choices by calculating the intrinsic BBD as well as the AAE for various restraint and structure optimization options, using ff99SB as a reasonable MM model.

Ideally, one should start with relatively close agreement between ff99SB

and MP2/6-31+G**//HF/6-31G* energies (low AAE), with similar errors between different backbone conformations (low BBD), if dihedral parameters are to reconcile errors without explicit coupling to the backbone conformation. According to the penultimate rotamer library [Lovell et al., 2000], the side chains of aspartate and asparagine depend most on backbone conformation; we thus chose them for initial testing of how the energy calculations impact coupling between side chain and backbone dihedral parameters.

Restraining all backbone dihedrals and re-optimizing the QM structure with MM before calculating energy yielded both the lowest AAE (2.55 ± 0.09 kcal mol^{-1} for Asp and 1.98 ± 0.01 kcal mol^{-1} for Asn, error bars reflect difference between α and β backbone context) and lowest BBD (1.35 ± 0.01 kcal mol^{-1} for Asp and 1.42 ± 0.03 kcal mol^{-1} for Asn, error bars reflect difference between two staggered halves of structures), as shown in Figure 2.2. At the opposite extreme, restraining just ϕ and ψ and using the QM structures to calculate MM energies resulted in the greatest AAE (3.45 ± 0.13 kcal mol^{-1} for Asp and 3.09 ± 0.74 kcal mol^{-1} for Asn) and BBD (2.23 ± 0.02 kcal mol^{-1} for Asp and 3.90 ± 0.08 kcal mol^{-1} for Asn). Fundamental differences in the modeling of bonded and non-bonded interactions between QM and MM are likely exacerbated when the QM-optimized structures are evaluated in MM without re-optimization; these differences manifest as larger errors in MM energy, as well as less transferability between backbone contexts. Restraining all possible combinations of four-atoms describing each dihedral in both the backbone and side chain was also attempted (this approach led to the ff12SB parameter set, see Appendix B), but restraining all 4-atom dihedrals in the side chains prevented relaxation of angle terms in MM optimization, leading to increased error that likely would not be present in a simulation when steric clashes can be alleviated through adjustment of covalent structure. Based on these results, we made the decision to restrain all backbone dihedrals during structure optimization, but only the defining dihedrals for side chains, and to re-optimize the QM structures with the MM model (using the same restraints) prior to calculation of MM energies. These energies were used in the calculation of the objective function in Equation (2.5).

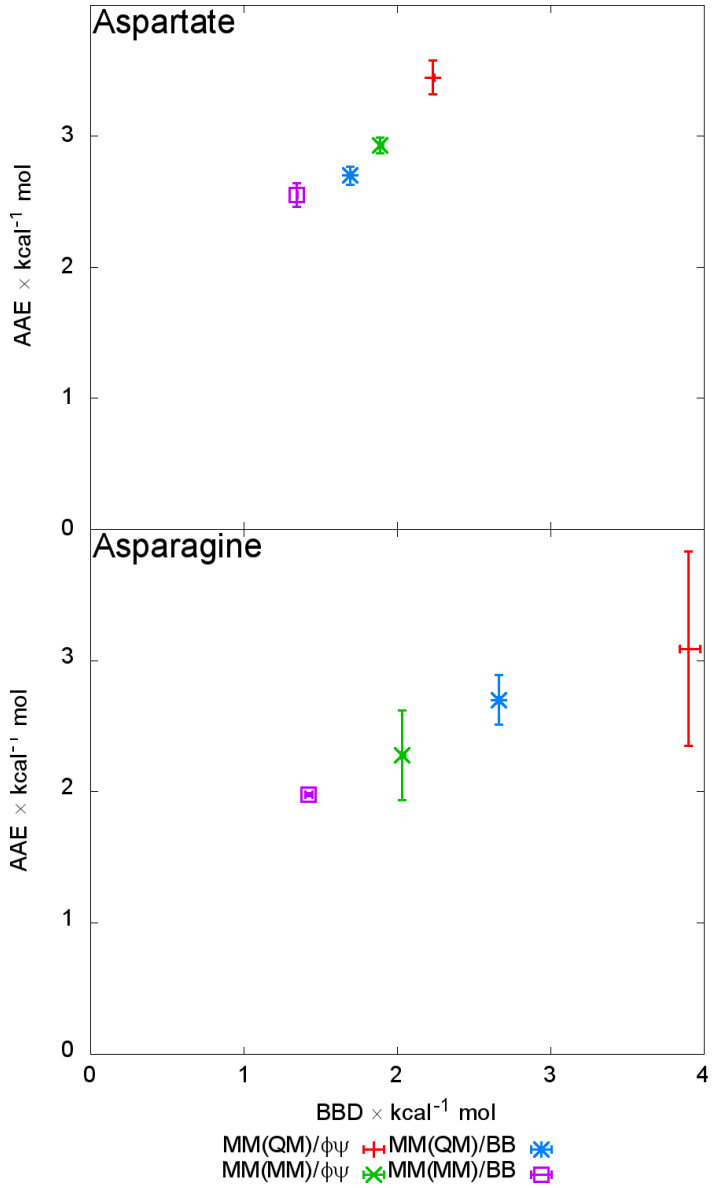

Figure 2.2: AAE vs. BBD of aspartate and asparagine, calculating MM energies of QM structures and restraining: ϕ and ψ (red crosses) or all backbone dihedrals (blue stars); or calculating MM energies of MM re-optimized structures and restraining: ϕ and ψ (green 'X's) or all backbone dihedrals (purple squares). The latter provides the lowest AAE values, as well as the best intrinsic transferability between backbone conformations. Error bars in AAE indicate difference between AAE calculated in the α or β backbone context, and BBD using half of the structures.

As discussed in Methods, each solving group was optimized separately (objective function $\mathscr{O}$ values for each solving group are provided in Table 2.16), but here we average the individual objective function $\mathscr{O}$ values, weighted by the number of amino acid-backbone combinations contributing to each, to facilitate their comparison between different parameter sets. The resulting $\mathscr{O}$ values quantify the magnitude of error in energy differences for conformation pairs, averaged over all amino acids and backbone conformations. In ff99SB, $\mathscr{O}$ was $1.52\,\mathrm{kcal\,mol^{-1}}$, while $\mathscr{O}$ for the final optimization parameter set was $0.98\,\mathrm{kcal\,mol^{-1}}$. This 35% improvement is decomposed by residue and by backbone conformation in Figure 2.3, and the distribution of all pair energy errors (REEs) is presented in Table 2.1. All of the amino acids with errors larger than $2\,\mathrm{kcal\,mol^{-1}}$ in ff99SB (tyrosine and protonated or deprotonated aspartic acid) were significantly improved with the new parameters. In addition to improvement for the ILDN residues previously addressed by Lindorff-Larsen et al. [2010], we observed better agreement with the QM training data for every residue compared to ff99SB. The only profile that didn't improve was α-backbone Phe, in which the initial ff99SB error was close to the average final AAE for all residues, limiting the potential for improvement. It is remarkable to see that the optimization procedure was able to find a solution that simultaneously improved performance for all amino acids, and with little resulting backbone dependence. We refer to the combination of ff99SB with new side chain dihedral parameters as ff14SBonlysc; adding a set of updated backbone parameters [Maier et al., 2015, Martinez, 2014] will result in the ff14SB model.

Although the ff14SBonlysc parameters show improved reproduction of the QM data, several caveats apply. First, the performance in Figure 2.3 measures the ability of the parameters to reproduce energies for structures that were used in the training shown above, but not training data used for the other force fields, thus better performance on the training data is expected. Second, closely reproducing gas-phase QM data does not guarantee reliable simulation properties [MacKerell et al., 2004b]. As discussed above, it is possible that training against gas-phase QM data might counteract some of the influence of the "pre-polarized" partial charges in our model, potentially worsening

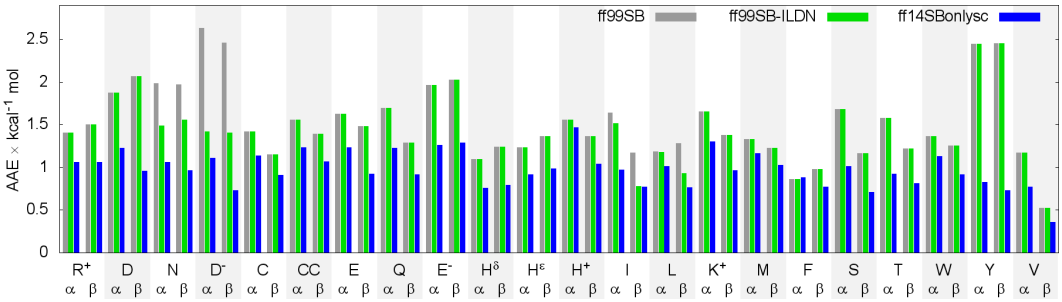

Figure 2.3: The AAE of each force field for each amino acid (single letter codes), with data for both α and β backbone conformations. Ionized residues are indicated by charge superscripts. CC indicates the disulfide bridge. Data are shown for ff99SB, ff99SB-ILDN, and ff99SB with the new side chain corrections (ff14SBonlysc).

performance for simulations in aqueous solution. Thus we followed the training against QM data with more rigorous testing in solution simulations, with comparison to experiments also in solution.

2.6 Testing Strategy

Key fitting assumptions raise important questions, like whether the diversity and designed backbone-independence of our side chain training set will improve side chain rotamer preferences for proteins in solution, despite training against in vacuo dipeptide energies at a modest level of QM theory. To investigate accuracy of side chain rotamer sampling, we compared against $\chi 1$ scalar couplings for a set of folded proteins including GB3, ubiquitin, lysozyme, and BPTI (collated by Lindorff-Larsen et al. [Lindorff-Larsen et al., 2010, Berndt et al., 1992, Grimshaw, 1999, Hu and Bax, 1997, Chou et al., 2003, Miclet et al., 2005, Schwalbe et al., 2001, Smith et al., 1991]). Importantly, we considered the performance of the new model relative to ff99SB [Hornak et al., 2006] and ff99SB-ILDN [Lindorff-Larsen et al., 2010] in different secondary structures, to evaluate whether the design of average coupling into our side chain parameters is expressed by improved transferability between different backbone conformations in proteins. We also tested the benefit of re-optimizing parameters for side chains other than ILDN (Ile, Leu, Asp, and Asn).

Table 2.16: Objective values O for each of the solving groups

Solving group	Amino acids	O_{ff99SB}	O_{ff14SB}
0	Cyx	1.5	1.2
1	Glh Gln Glu Met	1.6	1.1
2	Ile Thr Val	1.2	0.8
3	Ash Asn	2.0	1.1
4	Hid Hie Hip	1.3	1.0
5	Phe Tyr	1.7	0.8
6	Arg Lys	1.5	1.1
7	Cys	1.3	1.0
8	Trp	1.3	1.0
9	Asp	2.5	0.9
10	Leu	1.2	0.9
11	Ser	1.4	0.9

These side chain parameter evaluations are subject to numerous caveats as discussed in Chapter 1. For example, the Karplus curve assumes that a scalar coupling can be calculated from a single dihedral angle [Karplus, 1959, 1963], although the spin-spin coupling may be sensitive to the spin lattice structure. One important limitation is that many reported scalar couplings are outside the range of relevant Karplus curves that might be used to compare simulated ensembles to experiment. Thus reproduction of the experimental observations would be impossible regardless of the ensemble of conformations sampled in simulation. In these cases, we adjusted the target value by adopting the value on the Karplus curve lying closest to the experimental value; otherwise, the experimental value was used as the target:

$$ {}^3J^*_{i,\text{NMR}} = \begin{cases} \min(J_{i,\text{Karplus}}), & J_{i,\text{NMR}} < \min(J_{i,\text{Karplus}}) \\ \max(J_{i,\text{Karplus}}), & J_{i,\text{NMR}} > \max(J_{i,\text{Karplus}}) \\ J_{i,\text{NMR}}, & \text{otherwise} \end{cases} \tag{2.10} $$

Additionally, because H-H scalar couplings reporting on some residues have a much larger range than C-C scalar couplings reporting on others, deviations were normalized by Karplus curve range. The errors are summarized in terms

of the average normalized error,

$$\mathrm{ANE} = \frac{1}{N}\sum_{i}^{N} \frac{\left|\langle J_i \rangle_{\mathrm{sim}} - {}^{3}J^{*}_{i,\mathrm{NMR}}\right|}{\max(J_{i,\mathrm{Karplus}}) - \min(J_{i,\mathrm{Karplus}})} \tag{2.11}$$

The resulting metric is more intuitive than average error, as 0 indicates best possible agreement, whereas 1 indicates maximum deviation.

In the peptides and proteins tested here, backbone and side chain dihedrals are coupled to each other within and between residues, making it difficult to determine exactly why a particular scalar coupling may disagree with experiment (assuming the error is not because of the experimental measurement or the Karplus curve). Likewise, this hinders ascribing credit for improvement to any specific backbone or side chain update. To help aid in the decomposition, we tested $\chi 1$ scalar couplings with just side chain modifications and then introduced backbone updates, to help isolate the effects of intended and secondary changes. On the other hand, this dihedral coupling can mean that $\chi 1$ scalar couplings implicitly report on backbone, $\chi 2$ or $\chi 3$ torsions; thus reproducing $\chi 1$ data may suggest reasonable accuracy in other parameters as well.

Side chains are coupled to the backbone, but we do not expect folded proteins to globally unfold on the 100 ns timescale considered here for proteins due to changes in side chain parameters. The backbone parameters studied by Martinez [2014] increased the helical content of a small peptide, hydrogen bond surrogate peptide. This peptide was designed with a covalent modification that emulates a persistent helical hydrogen bond between the N-terminus and the fifth residue. Thus, this system can extend the helix especially quickly from its pre-nucleated core, despite having only ten residues. But the backbone modifications brought the helical content from 0.12 ± 0.01 with ff99SB to 0.26 ± 0.01 with ff14SB, compared to the experimental 0.46 [Wang et al., 2006]. As backbone dynamics may depend on side chain conformation, we therefore tested the effect of introducing the new side chain parameters on hydrogen bond surrogate peptide helicity.

As the ff14SB [Maier et al., 2015] backbone changes increased helical content relative to ff99SB [Hornak et al., 2006], we wanted to ensure that ff14SB did not compromise β-stability, but maintained secondary structure balance.

Thus, we compared the folding dynamics of a small β-hairpin between ff99SB and ff14SB.

Finally, we verified that the combined backbone and side chain updates, ff14SB, maintained the reasonably accurate protein order parameter reproduction reported previously for ff99SB [Hornak et al., 2006]. We calculated backbone NH order parameters for the same simulations used to analyze $\chi 1$ scalar couplings.

2.7 Testing Results

2.7.1 Agreement with side chain NMR scalar couplings is improved with ff14SB

We evaluated side chain dihedral parameter changes by comparing to three-bond scalar couplings that report on dihedral dynamics. In this evaluation, we simulated GB3, ubiquitin, lysozyme, and bovine pancreatic trypsin inhibitor (BPTI) to compare against scalar couplings aggregated by Lindorff-Larsen et al. [Lindorff-Larsen et al., 2010, Berndt et al., 1992, Grimshaw, 1999, Hu and Bax, 1997, Chou et al., 2003, Miclet et al., 2005, Schwalbe et al., 2001, Smith et al., 1991], and simulated GB1 in urea to compare against $\chi 1$ scalar couplings [Vajpai et al., 2010] in an unfolded context.

We tested ff99SB and ff99SB-ILDN as references, ff14SB which includes the backbone [Maier et al., 2015, Martinez, 2014] and side chain parameter updates described above, and also ff14SBonlysc, which includes the side chain updates described above while retaining the ff99SB ϕ and ψ parameters. This allows us to partially deconvolute the influence of improvements to the side chain and backbone. Simulations of each protein were carried out using each force field, and the ANE (Equation (2.11)) was calculated for each amino acid where experimental data is available (Figure 2.4). The average error was 0.160 ± 0.004 with ff99SB, 0.129 ± 0.003 with ff99SB-ILDN, 0.127 ± 0.003 with ff14SBonlysc, and 0.129 ± 0.003 with ff14SB. The average figures were within statistical uncertainty for ff99SB-ILDN and ff14SB, both of which show measurable improvement over ff99SB. Not surprisingly for these stably folded

proteins, there is little difference between ff14SB and ff14SBonlysc, suggesting that the improvement from ff99SB observed in this test is largely due to side chain parameter updates.

All of the variants significantly improved upon ff99SB in average, however the specific improvements of each force field differed. For example, the errors obtained using ff14SB (ff99SB-ILDN values given in parentheses after ff14SB values) in isoleucine, leucine, aspartate, and asparagine—the four residues modified by ff99SB-ILDN—were 0.11±0.01 (0.091±0.005), 0.16±0.02 (0.13±0.01), 0.111±0.009 (0.16±0.02), and 0.12±0.02 (0.154±0.009), respectively—slightly improved in 2 cases, and slightly worsened in 2 others.

As discussed above, ff99SB-ILDN was fit using β backbone conformations, while our fitting procedure was designed to improve side chain energetics for multiple backbone conformations. We investigated whether explicit inclusion of dipeptide α backbone conformations for QM calculations in the gas phase was successfully transferred to improvement in scalar couplings of helical residues in larger proteins. We analyzed residues refit by both ff99SB-ILDN and ff14SB that matched the following criteria: in a helix, solvent-exposed and therefore likely to represent the intrinsic preferences of the amino acid, and experimentally characterized by χ1 scalar couplings. Only three residues fit these criteria, N35 of GB3, D32 of ubiquitin, and N97 of lysozyme. Of the three, all are significantly better reproduced with ff14SB than ff99SB-ILDN, with ANEs for N35, D32 and N93 of $0.11 \pm 0.03/0.22 \pm 0.09$ (ff14SB/ ff99SB-ILDN), $0.15 \pm 0.04/0.47 \pm 0.02$, and $0.16 \pm 0.02/0.31 \pm 0.04$, respectively.

As ubiquitin D32 was the most statistically different between the two force fields, we attempted to decompose its simulation accuracy according to fitting method, potentially providing insight to guide future parameter optimization efforts. First, we used ff14SB aspartate side chain parameters in ff99SB-ILDN to ensure that aspartate side chain parameters are responsible for observed differences. This significantly reduced the D32 ANE to a value comparable to ff14SB (0.077 ± 0.006), confirming direct influence of the Asp parameters on D32 dynamics. The most obvious methodological difference that could explain this phenomenon is the inclusion of helical backbone structures in training.

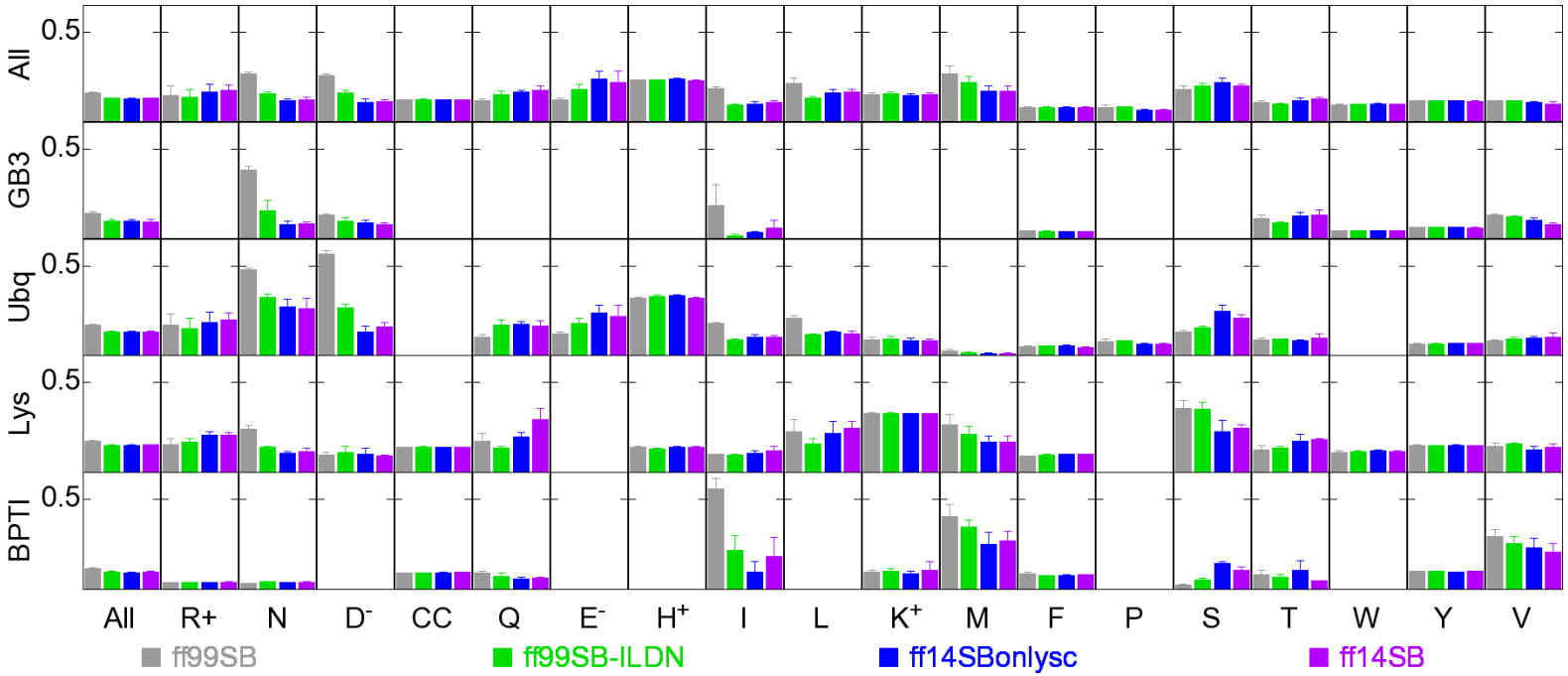

Figure 2.4: Average normalized errors (ANE) in side chain scalar couplings for each and for all amino acids in GB3, ubiquitin (Ubq), lysozyme (Lys), bovine pancreatic trypsin inhibitor (BPTI), and all proteins, according to ff99SB, ff99SB-ILDN, ff14SBonlysc, and ff14SB. Amino acids are shown with single letter code, with charge state noted for ionizable side chains. Error bars are calculated from four independent simulations.

We repeated the optimization for solving group 10 (containing just aspartate), but using only β backbone conformations in Equation (2.5). Simulations using these parameters still resulted in rather low ANE of 0.070 ± 0.008, suggesting that this particular improvement was not due to inclusion of fitting for both backbone conformations. A second possibility is that our fitting protocols had several differences compared to ff99SB-ILDN, such as weighting of squared QM-MM energy differences by QM energy in the ff99SB-ILDN fitting, which could introduce bias if positions of side chain rotamer minima are coupled to backbone conformation. Another protocol difference is that ff14SB parameters used two 4-atom dihedrals to describe χ1, while ff99SB-ILDN fit only one set. We refit parameters for solving group 10 using the ff14SB protocol but with the aspartate QM and MM energies published by Lindorff-Larsen et al. [2010]. With the resulting parameters, sampling of D32 was not improved compared to ff99SB-ILDN (ANE = 0.42 ± 0.07), suggesting that the D32 performance is related to differences in the QM benchmarks used to train the two force fields rather than the optimization protocol. Although both data sets used a 2D scan of χ1 and χ2, the resulting energies are influenced by the level of QM theory and the restraints used to generate potential energy surfaces for fitting. To test the influence of restraints, we used the potential energy surfaces that we generated with different structure optimization methods to test backbone-dependence (Figure 3). We therefore retrained solving group 10 parameters based on potential energy scans of only β, or α and β conformations, using the combinations of restraints and QM or MM optimized structures as carried out for Figure 2.2. The ANE of D32, and of all aspartates in ubiquitin, were plotted against the BBD of each method in Figure 2.5. For the parameters trained using only β dipeptides, the BBD of each method forwardly predicts the error of D32 obtained using those parameters. This suggests that how the structures in a potential energy surface are generated can significantly alter their transferability. In fact, the combination matching ff99SB-ILDN (restraining only ϕ and ψ, using QM structures for MM energies) also provided comparable results to ff99SB-ILDN (ANE of 0.32 ± 0.11), suggesting that the deviation in MD from solvated protein NMR data can be traced to the restraint method used during parameter development. The largest errors arise

when using QM structures for MM energies with fewer restraints; these errors are reduced when both α and β dipeptides are used in fitting, perhaps because artifacts from backbone interactions are lessened when requiring that the parameters work in multiple backbone contexts (for example, D32 ANE is reduced from 0.32 ± 0.11 with β-only training to 0.067 ± 0.007 using both α and β). Although we examined only one location in detail (D32), the same trend holds when considering all Asp residues in ubiquitin (Figure 2.5). These results on agreement between MD and NMR also mirror the findings in Figure 2.2, where the ability to reproduce QM energies (AAE) showed a similar dependence on BBD for different restraint methods. Taken together, the results show not only the sensitivity of the model to restraint method, but also reinforce that improved reproduction of gas-phase QM dipeptide energies leads to a better match to NMR data for proteins simulations in water.

If we expand the backbone-dependent comparison to consider all ILDN residues within helices versus all those without, we observe the same trends of ff14SB backbone-independence. The average errors of all I, L, D, and N residues in helices were 0.18 ± 0.01 with ff99SB-ILDN and 0.13 ± 0.01 with ff14SB. On the other hand, the average errors of all ILD and N residues not in helices were 0.11 ± 0.01 with both force fields. This indicates that overall errors for side chains in helical context are improved with ff14SB relative to ff99SB-ILDN, and that in ff14SB these errors are similar in magnitude to the non-helical side chain errors for both force fields. It seems reasonable to conclude that this improved transferability in ff14SB arises directly from the training of ff14SB against more transferable energy targets than other options tested, with multiple backbone conformations.

Overall, the results suggest that more careful consideration of these issues should be a factor in future force field efforts, as these measures can impact performance of simulations using the resulting parameters. These choices include how finely geometric changes outside the scan region are controlled, what level of variation in this geometry is desirable between QM and MM energy evaluations, and how these decisions are affected by intentional inclusion of diversity in neighboring regions.

The performance of individual amino acids with ff14SB is not always better

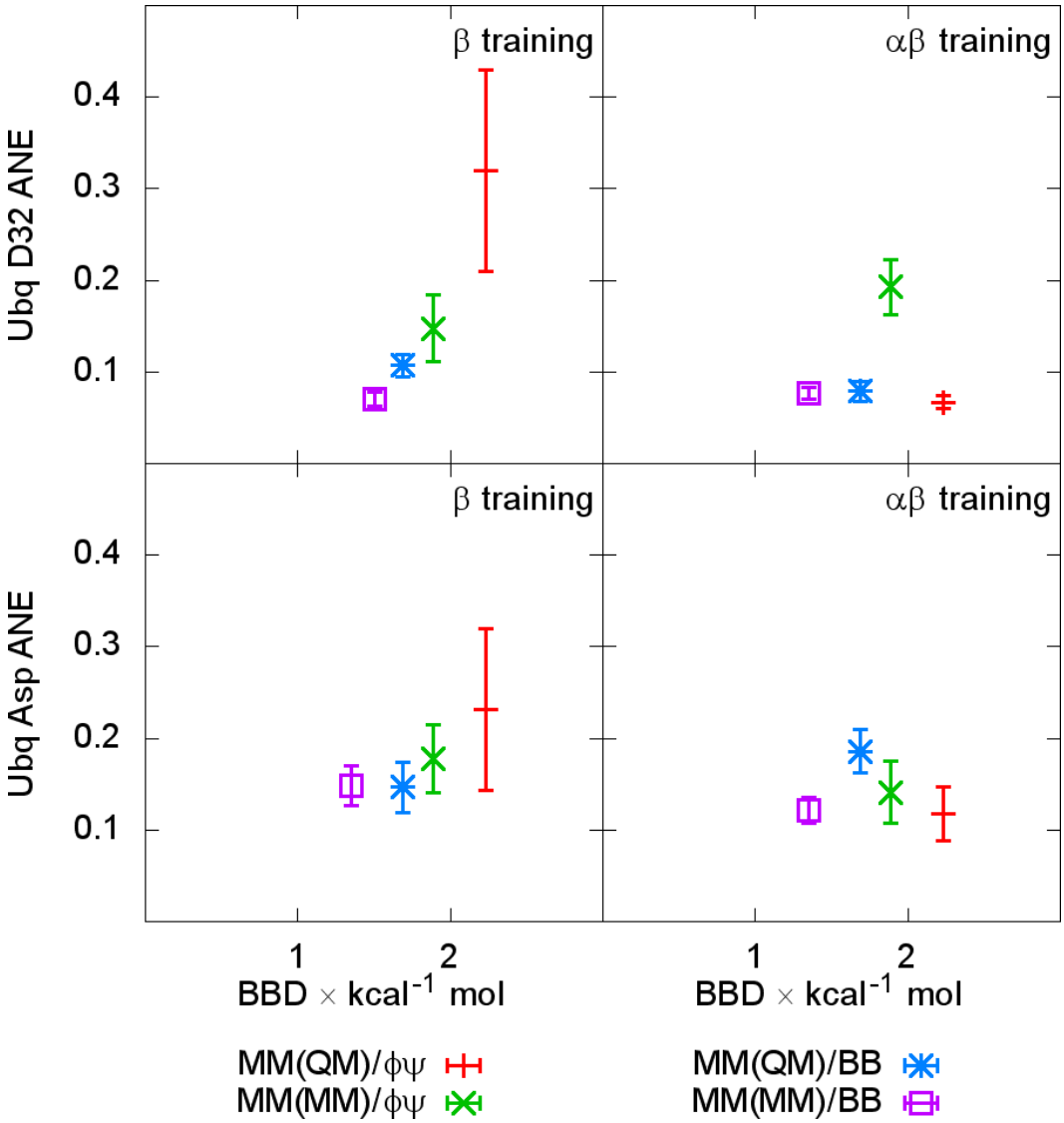

Figure 2.5: Simulated average normalized error (ANE) of ubiquitin (Ubq) D32 and all Ubq aspartate (Asp), with parameters developed from β or α and β conformations of aspartate dipeptides with restraints on all backbone dihedrals (BB) or only ϕ and ψ (ϕψ), and with molecular mechanics energies calculated for molecular mechanics structures (MM(MM)), or quantum mechanics structures (MM(QM)), versus backbone dependence on the x-axis. Parameters from β dipeptides with less backbone-dependent errors against quantum mechanics also exhibit lower errors against helical D32 scalar couplings. Training with α and β conformations performs comparably or, as in MM(QM)/ϕψ, better against scalar couplings.

than with other force fields, however. Lysozyme Q41 had a greater ANE with ff14SB (0.29 ± 0.06) than ff99SB (0.17 ± 0.02) or ff99SB-ILDN (0.13 ± 0.01). With ff14SBonlysc, Q41 had an ANE of 0.20 ± 0.02. With ff14SBonlyILDN (ff14SB but without side chain updates for amino acids other than I, L, D and N), the Q41 ANE was 0.12 ± 0.03. Looking at individual errors influencing the ANE, the experimental Q41 $H_\alpha H_{\beta 2}$ scalar coupling average values were too low with all force fields. Compared to the experimental value of $10.0\,s^{-1}$ (suggesting predominantly g^- $\chi 1$), ff99SB-ILDN, ff99SB and ff14SBonlysc were very similar at 9.2 ± 0.2 Hz, 8.8 ± 0.5 Hz, and $8.7 \pm 0.3\,s^{-1}$, respectively, while ff14SB was even lower at 7.3 ± 0.5 Hz due to additional sampling of trans $\chi 1$. Oversampling trans with ff14SB also resulted in $H_\alpha H_{\beta 3}$ scalar couplings of 5.4 ± 0.9 Hz, compared to experimental 1.2 Hz. Replacing the ff14SB Gln parameters with those from ff99SB had little effect on Q41, (7.7 ± 0.6 Hz for $H_\alpha H_{\beta 2}$ and 5.6 ± 0.5 Hz for $H_\alpha H_{\beta 3}$, with an ANE of 0.29 ± 0.06). This is not surprising, since Q41 ANEs were comparable between ff99SB and ff14SBonlysc. While this suggests that the error is related to backbone parameter changes, attempts to decompose dynamics are hampered by the fact that Q41 ANEs are not statistically worse for ff14SB than for ff99SB.

In fact, only 19% of all normalized errors were statistically different between ff99SB-ILDN and ff14SB at a significance of $p < 0.05$. As differences of varying statistical significance may be worth consideration, we determined the significance of each difference between ff99SB-ILDN and ff14SB by performing t-tests. As some of the insignificant differences may be small in magnitude, we calculated the average differences in normalized errors satisfying a range of p values from 0 to 1 (Figure 2.6). With the exception of the second most significant difference in BPTI, ff14SB was more accurate than ff99SB-ILDN for all of the most significant differences (up to $p < 0.1$).

It is also of interest to evaluate side chain dynamics in an unfolded protein. There are experimental scalar couplings describing GB1 and ubiquitin denatured in 8 M urea [Vajpai et al., 2010]. These experiments were carried out at low pH (2.5 and 2.0 for ubiquitin and GB1, respectively), allowing us to test the parameters for the protonated side chains of amino acids Asp and Glu. As the computational cost of simulating unfolded proteins is quite high, we

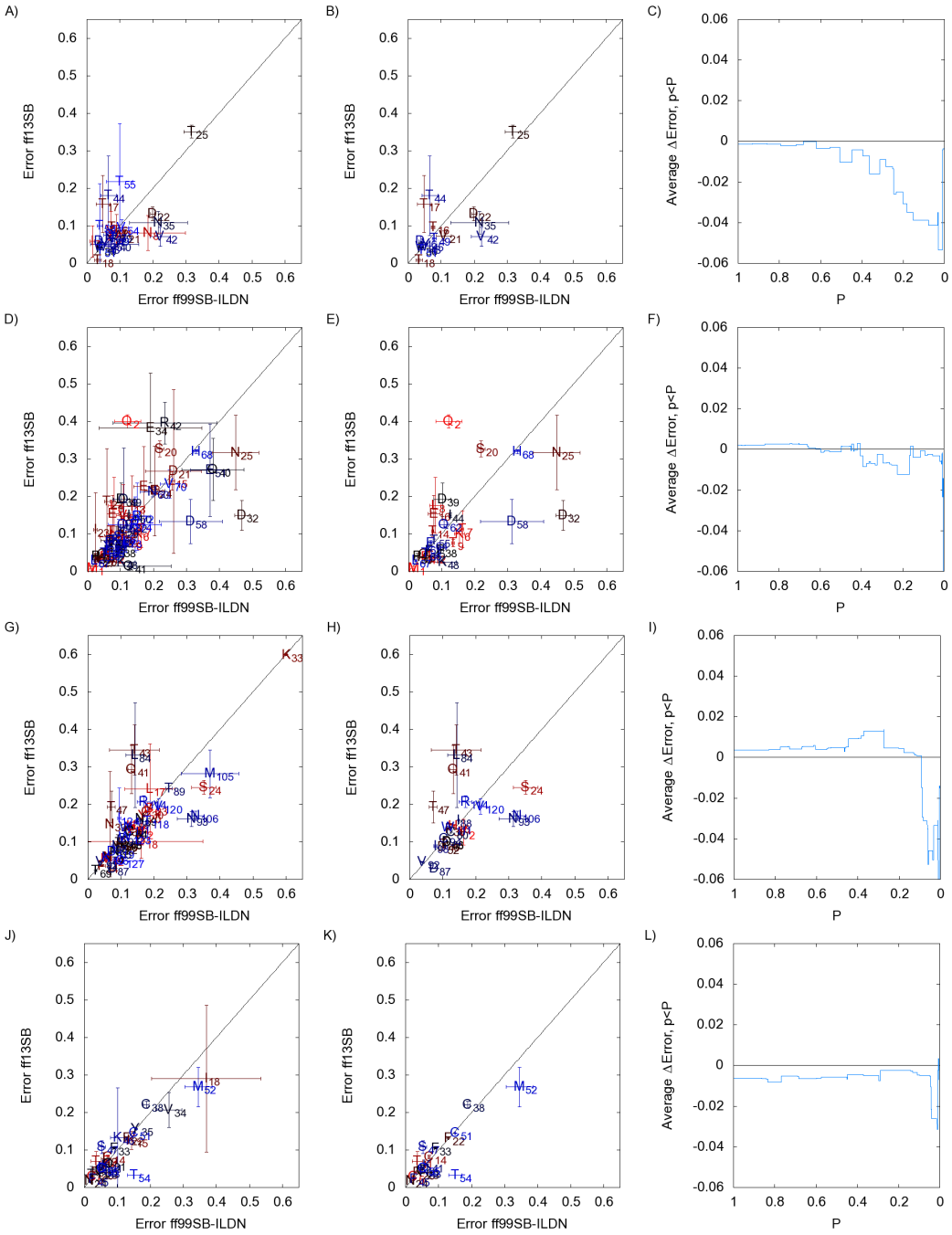

Figure 2.6: All normalized errors according to ff99SB-ILDN [Lindorff-Larsen et al., 2010] (x-axis) and ff14SB (y-axis), a subset of the errors where the uncertainties do not cross the equivalence line, and the average difference in normalized error from ff99SB-ILDN to ff14SB (y-axis) for normalized errors with significance of p<P (x-axis) for GB3 (A-C), ubiquitin (D-F), lysozyme (G-I), and BPTI (J-L).

limited this analysis to GB1, with 56 residues versus ubiquitin's 76 residues. Initial structures were generated using an implicit solvent [Nguyen et al., 2013] to collapse initially fully linear conformations built using LEaP [Zhang et al., 2010], performing cluster analysis on this collapse trajectory. The cluster centroids were then solvated with the same number of solvent molecules that was required for an 8 Å buffer for all of them. The centroids for the four most populous clusters were then simulated, with all the details provided in Section 2.4.

The ANEs for GB1 in urea (Figure 2.7) show similar improvements for ff14SB and ff99SB-ILDN in isoleucine, leucine, and asparagine. Several other residues improved with ff14SB. Most notably, ff14SB includes parameters for protonated aspartate, which was an outlier with ff99SB and ff99SB-ILDN, which did not update protonated aspartate. Additionally, improvements are noted for tyrosine and valine. In ff99SB-ILDN and ff14SB variants, tryptophan agreement got slightly worse, but the difference from ff99SB is small considering uncertainties. Overall, ff14SB best reproduced side chain dynamics in the folded and unfolded proteins tested here.

2.7.2 Helical stability is improved with ff14SB backbone changes and further improved with updated side chain parameters

Before a single backbone correction was chosen for ff14SB [Maier et al., 2015], Carmenza Martinez and Koushik Kasavajhala tested whether any of several candidate ff14SB backbone parameters that improved reproduction of Ala_5 scalar couplings also addressed ff99SB's helical limitation by examining a ten-residue hydrogen bond surrogate peptide (HBSP), where the geometry of the first helical hydrogen bond is enforced by a covalent bond between residues 1 and 5 [Chapman et al., 2004, Wang et al., 2006, Patgiri et al., 2008]. As mentioned, this alleviates the entropic cost of aligning the first four residues in a helix, allowing rapid testing of helical propagation according to each force field. Wang et al. reported $\sim 46\%$ helical content in PBS [Wang et al., 2006], but due to the potential for aggregation in that experiment, we followed the suggestion [Arora, 2015] of the authors and used the value of 70.13% helical

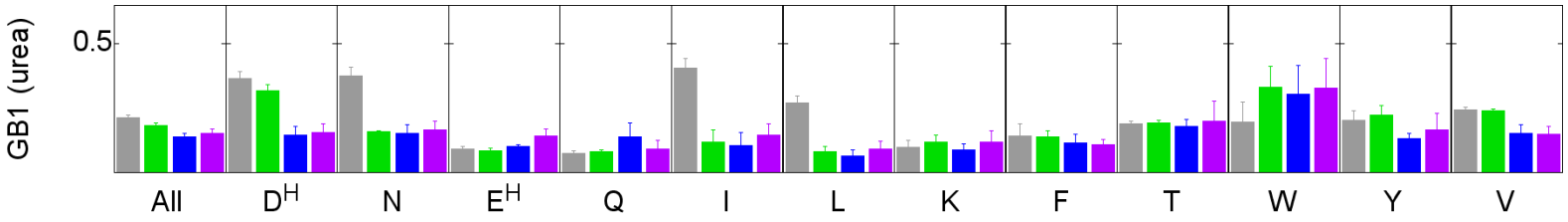

Figure 2.7: Average normalized errors (ANE) in side chain scalar couplings for each and for all amino acids in urea-denatured GB1 according to ff99SB, ff99SB-ILDN, ff14SBonlysc, and ff14SB. Amino acids are shown with single letter codes, with the charge state noted for ionizable side chains as a superscript. Error bars are calculated from four independent simulations.

content in 10% TFE, adjusted downward by $\sim$ 5–10% to obtain an estimate in water of $\sim$ 65%.

Simulations with ff99SB, which reproduced Ala_5 scalar couplings better than any of the potential ff14SB corrections when using DFT1 Karplus parameters, exhibited 0.17 ± 0.01 fraction helix. The mod1ϕ correction that performed best with Orig parameters, with the lowest χ^2 overall, increased this number to 0.51 ± 0.01. Further adding the mod2ψ correction to create mod1ϕ2ψ—the second lowest χ^2 against Ala_5 scalar couplings using the Orig Karplus parameters—yielded 0.72 ± 0.01 helical content, slightly higher than 0.65, but quite close. Meanwhile, mod3ϕ, which most closely reproduced Ala_5 scalar couplings with the DFT2 Karplus parameters [Case et al., 2000], only achieved 0.26 ± 0.01 helical content.

Adding the new side chain parameters increased the helical content in each case: ff99SB from 0.17 ± 0.01 to 0.26 ± 0.01, mod1ϕ from 0.51 ± 0.01 to 0.60 ± 0.01, mod1ϕ2ψ from 0.72 ± 0.01 to 0.79 ± 0.01, and mod3ϕ from 0.26 ± 0.01 to 0.46 ± 0.01. Whereas using mod1ϕ2ψ with the side chain corrections produced too much helix, mod1ϕ with the side chain corrections produced very close to the experimental target of 0.65. This increase in helical content by the side chain parameters suggests that a correction of only the backbone parameters to quantitatively address a system with side chains like HBSP, without considering side chain errors, may have led to an overcorrection in the backbone to cancel errors in the side chains. In that case, the transferability of such a setup to other amino acids might be challenging. Instead, the new side chain parameters improved the experimental agreement of HBSP simulations using mod1ϕ, which became the ff14SB backbone parameters. This increase of stability with the side chain parameters suggests that the side chain parameters have wanted effects on secondary structure.

2.7.3 Testing hairpin stability and structure

It is possible that the improvement in helical content was obtained at the cost of less accurate performance on β systems, whereas the side chain modifications, as for HBSP, may also affect stability. As a model system to carry

out initial tests of secondary structure balance, we turned to CLN025 [Honda et al., 2008, Davis et al., 2012], an engineered fast folding hairpin that is a thermally optimized variant of Chignolin. CLN025 contains N- and C-terminal glycine-to-tyrosine substitutions from Chignolin, which already possesses one tyrosine and one tryptophan. The presence of four aromatic side chains in a short peptide suggests the potential for strong sensitivity of observed stability to accurate treatment of side chain conformational energy profiles, as well as of hydrophobicity. The system also presents a challenge due to the relatively slow folding of β-sheets compared to the helical systems (although estimates of 100 ns for CLN025 were obtained from T-jump IR experiments [Davis et al., 2012]), and obtaining precise measures of population may be difficult. Still, use of CLN025 as a model presents a reasonable route to obtaining a qualitative view of whether ff14SB's increased helical propensity also compromises β stability.

Experimentally, CLN025 has both crystal and NMR structures [Honda et al., 2008], with 80 NOEs available for comparison. Honda et al., who derived CLN025 from chignolin, suggested that CLN025 is $> 90\%$ folded at 300 K. As mentioned, Davis et al. have probed CLN025 using infrared, confirming Honda et al.'s claim that CLN025 is a fast folder [Davis et al., 2012]. The change in nonpolar accessible surface area on folding has been estimated to be 376 Å^2 for CLN025, compared to 222 Å^2 for chignolin, thus indicating that the nonpolar side chains may be important in controlling stability [Honda et al., 2008].

For ff99SB and for ff14SB, we performed four MD runs starting from the NMR-based structure [Honda et al., 2008] closest to the ensemble average, and four additional runs starting from fully linear structures to quantify convergence. We compared simulation snapshots against the initial NMR structure using all non-symmetric atoms (Figure 2.8). We also performed cluster analysis on the combined trajectories from both force fields so that the influence of force field on cluster populations could be directly compared. Simulations with ff99SB predominantly sampled structures within cluster 0 at around 3 Å RMSD ($59 \pm 10\,\%$), or within cluster 1 at around 4.8 Å RMSD ($34 \pm 9\,\%$; all uncertainties reflect the difference between initially linear and native ensembles). Compared to ff99SB, the ff14SB simulations sampled cluster 0 with similar

frequency ($57 \pm 14\,\%$), but sampled cluster 1 much less than ff99SB ($5 \pm 3\,\%$), though the comparisons are qualitative due to the uncertainties. Instead, the ff14SB simulations are more diverse when unfolded, sampling structures ranging from 4 to just over 9 Å RMSD. Whereas ff99SB simulations sampled 194 clusters with non-zero frequency, ff14SB simulations sampled 843.

Inspection of the second major cluster of ff99SB (cluster 1, blue in Figure 2.8) reveals a shift of the C-terminal strand one residue out of phase relative to the N-terminal strand (representative structures for clusters 0 and 1 are shown overlaid on the NMR-based structure in Figure 2.9). The populations suggest that ff14SB destabilizes this alternate conformation, although the populations are not well converged; however, the difference is also qualitatively apparent in observing that this cluster is significantly sampled in 6 of 8 ff99SB simulations, but only 2-3 of 8 ff14SB simulations, with typically shorter persistence time than with ff99SB (Figure 2.8). Whether the ff14SB parameter changes favor the native-like cluster over the alternate cluster can be probed by decomposing the dihedral energies of each cluster according to each force field. In particular, we evaluated how the difference in energies of the two main clusters evolved from ff99SB to ff14SB:

$$\Delta\Delta E = (\langle U_{\text{ff14SB}}\rangle_{\text{cluster0}} - \langle U_{\text{ff14SB}}\rangle_{\text{cluster1}}) - (\langle U_{\text{ff99SB}}\rangle_{\text{cluster0}} - \langle U_{\text{ff99SB}}\rangle_{\text{cluster1}}) \tag{2.12}$$

Analysis using Equation (2.12) indicates that the dihedral changes in ff14SB favor the native cluster over the alternate by $2.9\,\text{kcal}\,\text{mol}^{-1}$ relative to ff99SB. Further decomposition of this difference reveals that parameter changes applied to Asp3 $\chi 2$ favor this native structure by $1.2\,\text{kcal}\,\text{mol}^{-1}$, and then ϕ modifications favor the native structure by $0.5\,\text{kcal}\,\text{mol}^{-1}$ in the backbone of Glu5. What's interesting is that both these changes adjust the barriers found at the level of individual amino acids. Asp3 was pushed away from $\chi 2$ of $\pm 90°$. Our training suggested that this barrier was too low with ff99SB, thus destabilizing this conformation is an expected result of considering local maxima to

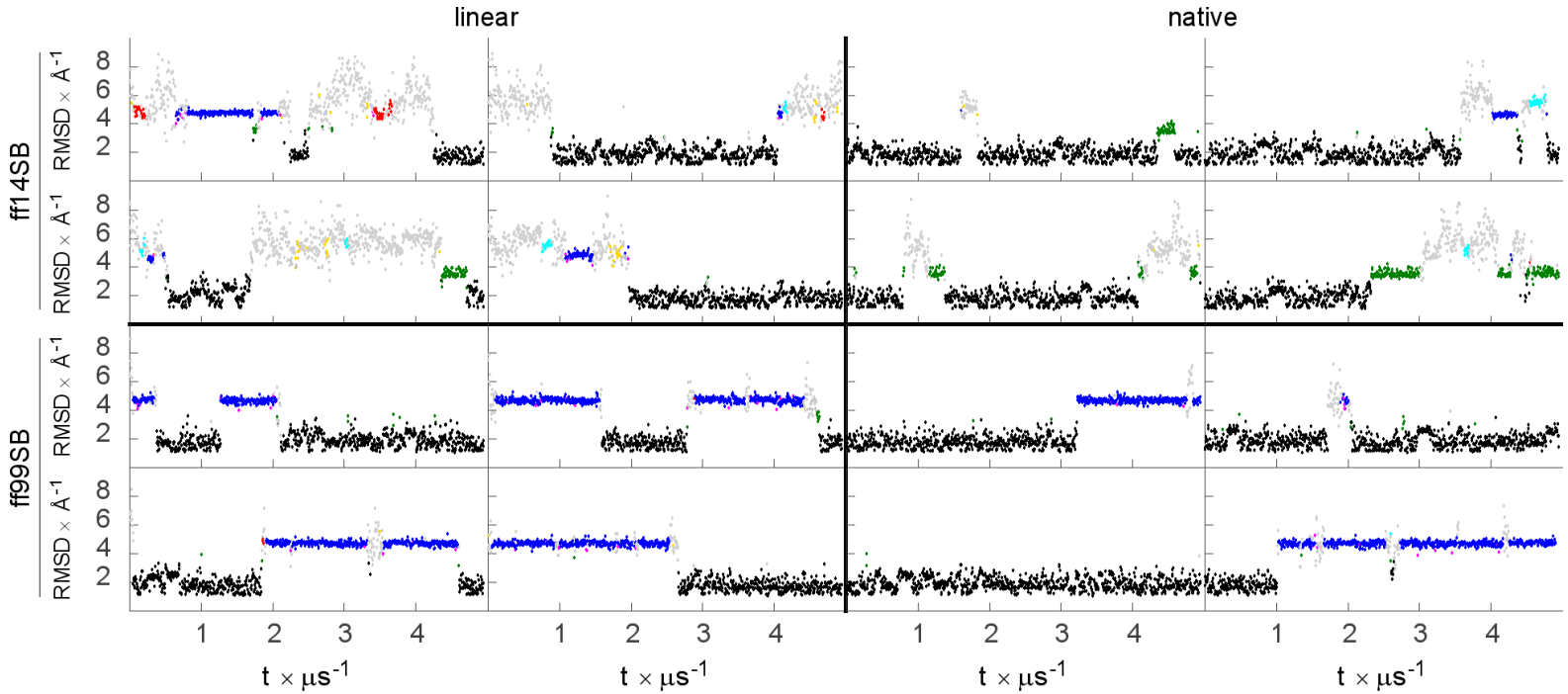

Figure 2.8: RMSD to native of the four linear and four native runs of CLN025 with ff14SB and ff99SB, colored by cluster: black=0, blue=1, green=2, cyan=3, red=4, fuchsia=5, gold=6, and all other clusters are light gray.

Table 2.17: Sum of the NOE violations from the restraints determined by Honda et al. [2008] for CLN025, for each simulation using ff14SB and ff99SB starting from linear and native conformations. *FF=force field

	linear				native			
	1	2	3	4	1	2	3	4
ff14SB	3.6	1.8	3.7	2.4	2.1	1.8	1.5	2.1
ff99SB	2.4	3.8	3.3	2.7	3.3	1.8	2.5	4.9

be important in side chain parameter derivation. Additionally, Glu5 had sampled conformations to the left of $\alpha_R$with ff99SB, but with ff14SB was confined to a more conservative α-helical basin, consistent with the ff14SB backbone goal of increasing ϕ barriers. This case illustrates the importance of reasonable barrier heights as well as the collaboration of side chain and backbone parameters.

Although it may appear desirable for ff14SB to favor the native conformation more than in ff99SB, the alternate structure echoes findings from simulations of Chignolin where presence of a similar strand-shifted structure actually improved agreement of simulations with experimental NOEs [Kührová et al., 2012]. We therefore estimated NOE buildup in the simulations, using the 'naïve' approach [Feenstra et al., 2002] used previously for Chignolin [Kührová et al., 2012]. Briefly, r^{-6} was computed for all interproton vectors and $\langle r^{-6}\rangle^{-1/6}$ of each vector compared to the NOE restraints published by Honda et al. [2008], downloaded from the BMRB [Ulrich et al., 2008]. For ambiguous restraints, contributions from each proton pair to the NOE were summed [Nilges, 1995]. The sum of NOE violations from each simulation were tabulated (Table 2.17)

Unlike Chignolin, however, better agreement is found for ff99SB native run 3, which sampled only the native cluster, than for ff99SB native runs 1 and 4 or extended runs 2 through 4, which sampled comparable amounts of the two clusters. In fact, the ff14SB extended simulations with their large variability agreed with experimentally derived NOE restraints [Honda et al., 2008] just as well as ff99SB extended simulations (2.9 ± 0.8 Å ff14SB vs. 3.1 ± 0.6 Å ff99SB).

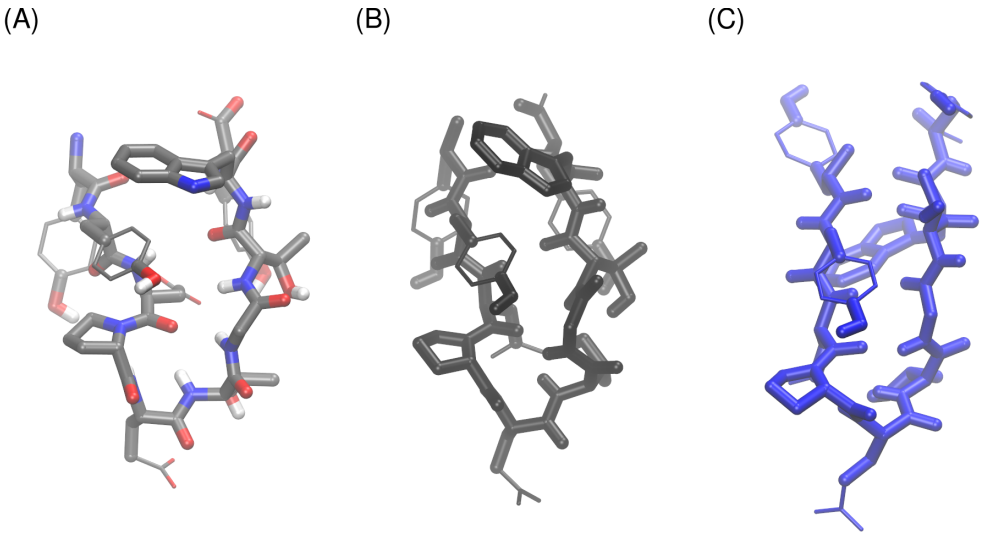

Figure 2.9: Licorice structures of CLN025. (A) The NMR structure closest to the ensemble averageHonda et al. [2008], colored by atom; (B) the centroid of cluster 0, the native-like cluster, colored black; (C) the centroid of cluster 1, where a shift in hydrogen bonding accompanies extension of the C-terminus of the second strand past the N-terminus of the first strand and flipping of W9 to the opposite side of the hairpin from Y2, against which it stacks in the NMR structure and cluster 0, and P4, which stacks against Y2, colored blue. Large licorice indicates atoms used in the clustering mask; smaller licorice atoms were omitted from clustering.

The best NOE deviations of 1.9 ± 0.2 Å were found for the ff14SB native simulations, though the difference from the 3.1 ± 1.2 Å deviations of ff99SB native simulations are only significant at $p<0.37$. Although high uncertainties suggest that no definitive conclusions can be drawn about potential improvements of ff14SB based on these simulations alone, only 1 of the ff99SB native simulations had NOE deviations (1.8) lower than the maximum ff14SB deviations (2.1). Thus the simulations together with energy analysis suggest that ff14SB is at least as reasonable as ff99SB at hairpin modeling, and thus the desirable increase in α-helical content with ff14SB did not worsen β-hairpin simulation accuracy.

Examination of NOE violations for the clusters alone shows that the first, native-like cluster violates NOE restraints by 2.5 Å, but addition of the second (whose violations are 53.5 Å) causes violations of 3.3 Å, comparable to violations for ff99SB simulations comprising these two structures (2.4 Å–4.9 Å). Interestingly, there is a third cluster most present in the ff14SB native-initiated simulations, which does improve agreement with experimental NOEs, down to 3.1 Å with the first two clusters and to 2.4 Å with only the first cluster. The first eight clusters, except the second, violate NOE restraints by 2.1 Å, and the first 15, again except the 2nd, arrives at the ff14SB native average violation of 1.9 Å. This steady but subtle downward trend continues with the addition of more clusters (by cluster 37, the violations are 1.8 Å). Of course, analyzing the subtle differences caused by the many small clusters is not necessarily statistically rigorous, especially as the source simulations have not converged. The point, however, is that the structural diversity observed with ff14SB may not be totally undesired, and actually improves the ff14SB match to experiment.

2.7.4 High quality of dynamics in the native state is maintained

We also evaluated the ability of ff14SB to reproduce local dynamics in well-folded proteins as measured by NH S^2 Lipari-Szabo order parameters. We calculated NH order parameters from the same simulations used for side chain scalar coupling evaluation. This calculation was performed using iRED,

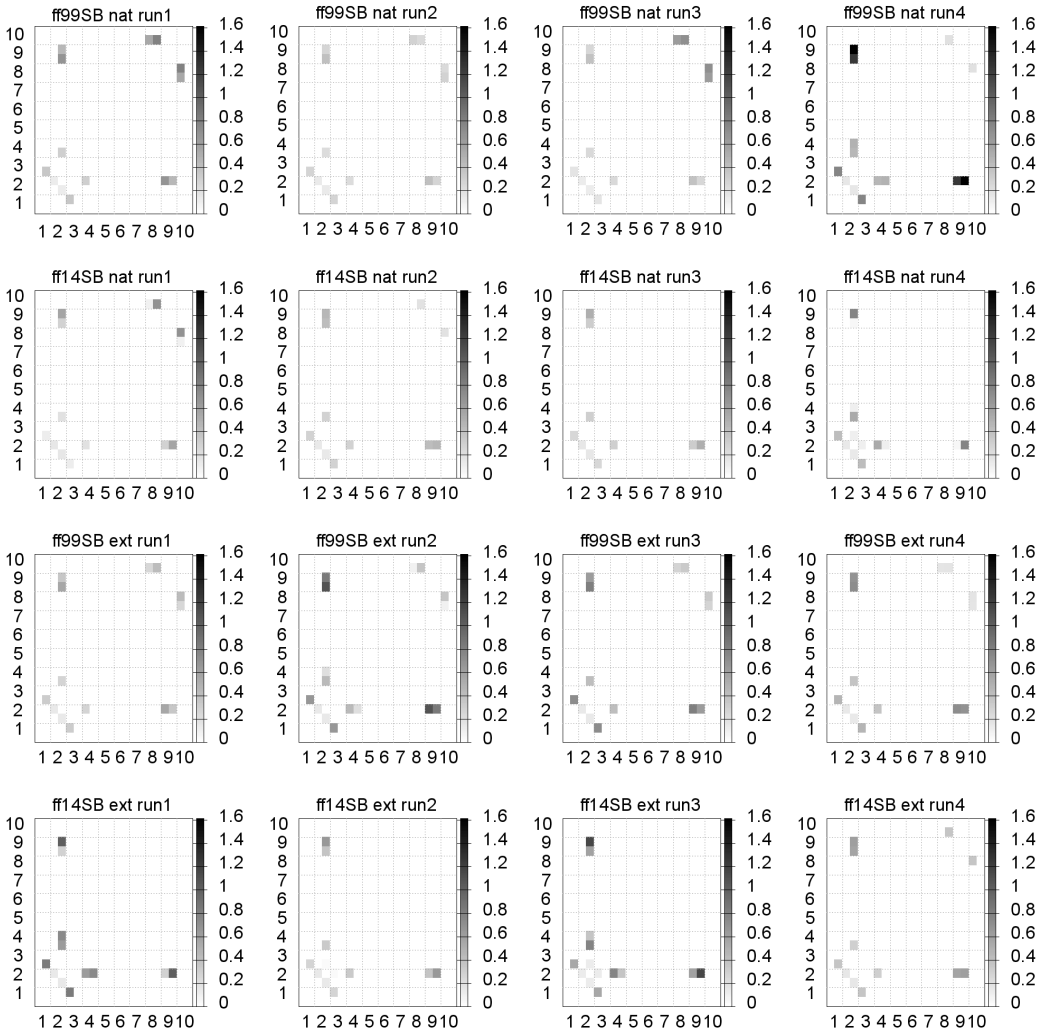

Figure 2.10: NOE [Honda et al., 2008] violations of CLN025 simulations with ff99SB and ff14SB, of each amino acid backbone and side chain to each other amino acid backbone and side chain. Nearly all simulations have minor violations within the N-terminus and between the N- and C-termini, whereas ff99SB simulations have greater violations within the C-terminus than ff14SB.

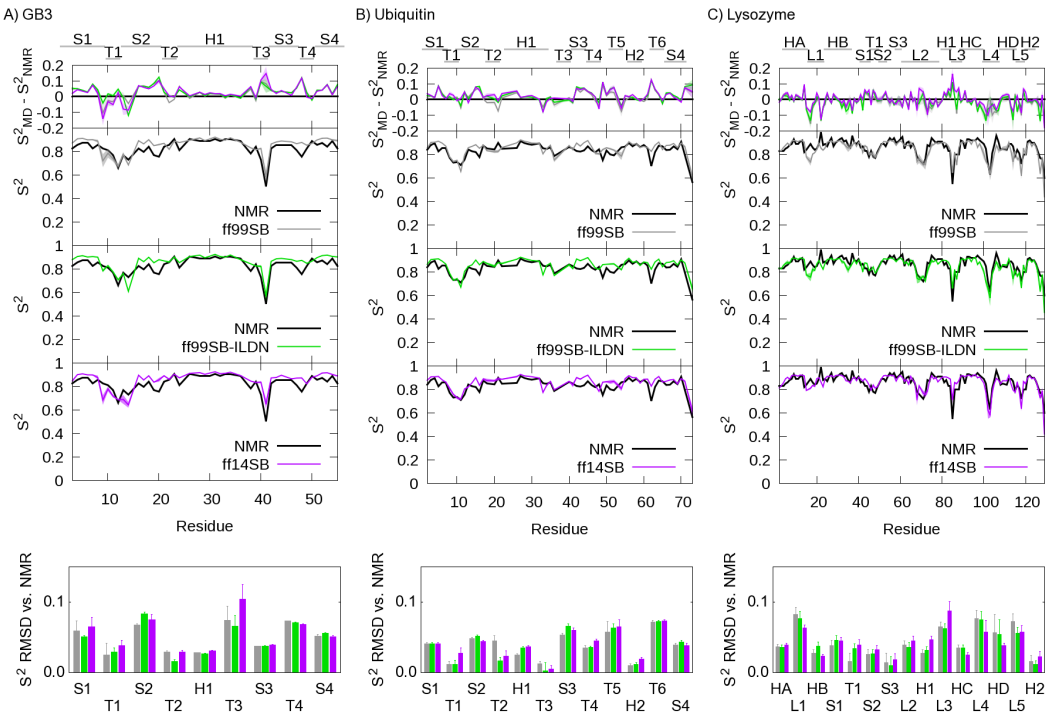

Figure 2.11: Order parameters from NMR compared to those backcalculated by iRED for ff99SB, ff99SB-ILDN, and ff14SB simulations of GB3, ubiquitin, and lysozyme. The top panels show differences between simulation and experiment, while the lowest panels show average data for each secondary structure region, following Hornak et al. [2006].

which does not require separability of local and global motions [Prompers and Brüschweiler, 2002]. The iRED-calculated order parameters, shown in Figure 2.11, are within 0.05 RMSD of NMR for all systems and force fields. We conclude that the high quality order parameter reproduction previously reported for ff99SB [Hornak et al., 2006] is maintained with ff14SB. There are, however, subtle differences worth noting.

Several turns or loops increased in order with ff14SB. In the cases of loops L1 and L4 in lysozyme, this increased order better reproduces experimental order parameters. Turn T3 in GB3 and loop L3 in lysozyme, however, may have become too ordered. L3 begins with S85, which is 0.16±0.04 too ordered. With ff99SB, S85 was already 0.10±0.02 too ordered, meanwhile D87 was 0.09±0.02 too disordered with ff99SB but only 0.05 ± 0.01 too ordered with ff14SB.

Hence, ff14SB is not statistically worse at modeling S85 spin relaxation, which may have already been too ordered with ff99SB. Naturally, these estimates depend on both the accuracy of the iRED method and the uncertainty in the simulations and experiments; absolute comparisons against the experimental S^2 differing by 0.06 ± 0.04 should not be overemphasized. But as a trend, there appears to be slightly less flexibility in loops with ff14SB compared to ff99SB, both aiding and lessening agreement between iRED and experimental order parameters. We conclude that ff14SB maintained ff99SB order parameter reproduction on average, but with subtle reduction in flexibility.

2.8 Conclusion

The weaknesses of ff99SB addressed by ff14SB [Maier et al., 2015] are its less than ideal agreement with polyalanine scalar couplings, insufficient helical propensity, and, as described here, inaccurate side chain preferences. We tackled the latter by de novo fitting against a backbone-independent MP2 training set. The new model ff14SB improved side chain rotamer distributions as suggested by scalar couplings, while augmenting helical content of small peptides and maintaining the reasonable reproduction of order parameters and hairpin structure. The ubiquity of ff14SB improvements and more thorough description of potential limitations will require further testing than possible here. But based on the benchmark reported, we recommend ff14SB for the simulation of proteins and peptides.

Chapter 3

Assessing the factors that may be used to improve the ff99SB protein backbone parameter training

3.1 Introduction

As discussed in Chapter 1, the accuracy of AMBER force field 99SB (ff99SB) [Hornak et al., 2006, Showalter and Bruschweiler, 2007, Li and Brüschweiler, 2009, Lange et al., 2010, Cerutti et al., 2010] resulted in its wide adoption by the simulation community. This wide adoption allowed the identification of trends in ff99SB's strengths and weaknesses—weaknesses that groups have sought to improve. Chapter 2 details the refinement of the ff99SB side chain parameters against quantum mechanics energy profiles, resulting in improved NMR scalar coupling reproduction. Additionally, issues with ff99SB backbone sampling have been reported [Thompson et al., 2010, Best and Hummer, 2009, Maier et al., 2015]. The accuracy of ff99SB has been directly evaluated by comparison against solution NMR data [Best et al., 2008, Li and Brüschweiler, 2009, Wickstrom et al., 2009]. Thus recent refinements of ff99SB

backbone parameters have focused on empirical adjustments to better reproduce solution NMR data [Best and Hummer, 2009, Li and Brüschweiler, 2011, Maier et al., 2015]. This is reasonable as deviations from NMR suggested that only small changes to the ff99SB backbone energy landscape may be needed. Moreover, the solution NMR data likely includes effects that are difficult to incorporate by training against ab initio energies of small model systems.

These small empirical corrections, while abating their target errors, may miss less apparent shortcomings of the ff99SB energy surface. For example, the ff99SB fitting against only in vacuo energy minima left much of the backbone energy surface unconstrained. Whereas ff14SB is unique among ff99SB backbone modifications in that it adjusted not only energy minima, but specifically targeted the β-ppII transition [Maier et al., 2015], it is still not clear whether other changes might be beneficial.

Another concern when comparing in vacuo energies between QM and MM has been that most biomolecular force fields do not have true gas-phase charges, but charges that are more suitable for simulations in water. Thus artifacts can arise in vacuo simply because the QM electron density polarizes only in response to the molecule, in the absence of any dielectric medium, whereas fixed MM charges may be "pre-polarized" in anticipation of such a medium. Although empirical corrections may help compensate for this weakness, work on protein force fields [Duan et al., 2003] and more recent work on nucleic acid force fields [Zgarbova et al., 2011, 2013] suggest that calculation of QM and MM energies in the context of implicit solvent may also be a viable workaround to this problem.

Additionally, ff99SB and its derivatives used alanine as a model for all non-glycine amino acids. But the sequence dependence of structural preferences may require separate corrections for different amino acids. For example, Martinez showed that the ff14SB backbone updates, while improving agreement with Ala_5 scalar couplings, worsened agreement with scalar couplings for Val_3 [Martinez, 2014]. Notably, higher experimental $H_N H_\alpha$ scalar couplings for Val_3 than Ala_3 suggest more structures along the β-ppII transition for valine than for alanine. To better picture this conformational difference, histograms of ϕ/ψ distributions for alanine and valine extracted from the PDB

by Lovell et al. [2003] are plotted in Figure 3.1. These histograms corroborate the solution NMR data in suggesting that valine samples conformations more evenly along the β-ppII barrier than alanine. This difference is expected as there is a higher chance of steric clash between the backbone and valine's two γ-carbons than between the backbone and alanine's methyl hydrogens. For example, the helical conformation, where the backbone NH points towards the side chain, would be destabilized by valine's isopropyl side chain relative to alanine's methyl. Whereas this could reasonably be accounted for in force fields by the presence of the valine side chain and reasonable non-bonded parameters, simulations with ff99SB and ff14SB as compared to NMR scalar couplings suggest that the correction needed along this transition for alanine and valine may differ. If this is the case, it would also suggest that the alanine-based QM fitting in ff99SB [Hornak et al., 2006] may not apply to all amino acids. This result would not be surprising given that, as described in Chapter 2, different amino acids were found to need different side chain corrections. Still, another possibility is that a single correction yet undiscovered may apply to both alanine and valine reasonably well.

Missing sequence dependence was additionally suggested when Hai Nguyen and the author of this dissertation applied ff99SB together with the updated side chain parameters described in Chapter 2 (ff14SBonlysc) to the folding of seventeen proteins with diverse topologies [Nguyen et al., 2014]. Although most systems folded successfully, some failures occurred that did not suggest systematic trends in one secondary structure being too stable relative to another. In the native conformation of hypothetical protein 1WHZ, for example, there are a three-strand β-sheet and three helices. In the structure preferred by ff14SBonlysc, the first β-strand became an α-helix, and the last two α-helices became strands in a β-turn [Nguyen et al., 2014]. This suggests errors that are not systematic across all amino acids, but may depend on the specific amino acids in each of these secondary structure units.

One more concern with the Cornell et al. [1995] line of force fields (including ff99SB [Hornak et al., 2006] and ff14SB [Maier et al., 2015]) is that the shapes of the structure basins do not quite match the shapes suggested by the PDB Figure 3.1. Generally, whereas the PDB suggests that ϕ and ψ should be

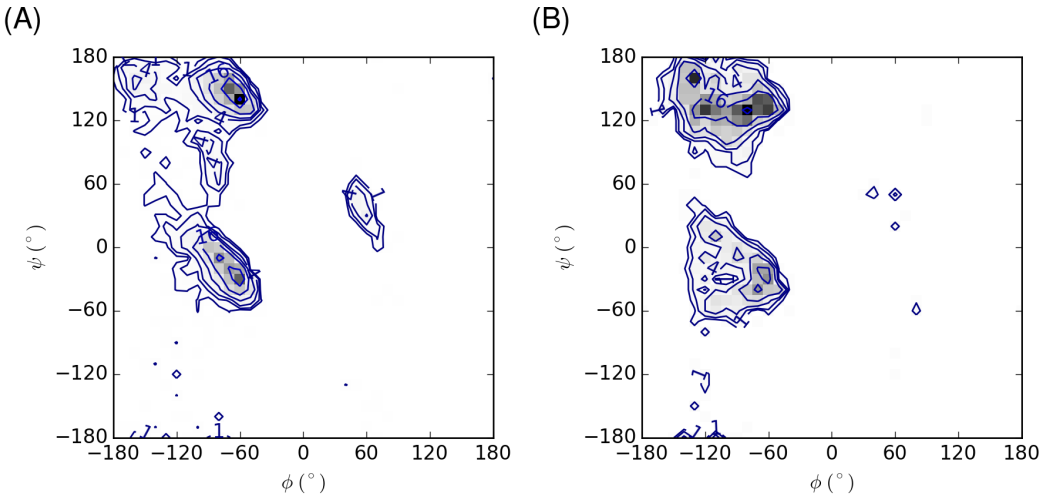

Figure 3.1: Histograms of (A) alanine and (B) valine backbone dihedrals ϕ and ψ based on the rotamer libraries provided by Lovell et al. [2003]. Each contour line represents a doubling in population, with labels indicating the fold enrichment compared to a completely flat distribution. Density is also shown as grids filled with white (no density) to black (maximum density).

anticorrelated, for example, in the α_R basin, ff99SB and ff14SB produce round blobs suggesting that ϕ and ψ are sampled somewhat independently. Furthermore, whereas PDB structures indicate a conformation centered around $\phi, \psi = (-75°, 75°)$, just below ppII, this conformation is not well sampled by ff99SB or ff14SB. As this work seeks to build upon the energy surface achieved by ff14SB, a goal is to more faithfully reproduce the fine details of each structure basin, not only the gross features like relative energy between α, β, and ppII.

One important tool for achieving greater correlation between ϕ and ψ would be coupled corrections that are a function of both dihedrals. The preeminent coupled correction in fixed-charge force fields is the CMAP [MacKerell et al., 2004b] employed by CHARMM force fields like CHARMM22/CMAP [MacKerell et al., 2004a] and CHARMM36 [Best et al., 2012]. The CMAP correction can reproduce any grid of energy differences spaced 15° across two dihedrals. The phase space between grid points is then

extrapolated using bicubic interpolation. CHARMM36 currently has individual CMAP corrections for glycine and proline, as well as a non-glylyl-non-prolyl correction based on alanine. It is trivial, however, to apply a CMAP correction to any combination of amino acids; in AMBER, one can simply specify a list of residues to be targeted by a CMAP when the CMAP is loaded. This dissertation will examine whether further CMAP amino acid demarcations could benefit simulations using force fields derived from ff94 [Cornell et al., 1995].

To investigate all these concerns, this chapter builds on the ideas of Chapter 2 that more, possibly residue-specific parameters, trained against quantum mechanics energies, with careful consideration of how the energies should be calculated and compared, may be a viable tool for retraining the full backbone energy surface of individual amino acids. We have two expectations if this method is viable. First, simulations of Ala_5 with new alanine force field parameters should agree with scalar coupling data [Graf et al., 2007] comparably to ff14SB [Maier et al., 2015] without empirical adjustment. Surpassing ff14SB would likely be an unreasonable challenge as ff14SB was empirically optimized to reproduce Ala_5 scalar couplings. Second, as the ultimate goal is to derive backbone parameters for non-glycine/non-alanine amino acids, agreement with Ala_5 scalar couplings should be accompanied by appropriate residue-dependent backbone preferences when the method is applied to other amino acids. In this case we use valine, as Val_3 is also characterized by solution NMR scalar couplings and valine exhibits different preferences from alanine that are not fully reproduced by ff99SB or ff14SB. Both alanine-derived parameters and valine-derived parameters will be considered, to evaluate whether residue-specific parameters are necessary.

Before attempting to retrain backbone parameters that might improve upon ff99SB, we first review the ff99SB assumptions that have been described so far. It was assumed that:

1. Alanine is an appropriate model for derivation of all nonglycine backbone parameters. Other non-glycine amino acids will be distinguished by their side chains. This assumption implies that the non-bonded effects of the side chain on the backbone are reasonable.

2. In vacuo energy minima would be a sufficient training set with constraints on the amplitude of some parameters. This assumption implies that transition energies can be ignored in training, as well as that the in vacuo minima obtained represent those sampled in dynamics simulations. It was acknowledged, however, that this assumption was necessary for computational tractability.

3. Vacuum energies would allow rigorous training of parameters for use in solution. As discussed in Chapter 1, the molecular mechanical charges of ff99SB are overpolarized by approximately 20% relative to gas-phase charges. This can lead to errors compared to gas-phase QM that result from using an incompatible charge model. Note that the same assumption was made in the side chain fitting of ff14SB [Maier et al., 2015] (Chapter 2).

4. Simple, uncoupled cosine corrections to ϕ and ψ dihedral torsion would adequately capture the differences of the true Ramachandran energy surface from those captured by bonded (bond and angle) and non-bonded (electrostatic and van der Waals) interactions.

5. Tetrapeptides that can form a helical hydrogen bond will provide an accurate means of accounting for secondary structure propensity in training (they are also needed when training against only energetic minima in vacuo because the dipeptide has no gas-phase helical minimum).

As a first step, this chapter evaluates assumptions 2–5 for alanine, for which there is a wealth of experimental scalar coupling data [Graf et al., 2007]. Importantly, alanine lacks side chains, simplifying the conformational space that needs to be mapped when fitting parameters. Moreover, its small size facilitates the many quantum mechanics calculations required for testing the variations to be described. The precedent of training against alanine [Hornak et al., 2006, Maier et al., 2015] means that the accuracy of other force field development approaches is readily available for comparison to a new training method.

To assess Assumption 2, three structure sets are developed, including: a set of variable structures extracted from high-temperature simulations; two-dimensional grids in ϕ and ψ; and, as in ff99SB [Hornak et al., 2006], energetic minima. It should be noted that for the grids, all amino acids in a molecule were assigned the same ϕ/ψ values. Thus, the grid is not a full grid spanning all energetically relevant conformations of molecules with more than one amino acid.

To investigate Assumption 3, energy calculations are carried out in vacuum as well as with implicit solvent. As mentioned above, ff03 [Duan et al., 2003] and recent RNA force fields [Zgarbova et al., 2011, 2013] employed implicit solvation in the QM and MM energy comparisons. This work tests comparing QM with COSMO [Klamt and Schuurmann, 1993] and MM with Poisson Boltzmann (PB) [Gilson et al., 1993], as done in the more recent RNA parameterization [Zgarbova et al., 2011, 2013].

To consider Assumption 4, corrections are derived using uncoupled cosine terms or coupled corrections. We use the CMAP [MacKerell et al., 2004b] well utilized by CHARMM force fields. Whereas CHARMM employs Ala-based CMAPs for all non-glycyl,non-prolyl amino acids, this dissertation tests whether AMBER force fields could benefit from additional partitioning of backbone parameter space.

And lastly, to test Assumption 5, structure sets for tetrapeptides and for dipeptides are considered. Against all expectations, we found that parameters based on Ala_1 allowed more accurate simulation results than those based on Ala_3. More remarkably, the results suggested that even smaller training compounds may be appropriate; thus we explored a scheme of extrapolating beyond the scale of the dipeptide. This extrapolation could be interpreted as either adjusting the effect of the peptide bonds in the training to achieve a correction based on a single peptide bond, which could be appropriate for polypeptides where the ratio of amino acids to peptides is approximately one, or perhaps serving to limit the effect of artifacts that depend on the length of the peptide used in training.

After first focusing on evaluating the latter four assumptions just reviewed

to develop a protocol for alanine, the ultimate goal is to reproduce the sequence-dependent preferences of different amino acids. Thus, Assumption 1 will be tested by applying the insights from alanine to deriving parameters for valine. Importantly, the scalar couplings for Ala_3 and Val_3 indicate different preferences that so far have not been captured by a single set of parameters with ff99SB or ff14SB. Our expectation is that a force field derived using a reasonable physics-based method should be able to capture these differences. Alanine parameters will also be tested in valine to evaluate the need for a separate set of parameters.

We note that we limited this analysis to goals that most directly stem from our experience with ff14SB. Particularly, we tried to use the tenets applied to the ff14SB side chain training to address limitations with the ff99SB backbone parameters as subsequently adjusted with ff14SB. The most concerning of these is the potentially poor transferability of the alanine correction to (in particular) β-branched amino acids like valine. But there are myriad other options and approaches that may be worth visiting in the not-so-distant future.

Firstly, there are several alternate charge models to the Cornell et al. model employed by ff99SB. For example, ff03 [Duan et al., 2003] and ff14ipq [Cerutti et al., 2014] both employ charges derived from electrostatic potentials in the context of solvent, with ff03 using PCM and ff14ipq using average charge distributions from TIP4P simulations. Such charge derivation schemes are promising, as the polarization in the charges is based on more readily understandable physics than the cancellation of error achieved by using gas-phase HF/6-31G* ESPs in the Cornell et al. charge set [Bayly et al., 1993, Cornell et al., 1993, 1995]. But neither of these force fields has been as widely adopted with ff99SB, and the author has more intimate experience with the specific shortcomings of ff14SB. Thus, as a follow-up to the first part of this dissertation, this work is restricted to optimizing ff14SB.

Even more promising than the above fixed charge force fields, in principle, are polarizable force fields that allow a rearrangement of charge based on molecular conformation. Such force fields are generally more expensive to use than fixed charge force fields like ff14SB, and have not been demonstrated to

the same point as their fixed charge contemporaries. One reasonable compromise between the accuracy of fully polarizable force fields and the simplicity of fixed-charge force fields is semi-polarizable models like the effective polarizable bond approach [Xiao et al., 2013]. In that model, only polar bonds are allowed to polarize, with a computational cost increase of only 10% relative to fixed-charge models. Although all these models promise to be very accurate when appropriately trained, again, we focused on evaluating what options could be applied to the simple and well-tested ff99SB as well as ff14SB. Hopefully, the insights gained may also apply to more advanced models.

Another thing that hasn't been done here is to optimize the parameters against experimental observations. One example of residue-specific parameters for AMBER, the Residue-Specific Force Field 2 (RSFF2), had parameters for each amino acid derived from coil conformation libraries [Zhou et al., 2015]. It was assumed that the distributions found in the PDB would match those in solution. RSFF2 actually can fold several systems with reasonable dynamics [Zhou et al., 2015]. Yong Duan's group has also added some parameters to AMBER for use with AMOEBA [Ren and Ponder, 2003], that include CMAPs for each amino acid based on PDB distributions. Despite the usage of such an approach by multiple groups and the apparent successes reported by Zhou et al. [2015], it is not clear to the author that coil libraries should indicate anything about physics or dynamics. Each structure is not guaranteed to be at the same temperature, there is usually only one conformation to represent a dynamic ensemble, the structures present are generally only those that could be crystallized, and usage of an amino acid in a protein may be as likely to reflect the requirements of biology as the actual energetic preferences of the same amino acid in a different context. These issues are more complicated than this chapter was intended to address, although based on the results reported by Zhou et al. [2015] may be worth investigating.

Although the above are some of the broad issues not addressed, there are many specific details to be worked out, at least as many as were examined for the side chain training discussed in Chapter 2. The landscape of backbone parameters is also much more complicated than that of side chain parameters; therefore, the accuracy of the backbone corrections must be of a very high

level to contribute to this field. With the many questions and potential issues involved, what follows is an examination of a subset of the issues that were considered most salient by the author of this dissertation. Issues that may need further examination include:

1. Potential inconsistency between the QM COSMO [Klamt and Schuurmann, 1993] and MM PB [Gilson et al., 1993] solvation models. Besides It is unclear yet whether the radii should be the mbondi radii suggested for PB, the radii suggested for COSMO, or whether each solvation model should use the radii suggested for it. Initially, the author considered the recommended options for each solvation model, including radii. But there is some evidence that differences in radii, and thus in formation of a solvent cavity, may be problematic when comparing energies for specific structures.

2. Potential inconsistency between COSMO and RESP charges. Part of the goal of the screening is also to reduce the level of electrostatic interactions to that expected in aqueous solution, thus lessening the effects of small differences in charge, but inconsistency could still affect the results.

3. The Ala_3 "grids" calculated here only include structures where all amino acids are in the same conformation (ϕ, ψ, χ). Whereas this limitation grossly simplifies the conformational space and includes extended and helical segments, there are many possible conformations that are absent. These missing conformations may be especially important in loops devoid of canonical secondary structure.

4. The conformations used were optimized in vacuo. The limiting step was the MM structure optimization in PB, which never completed within the 72-hour wallclock limit, compared to QM optimizations in COSMO that completed within 10 hours. Pursuing this further will require some investigation of compromises in the PB parameters for the sake of efficiency and/or a larger computational investment.

5. The QM optimizations considered here used the same level of theory

(HF/6-31G*) as the ff99SB training and the ff14SB side chain training. Using a more expensive level of theory may provide greater accuracy.

It is emphasized that this work is preliminary, and serves as much as a guideline for future work that can use the ideas presented herein, as it is a vehicle for testing the ideas themselves. But the results obtained for simulations of Ala_5 and Val_3 are promising. Thus the author believes the work that follows is worth building on with a more thorough effort, even one that may require a significant fraction of a Ph.D.

3.2 Methods

3.2.1 Structure generation by high-T simulations

To generate structures for a random conformational set and a minimum energy set, simulations were performed of tetrapeptides using ff99SB for 100 ns. The timestep was 1 fs, with structures saved to a trajectory every 2 fs. Temperatures were maintained at 500 K using the Langevin thermostat with $\gamma_{\ln} = 1.0$. Implicit solvation was provided by GB-Neck2 [Nguyen et al., 2013] to prevent favorable electrostatics from inhibiting efficient sampling.

The natural goal is to extract conformations with maximal spread of the torsional landscape and minimal coupling between dihedrals. This was accomplished by minimizing the torsional similarity (Equation (3.1)) between each conformation A and all other conformations B.

$$\text{similarity}_A = \sum_{B \neq A}^{\text{conformations}} \prod_{\phi}^{\text{torsions}} (1 + \cos(\phi_A - \phi_B)) \tag{3.1}$$

We minimized the sum over Equation (3.1) using a Monte Carlo approach, starting with 500 randomly selected conformations. We then chose one of the 500 at random and one from the remaining trajectory structures, and calculated the similarity of each to the rest of the population. We then exchanged the two if the similarity was lower for the structure selected from the trajectory, or if a random number between 0 and 1 was less than the ratio of the

similarity score for the selected conformation of the 500 chosen to the similarity score for the selected conformation of the many yet unchosen structures. This was repeated until the overall similarity summed over Equation 3.1 for all 500 structures stopped decreasing. In practice, 100 000 attempted Monte Carlo moves was found to be sufficient, with no detectable improvements after 30 000 iterations.

3.2.2 Energy calculations

All QM calculations were performed using ORCA [Neese, 2012], a flexible quantum chemistry software package that can calculate gradients while employing the resolution of the identity approximation [Vahtras et al., 1993] together with COSMO implicit solvation [Klamt and Schuurmann, 1993]. The 'TightSCF' keyword was used for energy calculations and the 'TightOpt' keyword for structure optimizations. Implicit solvation was provided by the COSMO model [Klamt and Schuurmann, 1993] via the 'COSMO(Water)' directive. Otherwise, all default options were used. QM optimizations were performed with HF/6-31G* theory, whereas single point energies were calculated with RI-MP2/cc-pVTZ.

All MM calculations were performed using AMBER14 [Case et al., 2014]. Minimizations were performed for a maximum of 10 000 000 cycles in vacuo with the cutoff set to 99.0 Å—larger than the system size to enable the direct calculation of all non-bonded interactions. Minimizations were performed with the AMBER force field 14SB (ff14SB) [Maier et al., 2015]. Implicit solvation was provided by PB [Gilson et al., 1993] (igb=10 in AMBER), with either the modified Bondi radii [Bondi, 1964] or the COSMO radii. The PB options were: $\epsilon_{\text{in}} = 1.0$; $\epsilon_{\text{out}} = 78.4$; a ratio of 4.0 between the longest dimension of the rectangular finite-difference grid and that of the solute; solvent probe radius of 1.4 Å; mobile ion probe radius of 2.0 Å; the solvent excluded surface as implemented by Wang et al. [2012]; trimer arc dots with 0.125 Å resolution; convergence criterion of 0.0001; maximum iterations of 10 000; and grid spacing of 0.1 Å.

All structure optimizations included restraints on ϕ, ψ, and (for valine)

$\chi 1$ defined as N–Cα–Cβ–Cγ1, with harmonic weights of $25\,000\,\mathrm{kcal\,mol^{-1}}$. In QM optimizations, these dihedrals were constrained, and thus were invariant, rather than restrained against an energy penalty.

3.2.3 Fitting

Analytical solution of parameters is an attractive option. Although the solution of parameters typically takes less time than generation of ab initio training data or testing of the parameters by rigorous MD calculations, it is still a critical step during which one would like to ensure an optimal solution. Global minimizers like genetic algorithms have the advantages of maximizing (or minimizing) arbitrary objective functions, such as the mean absolute relative energy error. But convergence is always a question; even if multiple genetic algorithm runs converge to the same solution, there is no guarantee of a global optimum. Analytical solution could alleviate this problem. Typically, however, least linear squares solvers minimize the square errors that emphasize outliers, and require fitting of a single offset between molecular mechanics and target energy surfaces.

The latter of these limitations can be alleviated, however. Traditionally, the problem of fitting absolute energy errors with an offset has been formulated:

$$\sum_{v}\sum_{n}^{N_v}\left(\sum_{d}\cos(n\phi_i^d)\right)V_{v,n}+C=E_i-E_i^0 \tag{3.2}$$

for conformation i, with v representing all dihedral corrections with N_v periodicities n and amplitudes $V_{v,n}$, d representing all dihedrals with values ϕ_d to which each dihedral correction v applies, and C representing an offset between E_i, the target energy, and E_i^0, the baseline energy to which the left half of the equation is added.

Meanwhile, the fitting of relative energy errors can be formulated by simply subtracting Equation 3.2 evaluated for the differences between conformations

i and j:

$$\sum_{v}\sum_{n}^{N_v}\left(\sum_{d}\cos(n\phi_i^d)-\cos(n\phi_j^d)\right)V_{v,n}=(E_i-E_i^0)-(E_j-E_j^0) \tag{3.3}$$

In this case, the offset C drops out of the fit, as errors in relative energies are minimized. The only limitation is the requirement to fit square errors, rather than absolute errors. This will emphasize outliers in the data set; therefore, it is more important to discard real outliers where energies may not be sensible. Given the exponential relationship between energy and population, however, a square mapping of energy differences to the objective function may actually be more relevant to dynamics than absolute values.

Hence the system of linear equations in Equation 3.4 becomes that in Equation 3.5.

$$\begin{pmatrix}\cos(n_1\phi_{1,1}) & \cdots & \cos(n_N\phi_{1,N}) & 1\\ \vdots & \ddots & \vdots & \vdots\\ \cos(n_1\phi_{M,1}) & \cdots & \cos(n_N\phi_{M,1}) & 1\end{pmatrix}\begin{pmatrix}V_1\\ \vdots\\ V_N\\ C\end{pmatrix}=\begin{pmatrix}\Delta E_1\\ \vdots\\ \Delta E_M\end{pmatrix} \tag{3.4}$$

$$\begin{pmatrix}\cos(n_1\phi_{1,1})-\cos(n_1\phi_{2,1}) & \cdots & \cos(n_N\phi_{1,N})-\cos(n_N\phi_{2,N})\\ \vdots & \ddots & \vdots\\ \cos(n_1\phi_{M-1,1})-\cos(n_1\phi_{M,1}) & \cdots & \cos(n_N\phi_{M-1,1})-\cos(n_N\phi_{M,1})\end{pmatrix}\begin{pmatrix}V_1\\ \vdots\\ V_N\end{pmatrix}$$
$$=\begin{pmatrix}\Delta E_1-\Delta E_2\\ \vdots\\ \Delta E_{M-1}-\Delta E_M\end{pmatrix} \tag{3.5}$$

where M is the number of conformations and N is the number of dihedral corrections.

3.2.4 Extrapolation of parameters to length-independence

The energy differences (ΔE_{A0}) for the hypothetical alanine monopeptide (A0) were calculated by a linear fit to dipeptide (A1) and tetrapeptide (A3) per-residue energy differences. For every conformation on a grid, the QM − MM energy differences for A1 (ΔE_{A1}) and for A3 (ΔE_{A3}) were computed and divided by the number of residues. The slope in the per-residue energy differences versus number of residues ($\frac{\Delta E}{\Delta N_{\mathrm{res}}}$) was determined as in Equation (3.6), subtracting the A3 differences minus the A1 differences, and dividing by the 2 residue difference between A1 and A3.

$$\frac{\Delta E}{\Delta N_{\mathrm{res}}} = \frac{\frac{1}{3}\Delta E_{\mathrm{A3}} - \Delta E_{\mathrm{A1}}}{2 \text{ residues}} \tag{3.6}$$

Then, from the energy differences for A1 were subtracted the slopes in the per-residue energy differences versus number of residues, multiplied by the 1 residue needed to extrapolate to the monopeptide (A0) level. Hence the target energy surface for Ala_0 was defined as:

$$\Delta E_{\mathrm{A0}} = \Delta E_{\mathrm{A1}} - \frac{\Delta E}{\Delta N_{\mathrm{res}}} \times 1 \text{ residue} \tag{3.7}$$

The A0 − A1 offset can be simply rearranged as

$$\Delta E_{\mathrm{A0}} - \Delta E_{\mathrm{A1}} = -\frac{\Delta E}{\Delta N_{\mathrm{res}}} \times 1 \text{ residue} \tag{3.8}$$

This offset was added to the energy differences for other amino acid dipeptides. The approximate extrapolation from V1 to V0, referred to as V0(A) to parenthetically indicate the use of the alanine offset, was calculated as in Equation (3.9).

$$\Delta E_{\mathrm{V0(A)}} = \Delta E_{\mathrm{V1}} - \frac{\Delta E}{\Delta N_{\mathrm{res}}} \times 1 \text{ residue} \tag{3.9}$$

3.2.5 Test simulations

Initial structures

Helical conformations were defined as all $(\phi, \psi) = (-60°, -45°)$. Linear conformations were defined as all $(\phi, \psi) = (180°, 180°)$. TIP3P water [Jorgensen et al., 1983] was added to fill truncated octahedra with at least 8 Å from the system to the water boundary or 12 Å for Val_3. A larger buffer was needed for Val_3 as the CUDA implementation of pmemd [Salomon-Ferrer et al., 2013] is unstable using the particle mesh Ewald approximation [Darden et al., 1993] with small systems.

General details

Equilibration was performed with a weak-coupling (Berendsen) thermostat and barostat [Berendsen et al., 1984], targeting 1 bar pressure with isotropic position scaling, as follows. With 100 kcal mol^{-1} Å^{-2} positional restraints on protein heavy atoms, structures were minimized for up to 10 000 cycles and then heated at constant volume from 100 K to 300 K over 100 ps, followed by another 100 ps at 300 K. The pressure was equilibrated for 100 ps and then 250 ps with time constants of 100 fs and then 500 fs on coupling of pressure and temperature to 1 bar and 300 K, and 100 kcal mol^{-1} Å^{-2} and then 10 kcal mol^{-1} Å^{-2} positional restraints on protein heavy atoms. The system was again minimized, restraining only the protein main chain N, C_α, and C positionally with 10 kcal mol^{-1} Å^{-2} for up to 10 000 cycles. Three 100 ps simulations with temperature and pressure time constants of 500 fs were performed, with backbone restraints of 10 kcal mol^{-1} Å^{-2}, 1 kcal mol^{-1} Å^{-2}, and then 0.1 kcal mol^{-1} Å^{-2}. Finally, the system was simulated unrestrained with pressure and temperature time constants of 1 ps for 500 ps with a 2 fs time step, removing center-of-mass translation every ps.

SHAKE [Ryckaert et al., 1977] was performed on all bonds including hydrogen with the AMBER default tolerance of 10^{-5} Å for NpT/NVT and 10^{-6} Å for NVE. Non-bonded interactions were calculated directly up to 8 Å with cubic spline switching and the particle-mesh Ewald approximation [Darden et al.,

1993] in explicit solvent, with direct sum tolerances of 10^{-5} for NpT/NVT or 10^{-6} for NVE. The timesteps for NpT/NVT and NVE simulations were 2 fs and 1 fs, respectively. Tighter convergence criteria and a shorter timestep facilitated the energy conservation required for NVE.

Production simulations were carried out in the NVE ensemble.

3.2.6 Analysis

Scalar couplings were calculated from simulations using Karplus relations [Karplus, 1959, 1963]. Backbone scalar couplings were calculated as by Best et al. [Best et al., 2008]: using the *Orig* parameters [Ding and Gronenborn, 2004, Hennig et al., 2000, Hu and Bax, 1997, Wirmer and Schwalbe, 2002] also used by Graf et al. [2007] and the *Dft1* and *Dft2* parameters from Case et al. [2000]. Experimental scalar couplings for Ala_5 and Val_3 were obtained by Graf et al. [2007].

3.3 Evaluating methods with alanine

As outlined above, the first goal is to test the assumptions in the ff99SB training against alanine, before moving on to the question of whether different amino acids could benefit from unique parameters. This investigation is conducted by simulating Ala_5 to evaluate ff99SB, ff14SB, and modifications to ff14SB that will be described. First, TIP3P is compared against a new water model [Izadi et al., 2014]. Then training against minimum energy conformations, stochastically chosen conformations, and a grid of conformations is presented. These trainings optimized cosine or CMAP [MacKerell et al., 2004b] corrections, and either used MM structures for MM and QM energy calculations or first re-optimized structures with QM before calculating QM energies. Additionally, parameterization was done against energies calculated in the presence or absence of implicit solvent. The effects of these choices on parameters are discussed and then evaluated in Ala_5, with the finding that most parameters presented overstabilize α-helices. After examining the trends in error versus length of the training compound, a protocol for extrapolating

smaller than alanine dipeptide, to a hypothetical alanine monopeptide, is presented. The monopeptide parameters performed comparably to ff14SB, against Ala_5 scalar couplings without empirical adjustment, and sampled secondary structure basins that better resemble the PDB in shape than ff99SB or ff14SB. Thus quantum mechanics is likely a viable target for training force fields that continue to improve consistency with NMR and PDB conformations.

3.3.1 An "optimal point charge" water model improves Ala_5 scalar couplings

The goal is to improve the dynamics of polypeptides. But to begin, a new, promising QM-based water model is tested. As discussed in Chapter 1, water is essential to polypeptide structure. The water models most commonly applied to biomolecular studies are rigid explicit water models of the three-point [Jorgensen et al., 1983, Berendsen et al., 1987] and four-point [Horn et al., 2004] flavors. Often these models start with atomic-centered charges. An alternate approach, called "optimal point charge" (OPC), begins with a Lennard-Jones center on the oxygen, but then places three point charges wherever they best reproduce electrostatic properties, resulting in better reproduction of bulk properties and hydration free energies than other rigid water models of similar complexity [Izadi et al., 2014].

OPC may very subtly destabilize helices relative to TIP3P and thus enhance ppII populations (Figure 3.2). As a result, the lowest χ^2 of 0.85 ± 0.02 is obtained for ff14SB in OPC solvent with the Orig parameters, compared with 0.90 ± 0.02 for ff14SB in TIP3P with the Orig parameters. These differences are not great when considering uncertainties. But ff14SB was empirically optimized in the context of TIP3P. Given the physical basis of the OPC water model, it is reasonable that it may improve the prediction of various properties, in this case agreement with three-bond scalar couplings. Thus OPC will be considered as methods are evaluated for alanine.

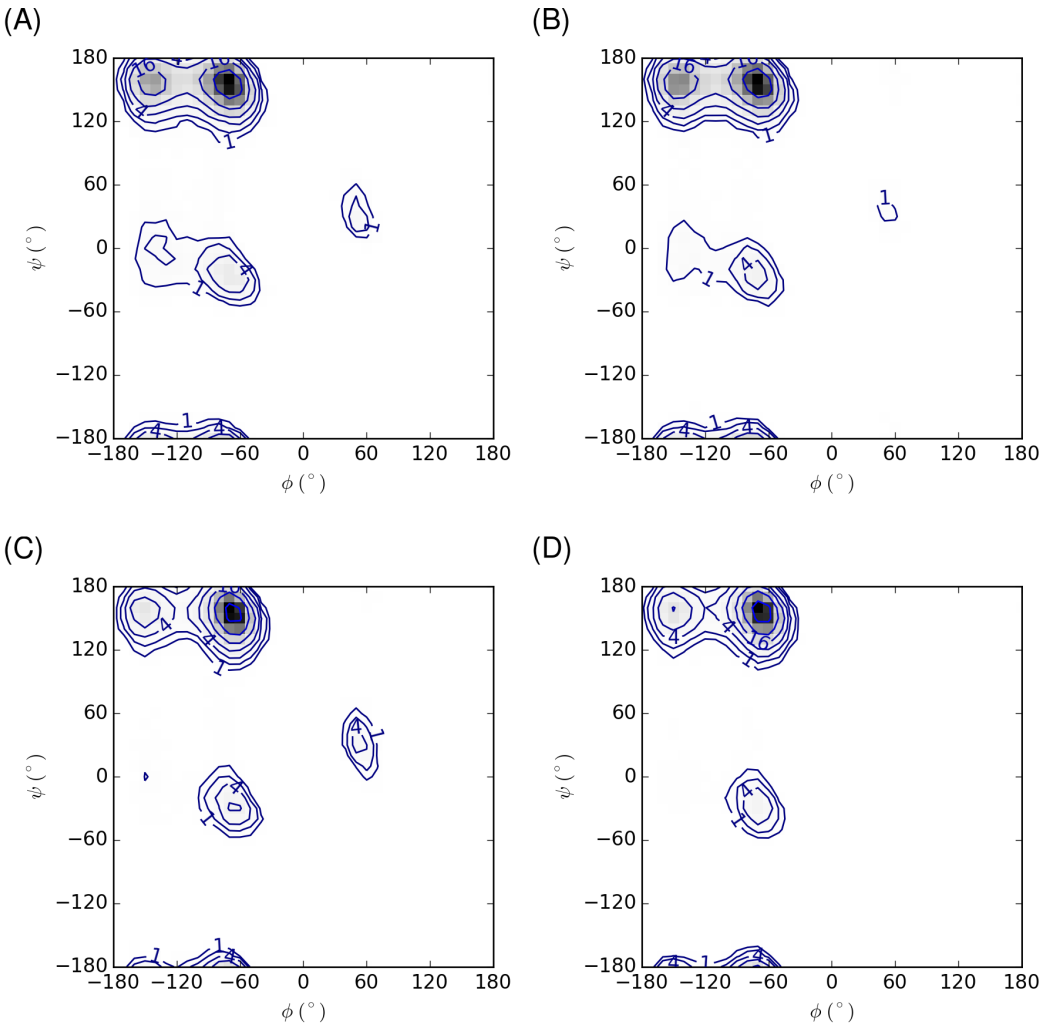

Figure 3.2: Ramachandran profiles of the second residue of Ala_5 in simulations with ff99SB and TIP3P (A) or OPC (B), as well as with ff14SB and TIP3P (C) or OPC (D).

3.3.2 Training conformations selected from high-T simulation

The first sets of structures for training were generated by simulating Ala_3 with N-terminal acetyl and C-terminal N-methyl caps (Ala tetrapeptide), to generate structures with the diversity that would be necessary for training backbone dihedrals. Two 100 ns simulations, beginning from α $(-60°, -45°)$ or linear $(180°, 180°)$ conformations, generated a total of 100 000 000 conformations, the histogram of which is shown in Figure 3.3A. Although the ff14SB force field was optimized for alanine [Maier et al., 2015], the ff99SB force field [Hornak et al., 2006] was employed here, as it offered a more even distribution of structures across the β to ppII transition. Although this may be unphysical, it facilitates the generation of varied structures for training.

Choosing 500 structures that minimized Equation (3.1) yielded the distribution of structures shown in Figure 3.3B, hereafter referred to as the *sim* structures. To compare against the ff99SB training, where HF/6-31G* minima were used, each of these conformations was minimized using HF/6-31G* in vacuo. Removing duplicates, there were the 151 minimum energy conformations shown in Figure 3.3C, hereafter referred to as the *min* structures. These overlap fairly well with the 51 minimum energy structures of the ff99SB training set, but are nearly three times as numerous. These sampling options are contrasted with conformations on a 24×24 grid, spaced every 15° in ϕ/ψ, hereafter referred to as the *grid* structures. The grid structures can be for peptides of any length. For dipeptides, grid simply refers to a grid across ϕ and ψ. For tetrapeptides, all residues were kept structurally homogeneous, i.e., all residues were given the same ϕ and ψ. Although this means the grid structures are not true grids across all relevant dihedral space, where some conformations may have residues with different (ϕ, ψ), the grid captures the all-helical and all-extended conformations and simplifies conformational space 331 776-fold (576 conformations, rather than 191 102 976 for the tetrapeptide, spaced every 15° in ϕ and ψ). Energy analysis and fitting of parameters to the sim, min, and grid structure sets are described below.

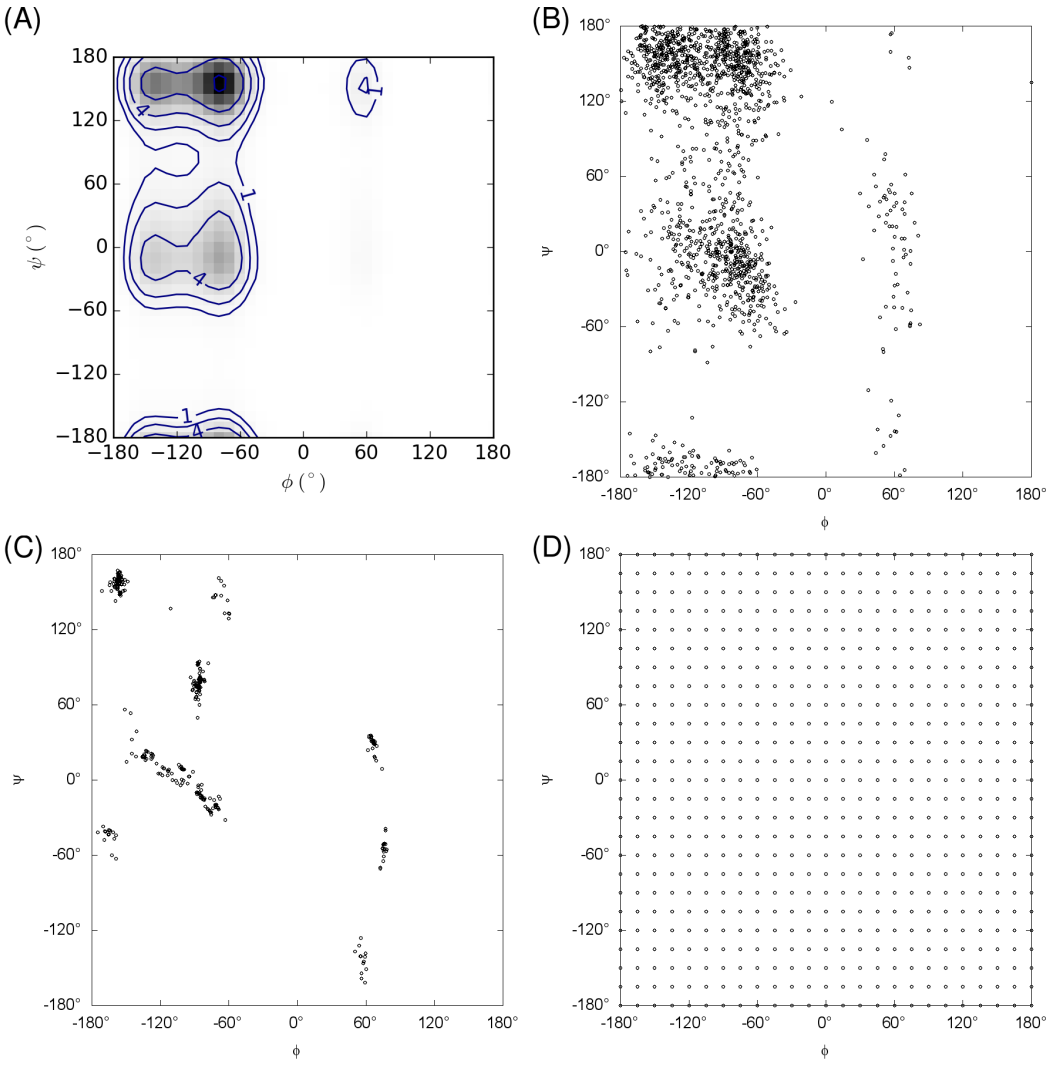

Figure 3.3: (A) Histogram of simulation of Ala_3 in GB-Neck2; (B) 500 Monte Carlo-selected pool of structures from Ala_3 simulation; (C) 151 unique in vacuo HF/6-31G* optimized conformations of Ala_3 from Monte Carlo structures; (D) 576 grid structures, spaced 15° in ϕ and ψ.

3.3.3 Energies of grids

Although it is difficult to visualize energy differences for irregular sets of conformations, like those of the sim and min data sets, the energies for the grid conformational sets were plotted, including each scheme for calculating energies. The vacuum energy surfaces for alanine dipeptide (A1) and tetrapeptide (A3) are shown in Figure 3.4 and Figure 3.7, respectively. Unsurprisingly, the A1 vacuum surface does not have an α-helical minimum, whereas the A3 vacuum surface does.

On the other hand, the solvated A1 and A3 energy surfaces are shown in Figure 3.5 and Figure 3.8, with MM solvation by PB [Gilson et al., 1993] and QM solvation by COSMO [Klamt and Schuurmann, 1993]. One potential issue in the COSMO treatment can arise from small molecular charge outside the solute cavity. To overcome the inconsistency of this charge, an outlying charge correction can be added to the potential energy. Two additional figures show the QM energy surfaces with the COSMO outlying charge correction, in Figure 3.6 and Figure 3.9. The implicit solvation effects result in A1 possessing an α-helical minimum according to all energy surfaces. In the helical conformation, the orientation of the NH and CO groups establishes a dipole. Polar solvents, like water, can polarize in response to this dipole, resulting in a stable α conformation even without a helical hydrogen bond. At least one historical weakness of using dipeptides in QM-based training—that they have no gas-phase minimum—is thus alleviated when solvation effects are considered.

The most obvious difference between the QM and MM profiles is that the MM profiles all strongly destabilize conformations with $\phi \approx 120°$, whereas the QM profiles destabilize this region less strongly. As this region is rarely sampled in simulations or the PDB [Lovell et al., 2003], having errors in this region may not be problematic in most cases, but is likely still worth addressing. Particularly, this region may play a role in the kinetics of sampling the α_L conformation.

Other differences relevant to dynamics pertain to the helical and ppII energy basins. Notably, the α basin stretches upward into 3_{10} helical conformations in the MM profiles and beyond, whereas the QM profiles tend to suggest

that the helical region should be more diagonal. Additionally, the QM profiles generally suggest the ppII region is too stable with ff14SB, but by how much ranges from $\sim$0.5 kcal mol^{-1} when QM energies are calculated for MM structures to $\sim$1 kcal mol^{-1} when QM energies are calculated for QM structures. To examine the effects of these differences on simulations, corrections first must be made to calibrate the MM energy surface to the QM energy surface for each energy calculation setup.

3.3.4 Dihedral corrections from each training set

First, cosine-based dihedral corrections were solved against the sim, min, and grid tetrapeptide datasets described above. One result was intuitive: minima tell nothing about transition energies. This is illustrated for cosine fits to energy differences for min (Figure 3.10) and sim (Figure 3.11) structure sets. The fits to QM energies of minimum energy structures stabilize the barrier in the center of the Ramachadran, continuing in a stripe across ψ near $\phi = 0°$, of between 5 and 12 kcal mol^{-1} relative to β. Additionally, some fits suggest that ϕ near 120° should be stabilized, whereas one fit (including implicit solvent with QM-reoptimized structures, pane C) suggests that the correction near $\phi = 120°$ should be destabilized by more than 6 kcal mol^{-1} relative to the average energy difference. Although one cannot say that any of these changes are reasonable or unreasonable based only on the energies of the minimum energy structures, it is clear that the structures in the training set do not prescribe such changes in the regions just discussed. The degree of variation in each of the corrections is also alarming, as in some cases there is up to $\sim$12 kcal mol^{-1} difference in the suggested relative energy change. I therefore no longer consider minimum energy structures on their own, as the parameter training landscape they provide is underdetermined.

Training with the sim set of structures resulted in much more similar changes for different energy calculation schemes. None of the sim fits indicated that any strong stabilization of ϕ near 0° is warranted. What's interesting is that the sim structure set possesses only 5 structures actually within

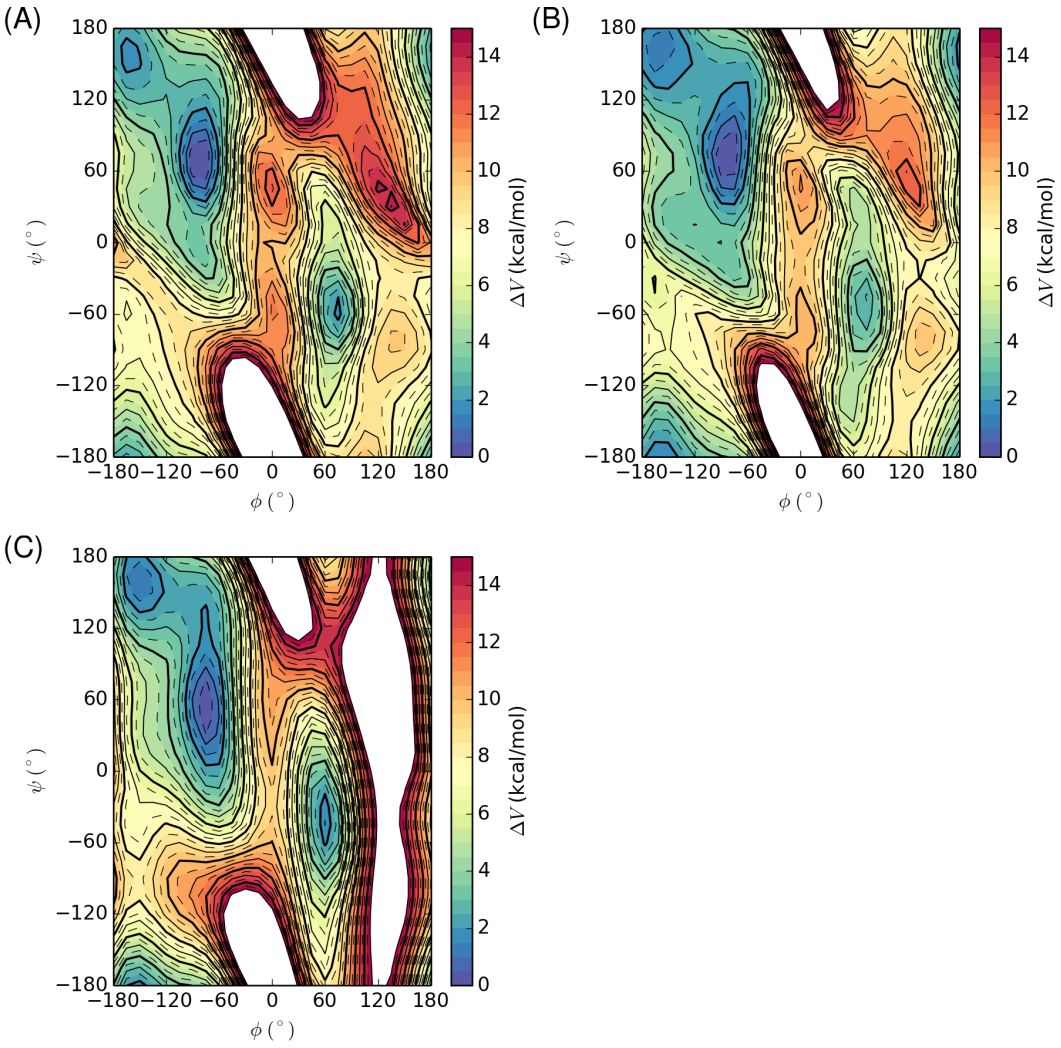

Figure 3.4: Ramachandran backbone energy surfaces for A1 in vacuo according to (A) QM energies of MM-optimized structures, (B) QM energies of QM-optimized structures, and (C) MM energies of MM-optimized structures. Solid, labeled contours indicate integer energy values in kcal mol^{-1}, whereas dashed contours indicate half-integer energies.

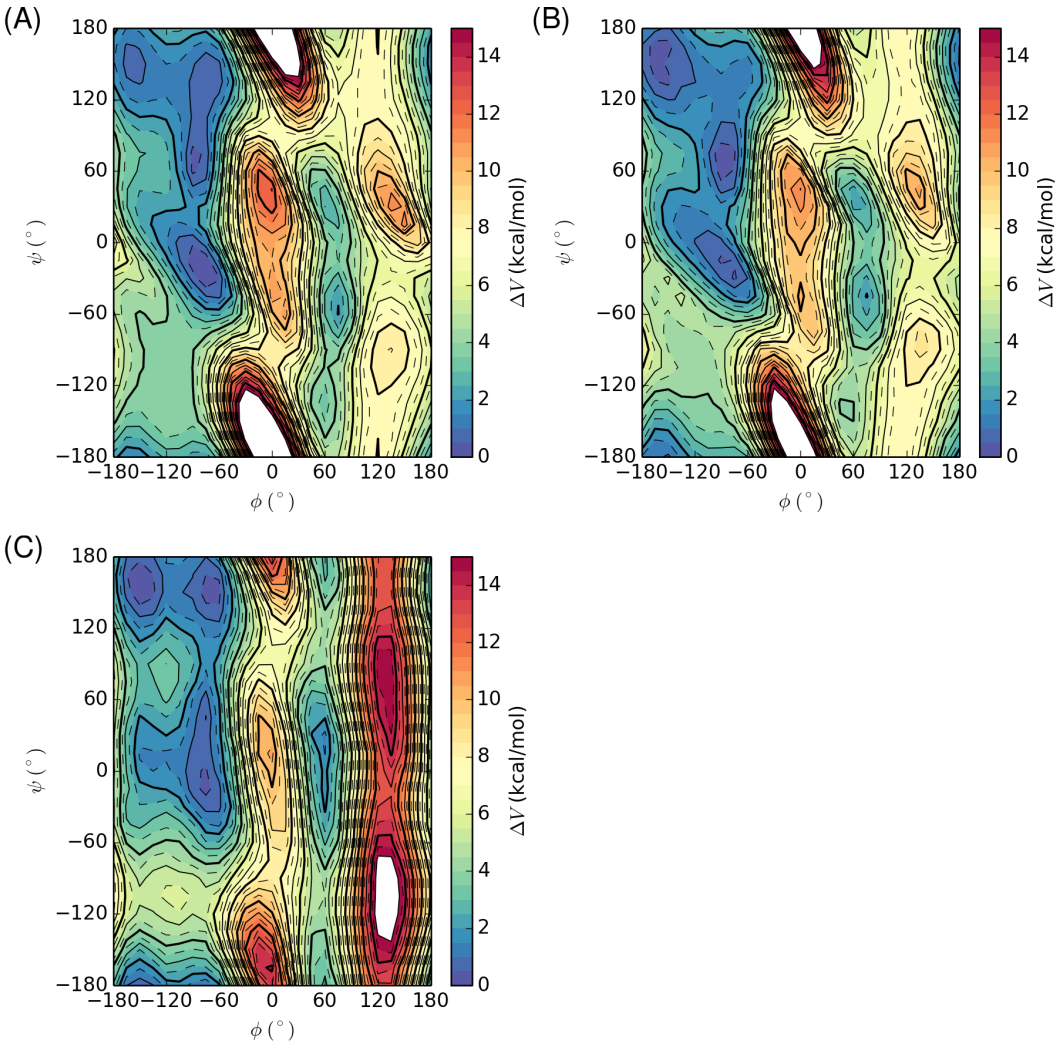

Figure 3.5: Ramachandran backbone energy surfaces for A1 in water according to (A) QM(COSMO) energies of MM-optimized structures, (B) QM(COSMO) energies of QM-optimized structures, and (C) MM(PB) energies of MM-optimized structures. Solid, labeled contours indicate integer energy values in kcal mol^{-1}, whereas dashed contours indicate half-integer energies.

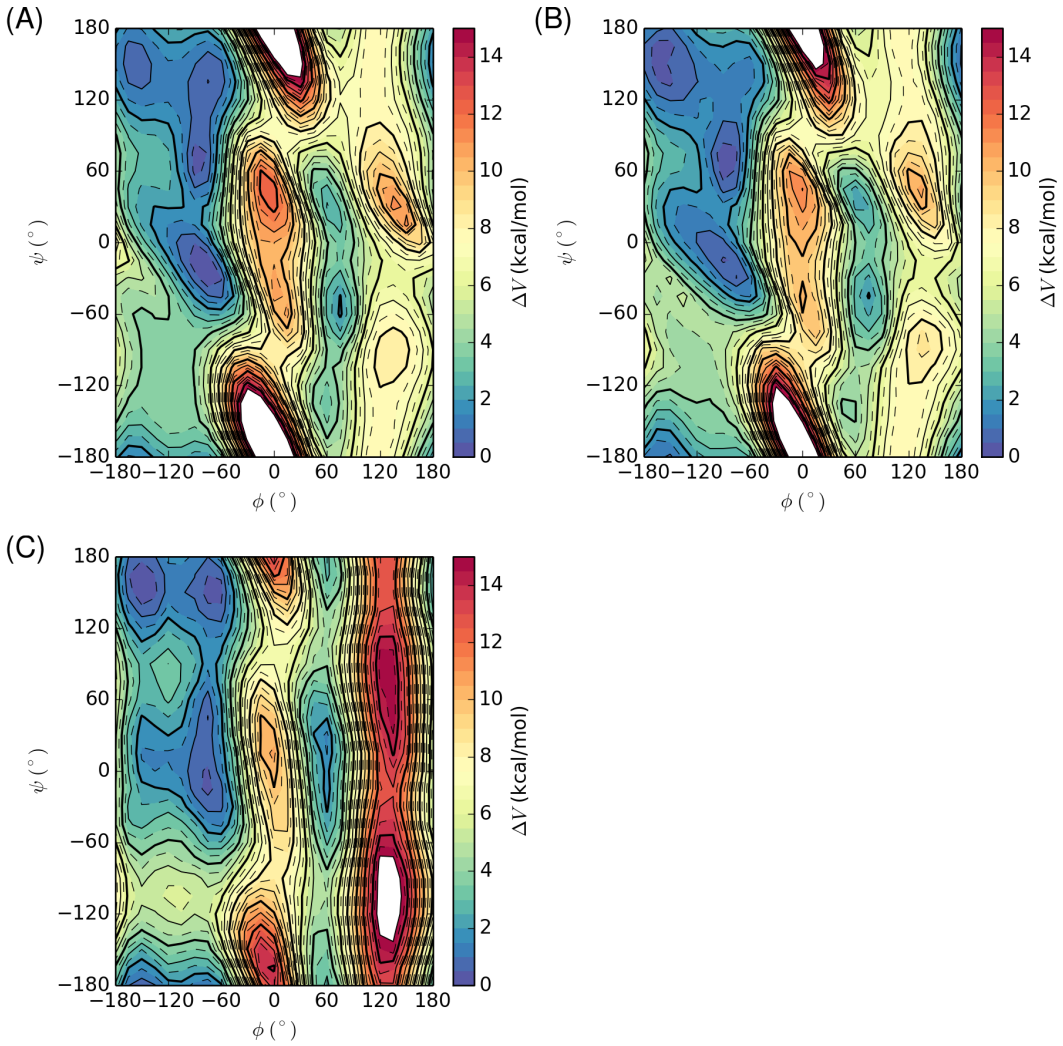

Figure 3.6: Ramachandran backbone energy surfaces for A1 in water, with the outlying charge correction (OCC) included for COSMO calculations, according to (A) QM(COSMO+OCC) energies of MM-optimized structures, and (B) QM(COSMO+OCC) energies of QM-optimized structures. Repeated here, for comparison, are (C) the MM(PB) energies of MM-optimized structures. Solid, labeled contours indicate integer energy values in $\mathrm{kcal\,mol^{-1}}$, whereas dashed contours indicate half-integer energies.

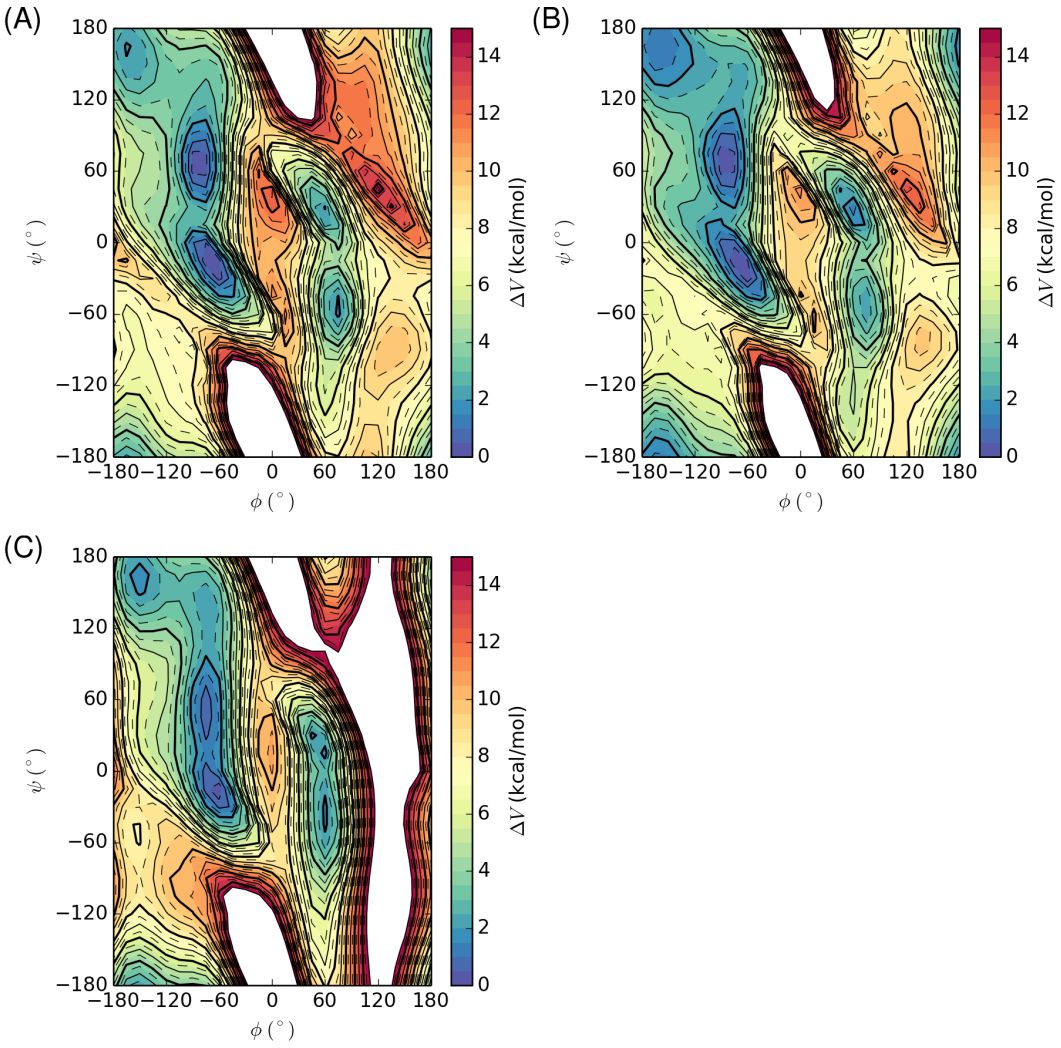

Figure 3.7: Ramachandran backbone energy surfaces for A3 in vacuo according to (A) QM energies of MM-optimized structures, (B) QM energies of QM-optimized structures, and (C) MM energies of MM-optimized structures. Solid, labeled contours indicate integer energy values in $\mathrm{kcal\,mol^{-1}}$, whereas dashed contours indicate half-integer energies.

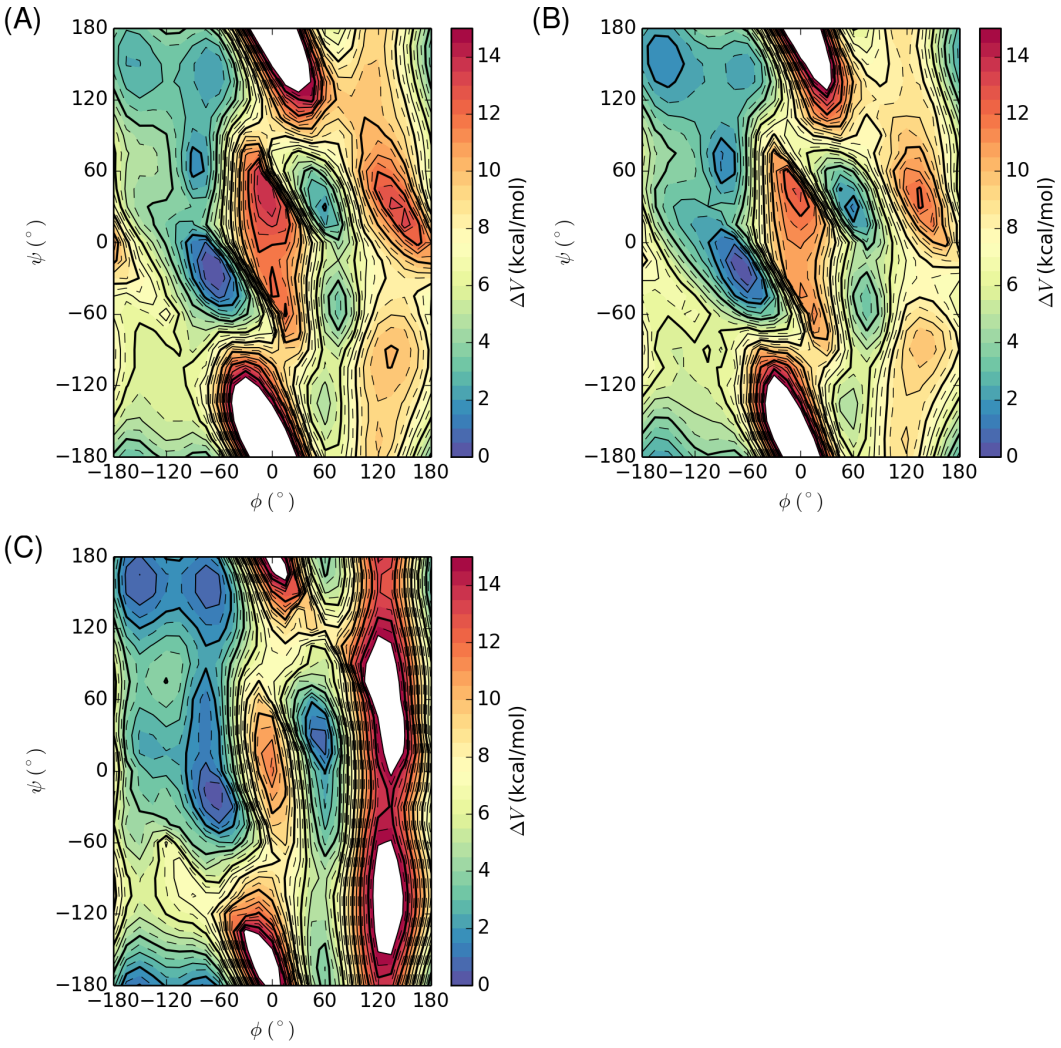

Figure 3.8: Ramachandran backbone energy surfaces for A3 in water according to (A) QM(COSMO) energies of MM-optimized structures, (B) QM(COSMO) energies of QM-optimized structures, and (C) MM(PB) energies of MM-optimized structures. Solid, labeled contours indicate integer energy values in kcal mol^{-1}, whereas dashed contours indicate half-integer energies.

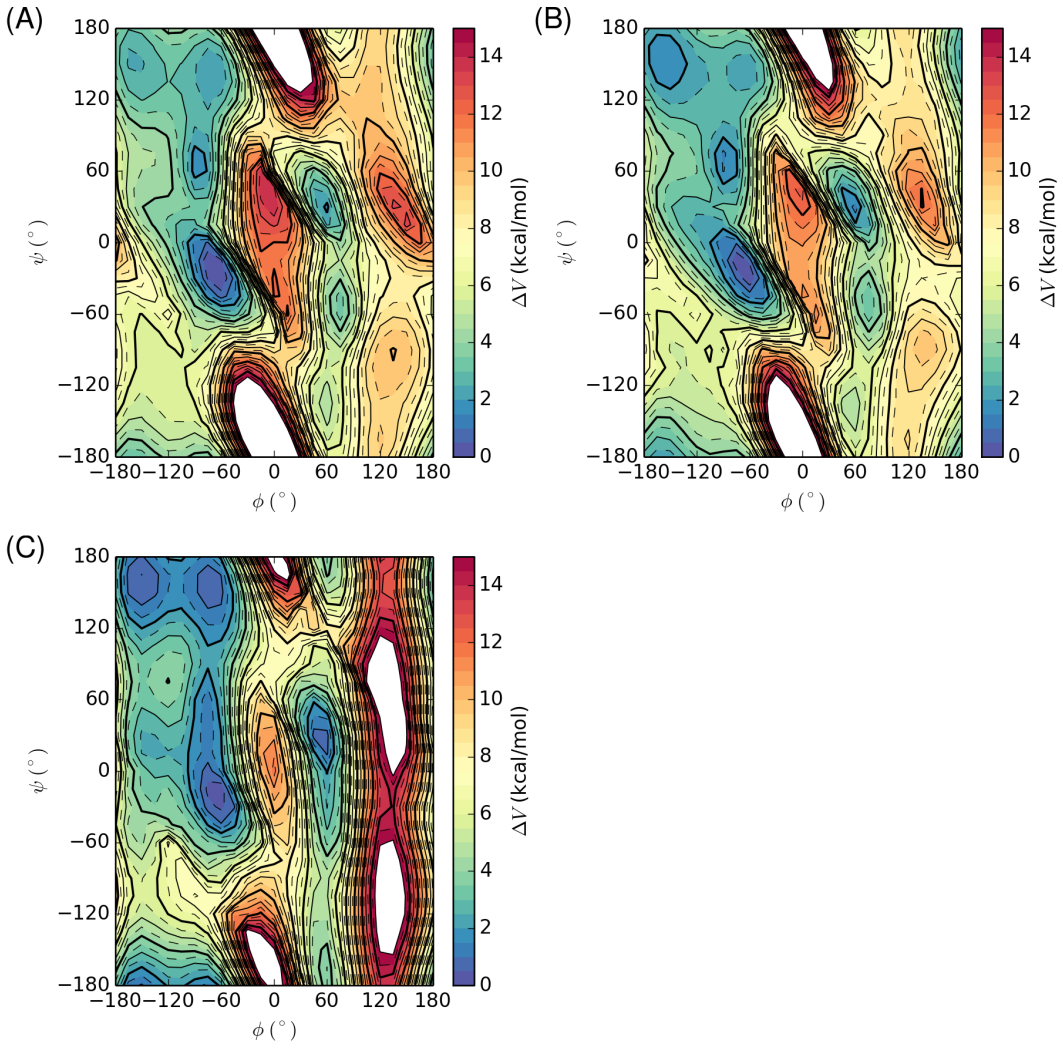

Figure 3.9: Ramachandran backbone energy surfaces for A3 in water, with the outlying charge correction (OCC) included for COSMO calculations, according to (A) QM(COSMO+OCC) energies of MM-optimized structures, and (B) QM(COSMO+OCC) energies of QM-optimized structures. Repeated here, for comparison, are (C) the MM(PB) energies of MM-optimized structures Solid, labeled contours indicate integer energy values in $\mathrm{kcal\,mol^{-1}}$, whereas dashed contours indicate half-integer energies.

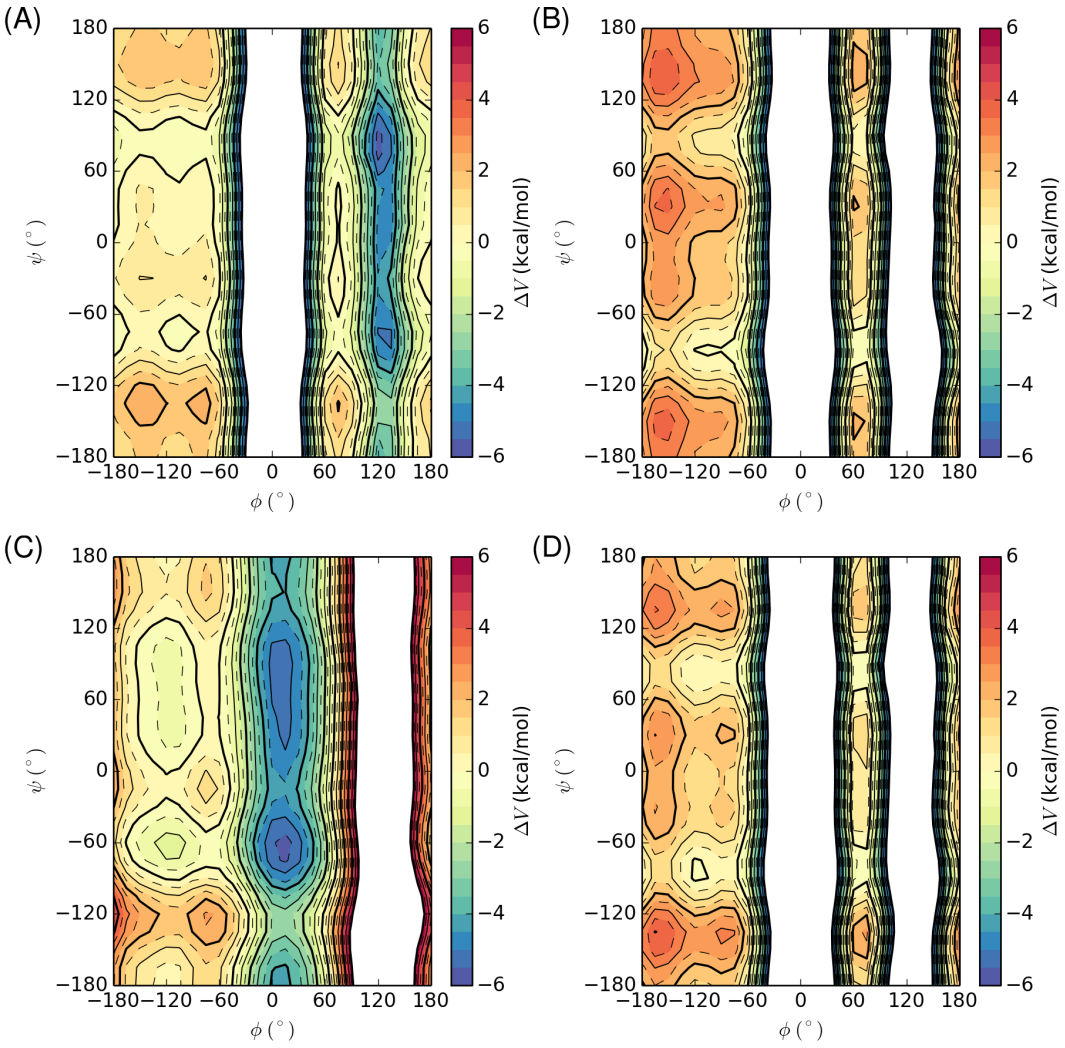

Figure 3.10: Effect of parameters fit to HF/6-31G* minimized structures (min) of alanine tetrapeptide, with energies calculated for MM-optimized structures in vacuo (A) or in implicit solvent (B), or with QM re-optimization of structures in vacuo (C) or in implicit solvent (D). Contour labels represent energy difference due to the parameters in kcal mol^{-1}.

$\phi \in [-30°, 30°]$, but with only 5 structures the energy correction in this region was constrained. An important difference between the min and sim sets of structures is that min has no structures within $\phi \in [-59°, 50°]$, while for sim this gap is $\phi \in [-21°, 5°]$. Thus the strong stabilization conferred by min parameters are likely artifacts of data scarcity.

Conversely, all sim-derived parameters suggested that ϕ around 120° should, in fact, be stabilized by about 6 kcal mol^{-1} relative to the average energy difference. This result is surprisingly consistent with three of the four min-derived parameter sets. One concern with the sim set is that there are no structures with $\phi \in [83°, 179°]$, and yet the correction is quite favorable in this region. What's interesting is that, although data are sparse in this rightmost region of the Ramachandran, this stabilization is common to all the sim fits performed. This was investigated further by fitting corrections to a more uniformly distributed set of structures with conformations in this region.

I considered grids of alanine tetrapeptide, where all three alanine residues were restrained to the same ϕ/ψ for each tetrapeptide conformation. The results of fitting these differences with cosines is shown in Figure 3.12. The feature of the sim corrections (Figure 3.11) where conformations with ϕ near 120° were stabilized exists in these grids, as well. Thus, this energy correction is not a result of poor sampling in that region in the sim training set. As it turns out, the issue is not that these new parameters overly stabilize that region; rather, the correction for ff99SB was arbitrarily high, as there were no data there in the ff99SB training. These conformations are not likely to be sampled frequently anyway, as they do not appear commonly in experimental protein structures [Lovell et al., 2003], but may be sampled even less with ff99SB than they already should be when, for example, transitioning to the α_L conformations.

The training quality with cosine fits to the sim and grid sets, however, is not very satisfactory (Table 3.1). The minimum root mean square error (rmse) achieved for these sets was 1.5 kcal mol^{-1}, for the sim structure set without QM re-optimization, both in vacuo and in the context of implicit solvent. The min structure set, that was poorly constrained, achieved rmses of less than 1 kcal mol^{-1}, but this is likely because good error can be achieved by fitting

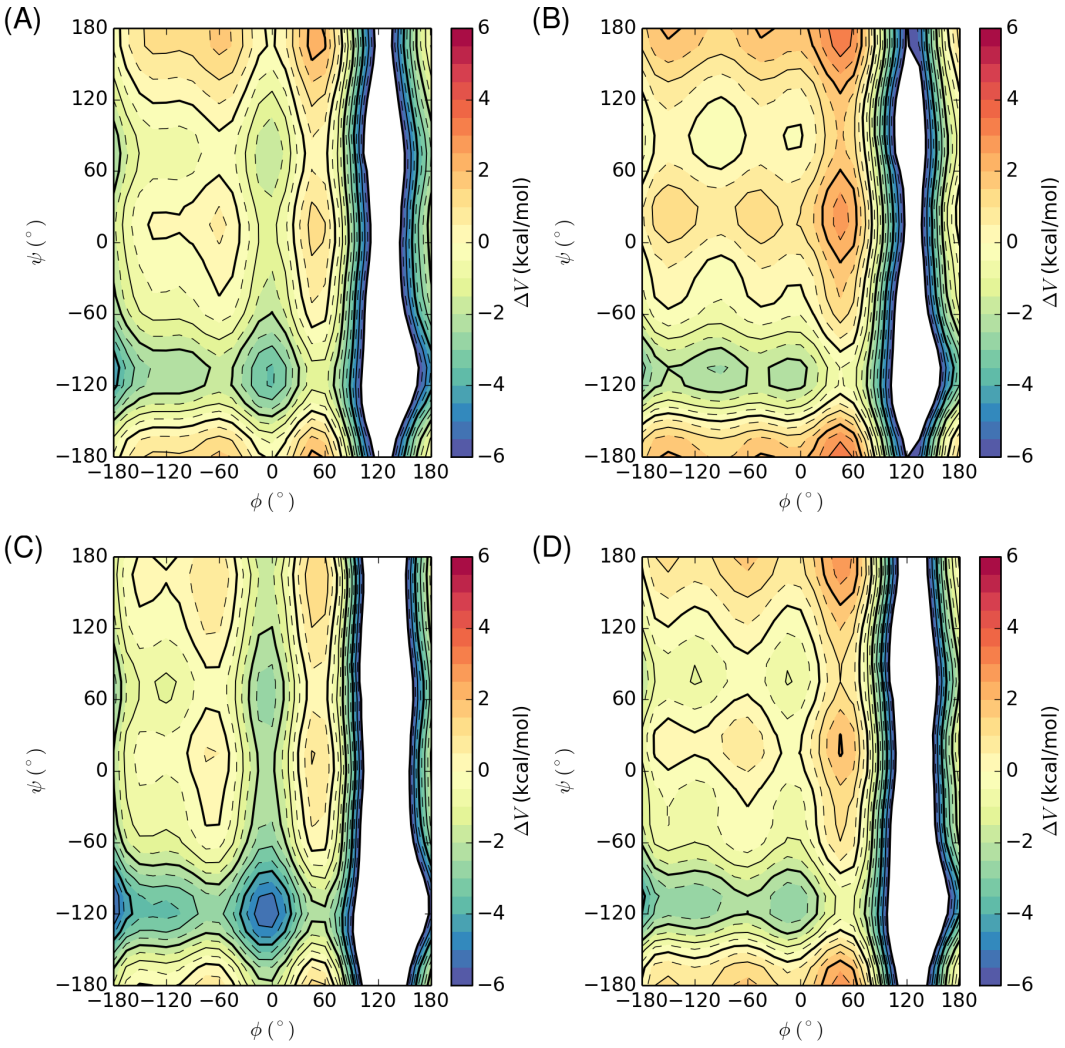

Figure 3.11: The correction profiles (least-squares optimized to reproduce QM − MM differences) fit to stochastically chosen simulation structures (sim) of alanine tetrapeptide, with energies calculated for MM-optimized structures in vacuo (A) or in implicit solvent (B), or with QM re-optimization of structures in vacuo (C) or in implicit solvent (D). Contour labels represent energy difference due to the parameters in kcal mol^{-1}.

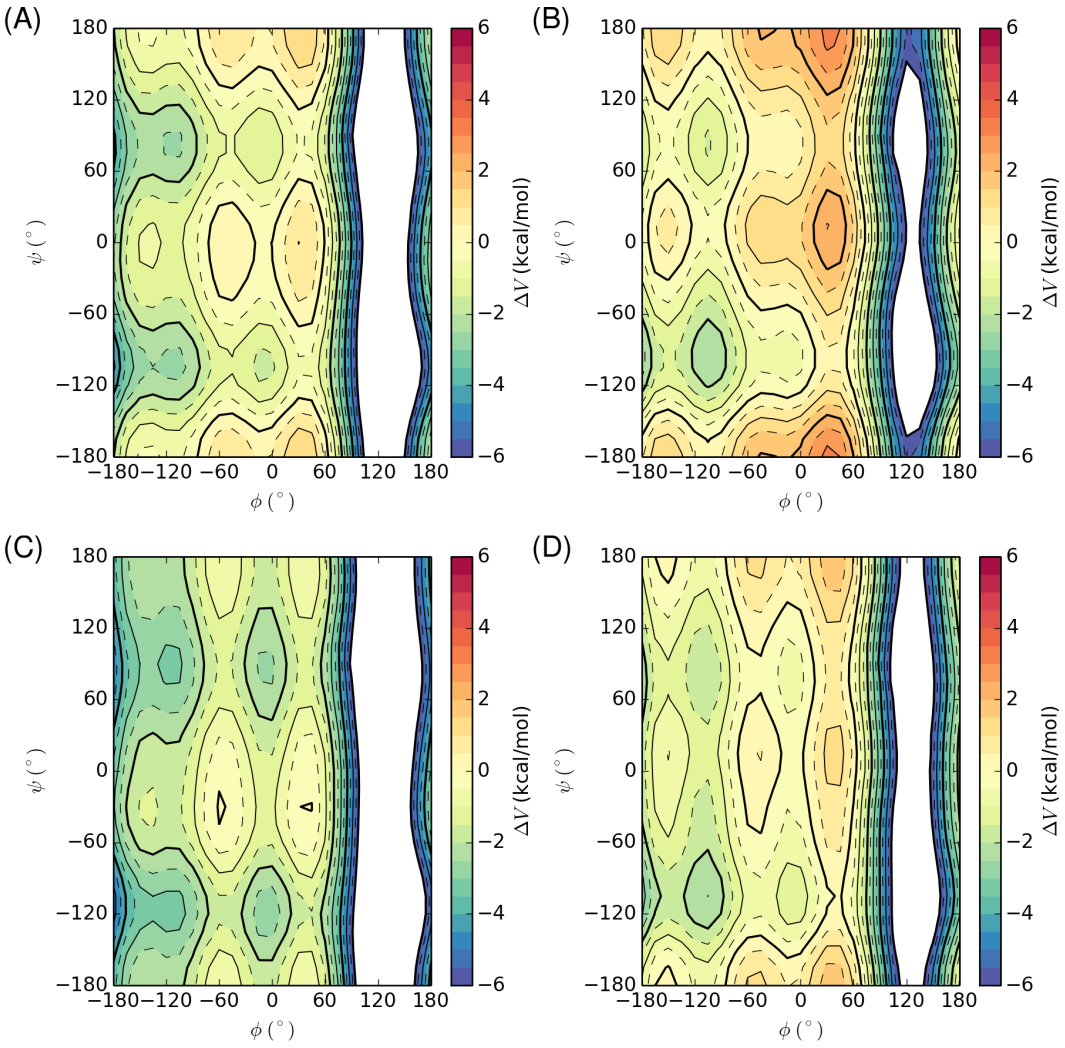

Figure 3.12: Effect of parameters fit to two-dimensional ϕ/ψ grids (grid) of alanine tetrapeptide conformations, with energies calculated for MM-optimized structures in vacuo (A) or in implicit solvent (B), or with QM re-optimization of structures in vacuo (C) or in implicit solvent (D). Contour labels represent energy difference due to the parameters in kcal mol^{-1}.

Table 3.1: Root mean square errors against Ala_3 QM energies with ff14SB and after fitting cosine dihedral parameters. rmse=root mean square error with ff14SB. new=root mean square error with new parameters

		min		sim		grid	
QM re-opt?	solvent	rmse	new	rmse	new	rmse	new
no	vacuum	2.0	0.6	2.2	1.5	13.0	3.6
no	water	3.1	0.8	2.3	1.5	11.9	4.6
yes	vacuum	1.9	0.9	2.5	2.0	12.7	4.5
yes	water	2.6	0.9	2.2	1.6	11.5	4.1

less of the Ramachandran surface.

The effects of these parameters were evaluated in Ala_5 simulations. These results are also concerning. None of the corrections derived thus far improved upon ff14SB, or even ff99SB, in Ala_5 simulations when compared to NMR scalar couplings (Tables 3.2 and 3.3). Two possibilities exist: either (1) the data have not been fit well enough, and with better quality fits, Ala_5 sampling should improve; or, (2) the data suggest trends that are problematic, unrelated to the quality of the QM fit. To identify patterns in errors in the underlying energy surface, more accurate corrections were created using a different functional form, and the effects on structure ensembles analyzed.

3.3.5 Training against tetrapeptides overstabilizes helices

To alleviate possible limitations in fitting that may arise from using uncoupled cosine corrections, I created CMAP corrections based on the QM − MM differences in the structure grids. As discussed, the CMAP can reproduce any two-dimensional energy surface exactly [MacKerell et al., 2004b]. Simulations of Ala_5 with the CMAPs derived from tetrapeptide fitting data resulted in considerable χ^2 errors of 3.98 ± 0.12 or higher with TIP3P solvent (Table 3.2) or 3.27 ± 0.17 or higher with OPC solvent (Table 3.3). This is surprising as tetrapeptides have been utilized in many force field training efforts, including ff99SB [Hornak et al., 2006], which achieves χ^2 of as low as 1.46 ± 0.02 with TIP3P, using Dft1 Karplus parameters. In the worst case, one of the new force

fields, derived against a grid of implicitly solvated, MM-optimized structures, achieved χ^2 of 9.04 ± 0.32 with the Dft1 Karplus parameters. A histogram of residue 2 in ϕ and ψ backbone dihedrals as sampled by this rogue force field is depicted in Figure 3.13. It is nearly entirely helical.

To examine why CMAPs based on the tetrapeptide data could be so detrimental, a CMAP was compared to cosine corrections trained to reproduce the same data set—a ϕ/ψ grid of solvated A3 energies of MM-optimized structures (Figure 3.14). From the CMAP (Figure 3.14B), which matches the underlying energy differences exactly, there is a clear stabilization near the α-helical region of the Ramachandran at about $(-60°, -45°)$, with destabilizations to the left and directly above. As the areas that need to be destabilized share the same ϕ (for the area above) or the same ψ (for the area to the left), fitting the α-helical energies together with these destabilizing features would be quite difficult with uncoupled cosine terms. The correction resulting from this fit is shown in Figure 3.14A, with the difference from the CMAP correction shown in Figure 3.14C. Although the cosine corrections capture some of the same overall features as the CMAP, it lacks the same helical stabilization that the CMAP can reproduce from the energy differences. If the energy differences in the source data are such that the helical region is overstabilized, this may actually result in more reliable energetics for cosine corrections that are tempered by the training data for nearby conformations. In other words, errors from ff99SB assumptions 4 (using cosine corrections) and 5 (using tetrapeptides) above may partly cancel in this case. In the context of a CMAP correction, however, one can fit the data too exactly, and the helical overstabilization in the training data will be captured too well to benefit simulations. Thus, the presence of a helical hydrogen bond in training may not be desired, when fitting corrections as exactly as allowed by CMAP.

Previous, successful force fields like ff99SB have been trained against tetrapeptide minimum energy structures [Hornak et al., 2006], whereas simulations with the poorly converged set of cosine-based corrections derived here against tetrapeptides have not been performed. It is thus difficult to assess what specific variations in the training procedures might have important effects in dynamics simulations. Training set sampling may play a significant role

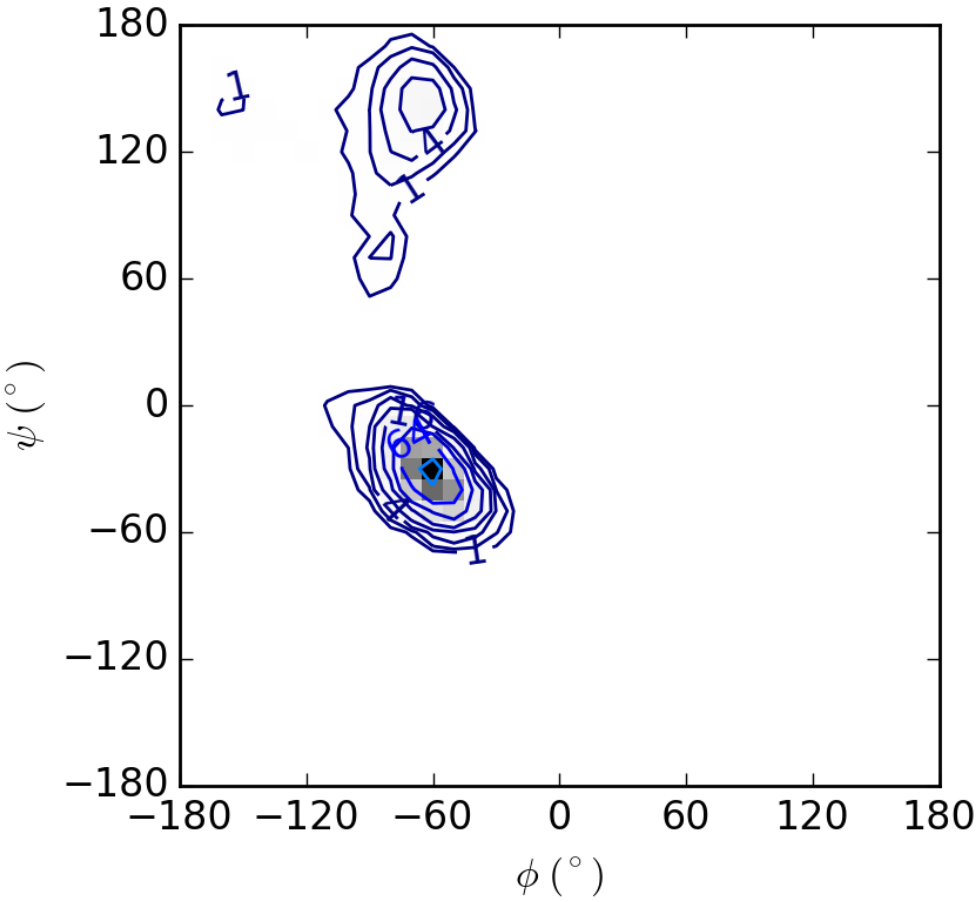

Figure 3.13: An impressively awful result. The second residue of Ala_5 is nearly entirely helical after training against QM energies for a grid of solvated MM-optimized tetrapeptides. Each square is colored from white for low population to black. Each overlaid blue contour represents a doubling in population, with 1, 4, 16, and 64 labels denoting how many times more populated each square is than if the same surface had a completely flat distribution. This simulation attained the enviable χ^2 of 9.04 ± 0.32, higher than any other simulation performed here (and hopefully anywhere else).

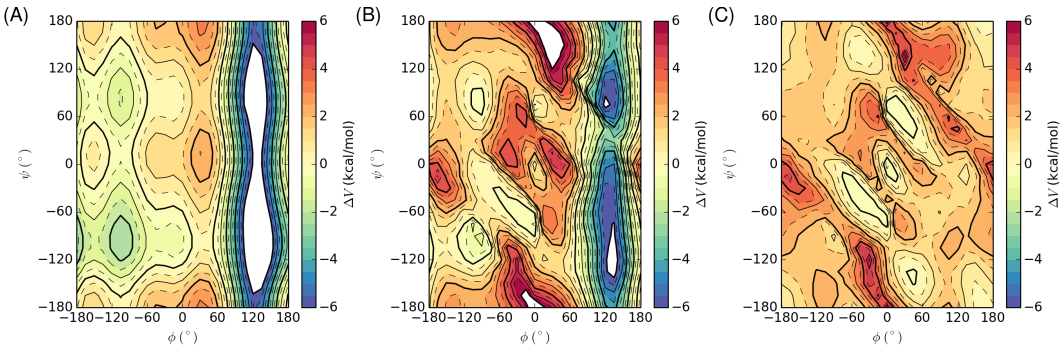

Figure 3.14: The corrections to an A3 ϕ/ψ scan, in the context of implicit solvent, with (A) cosine corrections and (B) a CMAP, as well as (C) the CMAP correction minus the cosine corrections.

in how the trained parameters stabilize different secondary structures. Imaginably, for example, training against a wide variety of structures including many transitions may limit the ability to fit minima. Or different assumptions employed in the ff99SB fitting to account for the limited set of conformations, such as assuming that the ϕ n=1 torsional term should have positive amplitude [Hornak et al., 2006], may also play a role. Regardless of other likely important differences, the cosine corrections here fit to tetrapeptide grid energies resulted in better agreement with experiment and closer behavior to ff99SB than fitting using CMAPs.

Although cosine-based corrections solved to reproduce Ala_3 energy differences may yield better results than CMAPs against Ala_5 scalar couplings by not completely fitting the quantum mechanics data, neither provides the level of performance that would be needed to replace ff14SB. Moreover, coupled ϕ/ψ corrections that can quantitatively reproduce any ϕ/ψ energy differences are still desirable to provide a more exact description of backbone dynamics. Thus I have pursued alternative means of generating data for training. In particular, I would like a method that stabilizes helices a little less than those examined thus far. One potential weakness in the use of tetrapeptides is that they allow the correction of errors in helical hydrogen bonds—errors that may be of a different magnitude in vacuum or when comparing COSMO and PB than in explicit solvent simulations. Of course, the hydrogen bond is only one aspect

of the helical conformation, in addition to the formation of a dipole that may have different effects in a QM model than in a fixed-charge MM model. Thus, I have turned to a model system without helical hydrogen bonding to examine whether it may provide a more reasonable training ground for solvated peptides.

3.3.6 Training against dipeptides stabilizes helices less

Dipeptides do not form helical hydrogen bonds. When, as here, correcting the helical hydrogen bond imparts too much stability to helical conformations in simulations, it may be because some error in the QM − MM comparison does not transfer precisely to a correction suitable for explicit solvent like TIP3P. As a result, it may actually be preferable to use smaller training compounds that cannot form hydrogen bonds. An alternative approach may be to incorporate both helical and interstrand hydrogen bonding in training to provide a balanced description. How to apply the latter to a dihedral correction is not trivial, however, as there is no guarantee that a single strand will have a neighbor. Moreover, if the error is in the α-ppII balance, it is not clear improving the α-β balance would help. It is additionally unclear whether fitting the errors in both α and β hydrogen bonds in QM-MM comparisons would even benefit the α-β balance in explicit solvent simulations, or whether the errors may include nontransferable artifacts.

Ultimately, I chose to first try the simplest option of applying the various methods described above to generating fitting data for alanine dipeptide (A1), where it is comparatively easy to map out the full 2D ϕ/ψ conformational space. As shown in Figures 3.4 to 3.9, the A1 energy differences stabilize the helical basin less than the A3 energy differences. These energy differences are explicitly presented in Figure 3.15, where QM is calculated with COSMO and the OCC for the MM-optimized structures, and compared to MM with PB.

The effect of this decreased helical stability can be seen in simulations of Ala_5. With decreased helical populations, the χ^2 values also decrease, in some cases by more than half (for the solvated MM parameters according to all sets of Karplus parameters, in both TIP3P and OPC solvents). This result

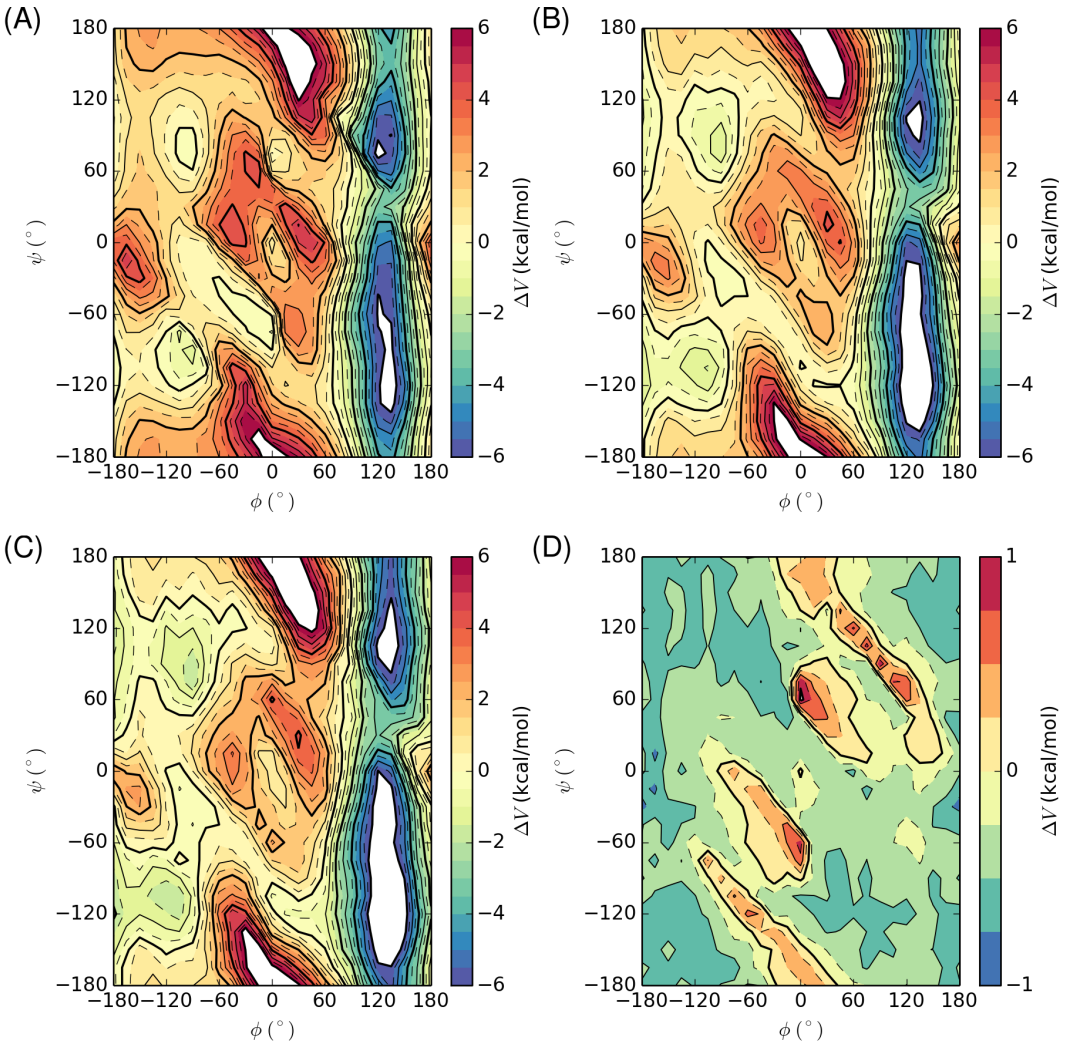

Figure 3.15: The QM-MM differences, with QM in the context of COSMO (with the outlying charge correction), and MM in the context of PB, for (A) alanine tetrapeptide, (B) alanine dipeptide, and (C) alanine "monopeptide." Also shown are (D) the differences from the dipeptide difference map (B) to the "monopeptide" difference map (C). From the tetrapeptide to the dipeptide, and on, the differences stabilize α-helix less.

is interesting because one would intuit that Ala_3 should be more appropriate for Ala_5 than Ala_1 would be, based on more similar peptide length. But the situation is much more insidious.

Whereas some of the A1 parameters achieved lower χ^2 than ff99SB (Tables 3.2 and 3.3), still none of them approached the accuracy of ff14SB. Examining histograms of residue 2, these parameters still generate a significant quantity of α-helix. This helical stability may arise from weaknesses in comparing the different electrostatics or solvent effects between MM and QM, perhaps concerning the stabilizing effect of the helical dipole, that persist even at the scale of A1. To investigate this, I examined trends in error compared to QM versus peptide length.

3.3.7 Length-independent parameters

Based on the previous sections, it appears that there is a component of the per-residue error between QM and MM energies that increases with the length of the peptide. This error manifests as QM helical stability that's missing from ff14SB with A3, but that is missing less from ff14SB with A1. This shows that the per-residue error depends on the number of residues. Imaginably, there is also some component of the per-residue error that does not depend on the number of amino acids. Or there may be an optimal length for the training compound to alleviate errors that result from artifacts in the training, and this may be smaller than A1. Of course, A1 is the smallest size that can be achieved in real calculations, so to get any smaller would require extrapolation from real data on A1 or larger.

There are (at least) two ways to think about extrapolation to a length smaller than A1:

1. Extrapolating out effects from having peptide groups in the training; or

2. Extrapolating out effects from having amino acids in the training.

In either case, we would first want to get the slope for the change in the correction over the change in the number of peptides (case 1) or amino acids

(case 2). Mathematically, this can be described as in Equation (3.10) for the case of extrapolating peptides or as in Equation (3.11) for the case of extrapolating amino acids, where AA stands for amino acid and V is the potential energy for Ala_3 (tetrapeptide) or Ala_1 (dipeptide).

$$\frac{\Delta V}{\Delta\text{peptide}} = \frac{V_{\text{tetrapeptide}} - V_{\text{dipeptide}}}{4 \text{ peptides} - 2 \text{ peptides}} \tag{3.10}$$

$$\frac{\Delta V}{\Delta\text{AA}} = \frac{V_{\text{Ala3}} - V_{\text{Ala1}}}{3 \text{ AAs} - 1 \text{ AA}} \tag{3.11}$$

We could extrapolate to a per-residue correction assuming only one peptide group. If this correction needs to be in a protein where the ratio of peptide groups to residues is ~ 1, then it would make sense to add a correction to each residue that only accounts for one peptide group. This extrapolation to the "monopeptide" level would be:

$$V_{\text{monopeptide}} = V_{\text{dipeptide}} - \frac{\Delta V}{\Delta\text{peptide}} \times \text{peptide} \tag{3.12}$$

Alternatively, if we want to extrapolate to the correction that, assuming a linear trend in QM $-$ MM artifacts per amino acid, would remove the artifacts of having amino acids, we would want to extrapolate to zero amino acids:

$$V_{\text{Ala0}} = V_{\text{Ala1}} - \frac{\Delta V}{\Delta\text{AA}} \times \text{AA} \tag{3.13}$$

What's neat is that both of these equations involve exactly the same calculations. The critical difference is the concept: either the correction is for a system where the number of peptides is extrapolated to 1, or the correction is for a system where the presence of amino acids is extrapolated away. The former makes sense in terms of the essentials of molecular mechanics, not just considering possible artifacts in the QM $-$ MM comparisons. Likely, there's some combination of factors at play.

Another way of conceptualizing this, is that the derivation of A0 parameters serves to subtract off the component of the per-residue energy error that fluctuates with the number of residues or peptides. In other words, what is being removed is the component of the error that is not linear with peptide

length. One may arrive at the conclusion that this fluctuation in the per-residue error is cooperativity—something that could be appropriate when simulating peptides and proteins. But if this were the case, then it is surprising that including the cooperativity of A3 and even of A1 could result in such egregious behavior at the level of A5, where cooperativity should be greater than in A3 or A1.

More likely, these nonlinear errors may result from differing treatments of nonbonded interactions between the QM and MM models in a vacuum or in the context of different implicit solvent models. These discrepancies would not be inherent to alanine and may differ in explicit solvent. A simplistic hypothesis is that subtracting off these nonlinear additions would result in length-independent alanine backbone parameters that ought to depend less on the exact details of how energy comparisons are performed.

Of course, there are some limitations to this approach. Firstly, some would argue that larger fragments are more appropriate and that moving to an A0 training set is the opposite of what is desired. Force fields, as might be argued, should incorporate the cooperativity that the quantum mechanics may be exhibiting. I would counter, as argued above, that these calculations have not only true cooperative effects that, hypothetically, could yield a more robust force field, but also contain artifacts resulting from imperfect treatment of electron correlation, basis set incompleteness and superposition errors, different QM/MM charge/solvent modeling, and the different preferred geometries in the QM/MM comparison. Thus what's removed when one extrapolates to A0 is not necessarily cooperativity that ought to be captured. Whether it is better to subtract both cooperativity and such artifacts, or to maintain both, remains to be shown below.

Secondly, assuming that one would want to attain A0 parameters, there's no reason to expect that the extrapolation should be linear through A3 and A1. I would concede that this may be true. One could evaluate the issue with scans on A2, testing whether there even exists any simple trend from A1 to A3. In the present work, however, we simply assume that the per-residue differences being subtracted off are those that are linear with peptide length. We test the effects of this assumption later, and it turns out to be quite productive.

And thirdly, performing tetrapeptide and dipeptide backbone potential energy scans may be feasible for alanine, which is so small, but would become very expensive with larger amino acids that also have multiple side chain rotamers to consider. Thus, this method would appear to be ad hoc in the worst sense, in that it's not even necessarily generalizable. I suggest, however, as will be tested, that the trend in length-dependent differences for alanine may not be altogether that different for other amino acids. There may be some physical justification for expecting the $\mathrm{A0} - \mathrm{A1}$ differences to transfer to other amino acids if one considers the view that the extrapolation moves from the dipeptide to the monopeptide, and all amino acids form peptide bonds. This assumption would imply that the principal nonlinearity in the energy differences with respect to peptide length results from the hydrogen bond or the dipole of helical conformations, rather than the identity or presence of side chains. Still, extension to other residues may be a somewhat heuristic exercise where the Ala offset derived here is assumed to transfer, applied, and then tested. A more rigorous treatment could include extensive QM calculations on all other amino acids for which backbone corrections are to be derived. Here I have only considered the former.

Ideally, A0 parameters would result in good agreement with A5, indicated by low χ^2 values on the order of ff14SB's 0.90 ± 0.02 in TIP3P with the Orig Karplus parameters. Such agreement would suggest that using quantum mechanics, with a few tricks, like implicit solvent and an extrapolation that has some physical meaning, can produce parameters that are appropriate for simulations of peptides, and perhaps proteins, in aqueous solvent. As expected, the helical destabilization in the A0 energy corrections relative to A3 and A1 corrections resulted in the lowest χ^2 accessed by QM-derived parameters.

The A0 correction for one combination of methods that performed most similarly to ff14SB in A5 simulations—using MM structures for energy calculations without re-optimization (MM) in the context of water, including the COSMO outlying charge correction (water*)—is displayed in Figure 3.15. Pane C shows the A0 CMAP itself, whereas Pane D shows the difference between the A1 CMAP (Pane B) and the A0 CMAP (Pane C). The differences from A1 to A0 are small, less than $1\,\mathrm{kcal\,mol^{-1}}$ in magnitude. The notable feature

is that these differences destabilize the α-helical region of the Ramachandran, which is an expected result given the trend from A3 to A1.

The example shown in Figure 3.15C, the A0-grid-MM-water* parameters, had a χ^2 of 1.05 ± 0.02 in TIP3P solvent, or 0.96 ± 0.04 in OPC solvent, in both cases using the Orig Karplus parameters. Histograms of ϕ and ψ of the second residue are shown in Figure 3.16, where panes (A) and (B) illustrate A0-grid-MM-water* parameters in TIP3P and OPC, respectively. Analogous results, only for A1 parameters instead of A0 parameters, are shown in panes (C) and (D).

Notably, the four Ramachandran profiles in Figure 3.16 exhibit quite similar shapes for the α, β, and ppII basins. Importantly, these shapes match those illustrated by the PDB Figure 3.1 more closely than ff14SB. To make this comparison easier, Figure 3.17 contains just the Ala PDB histogram with histograms of the third residue in Ala_5 simulated with ff99SB, ff14SB, and A0 parameters. In particular, there is a more diagonal helical conformation with anticorrelation between ϕ and ψ, and the sampling of conformations below ppII on the Ramachandran map ($\phi, \psi \approx -75°, 75°$). This result is significant as this means that the details in backbone preferences found in experimental structures can also be obtained from quantum mechanics.

The difference in the four Ramachandran profiles in Figure 3.16 is that the simulations derived for A0, as expected, exhibit less α-helix. Presumably, further tweaking of the A0-MM-water* parameters, as was done for ff14SB, could result in slightly closer agreement with A5 scalar couplings, but the Karplus parameters may be limiting in this comparison. Having a method that can result in performance better than ff99SB and similar to ff14SB, I instead turn to the point of this exercise. The method must be able to work not only for alanine, which already agreed quite well with scalar couplings when using ff14SB. I apply the method to generate Val_0 (V0) parameters, using the A0 − A1 offset, as well as test whether the new alanine parameters can reproduce the preferences suggested by experiments for valine.

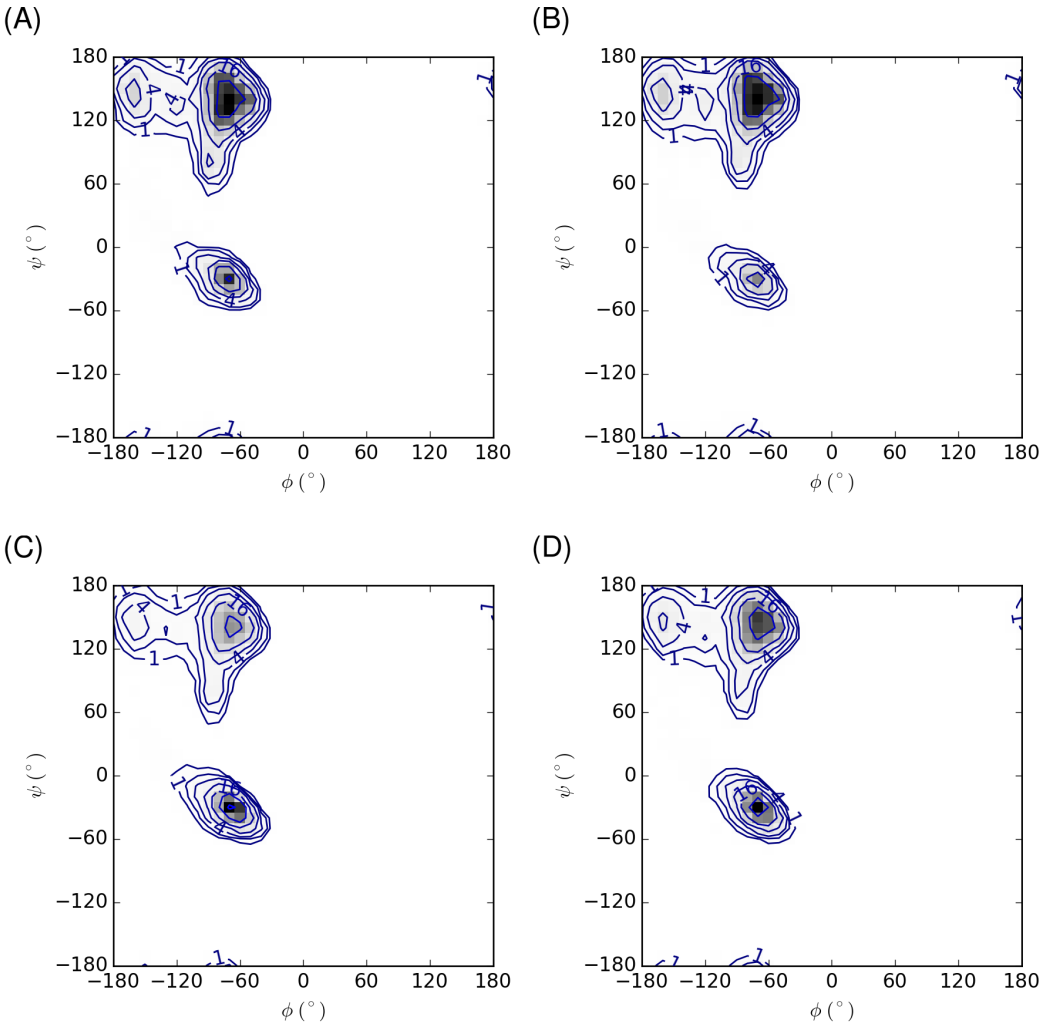

Figure 3.16: Ramachandran histograms of the second residue in A5, using force fields trained against MM-optimized structures in the context of water*, with extrapolation to A0, simulated in (A) TIP3P and (B) OPC, and without extrapolation (using A1 parameters), simulated in (C) TIP3P and (D) OPC. The most lucid difference is that the A0 plots in (A) and (B) exhibit greater sampling of ppII conformations, indicated by darker shading.

Table 3.2: χ^2 deviations from experimental scalar couplings [Graf et al., 2007] for simulations of Ala_5 in TIP3P solvent, according to Orig, Dft1, and Dft2 sets of Karplus parameters (described in text). Vacuum indicates training in a vacuum, as expected, whereas water indicates PB solvation for MM and COSMO solvation for QM, and water* indicates the same as water, except with the outlying charge correction added to the COSMO energies. A0, A1, and A3 models were developed based on grids of conformations.

Model	Optimized	Solvent	Orig	Dft1	Dft2
ff99SB			1.74 ± 0.10	1.46 ± 0.02	1.56 ± 0.09
ff14SB			0.90 ± 0.02	2.78 ± 0.20	1.26 ± 0.08
A0	MM	vacuum	3.05 ± 0.27	4.00 ± 0.02	3.15 ± 0.21
A0	MM	water	1.13 ± 0.01	2.94 ± 0.01	1.52 ± 0.02
A0	MM	water*	1.05 ± 0.02	2.89 ± 0.03	1.45 ± 0.00
A0	MM/QM	vacuum	3.37 ± 0.16	3.69 ± 0.02	3.26 ± 0.12
A0	MM/QM	water	2.08 ± 0.05	2.48 ± 0.05	2.12 ± 0.07
A0	MM/QM	water*	2.09 ± 0.03	2.48 ± 0.07	2.14 ± 0.05
A1	MM	vacuum	3.61 ± 0.13	4.57 ± 0.03	3.71 ± 0.08
A1	MM	water	1.85 ± 0.13	4.30 ± 0.20	2.49 ± 0.16
A1	MM	water*	1.80 ± 0.09	4.22 ± 0.16	2.42 ± 0.12
A1	MM/QM	vacuum	3.37 ± 0.44	3.42 ± 0.26	3.21 ± 0.39
A1	MM/QM	water	1.98 ± 0.00	2.58 ± 0.06	2.00 ± 0.02
A1	MM/QM	water*	2.13 ± 0.18	2.58 ± 0.10	2.12 ± 0.17
A3	MM	vacuum	4.56 ± 0.20	6.05 ± 0.34	4.89 ± 0.27
A3	MM	water	5.15 ± 0.26	9.04 ± 0.32	6.42 ± 0.29
A3	MM	water*	5.05 ± 0.19	8.87 ± 0.27	6.29 ± 0.23
A3	MM/QM	vacuum	4.04 ± 0.01	3.98 ± 0.12	4.15 ± 0.03
A3	MM/QM	water	4.39 ± 0.01	6.82 ± 0.05	5.07 ± 0.01
A3	MM/QM	water*	4.25 ± 0.18	6.57 ± 0.27	4.88 ± 0.22

Table 3.3: χ^2 deviations from experimental scalar couplings [Graf et al., 2007] for simulations of Ala_5 in OPC solvent, according to Orig, Dft1, and Dft2 sets of Karplus curves (described in text). Model, Optimized, and Solvent columns follow the same conventions as Table 3.2.

Model	Optimized	Solvent	Orig	Dft1	Dft2
ff99SB			2.07 ± 0.42	1.73 ± 0.30	1.89 ± 0.42
ff14SB			0.85 ± 0.02	2.72 ± 0.11	1.20 ± 0.03
A0	MM	vacuum	2.72 ± 0.00	3.60 ± 0.17	2.77 ± 0.06
A0	MM	water	1.01 ± 0.02	2.85 ± 0.07	1.39 ± 0.04
A0	MM	water*	0.96 ± 0.04	2.71 ± 0.06	1.30 ± 0.05
A0	MM/QM	vacuum	3.92 ± 1.10	3.93 ± 0.65	3.72 ± 0.98
A0	MM/QM	water	2.23 ± 0.03	2.73 ± 0.09	2.40 ± 0.07
A0	MM/QM	water*	2.18 ± 0.07	2.67 ± 0.11	2.33 ± 0.12
A1	MM	vacuum	2.85 ± 0.08	3.47 ± 0.07	2.87 ± 0.10
A1	MM	water	1.41 ± 0.02	3.63 ± 0.04	1.94 ± 0.03
A1	MM	water*	1.26 ± 0.04	3.46 ± 0.06	1.77 ± 0.05
A1	MM/QM	vacuum	3.64 ± 0.81	3.63 ± 0.51	3.47 ± 0.74
A1	MM/QM	water	1.89 ± 0.05	2.41 ± 0.08	1.89 ± 0.09
A3	MM	vacuum	3.61 ± 0.08	4.42 ± 0.09	3.67 ± 0.08
A3	MM	water	3.90 ± 0.20	7.43 ± 0.00	4.97 ± 0.13
A3	MM/QM	vacuum	3.95 ± 0.10	3.75 ± 0.14	4.03 ± 0.13
A3	MM/QM	water	3.27 ± 0.17	5.03 ± 0.23	3.63 ± 0.19

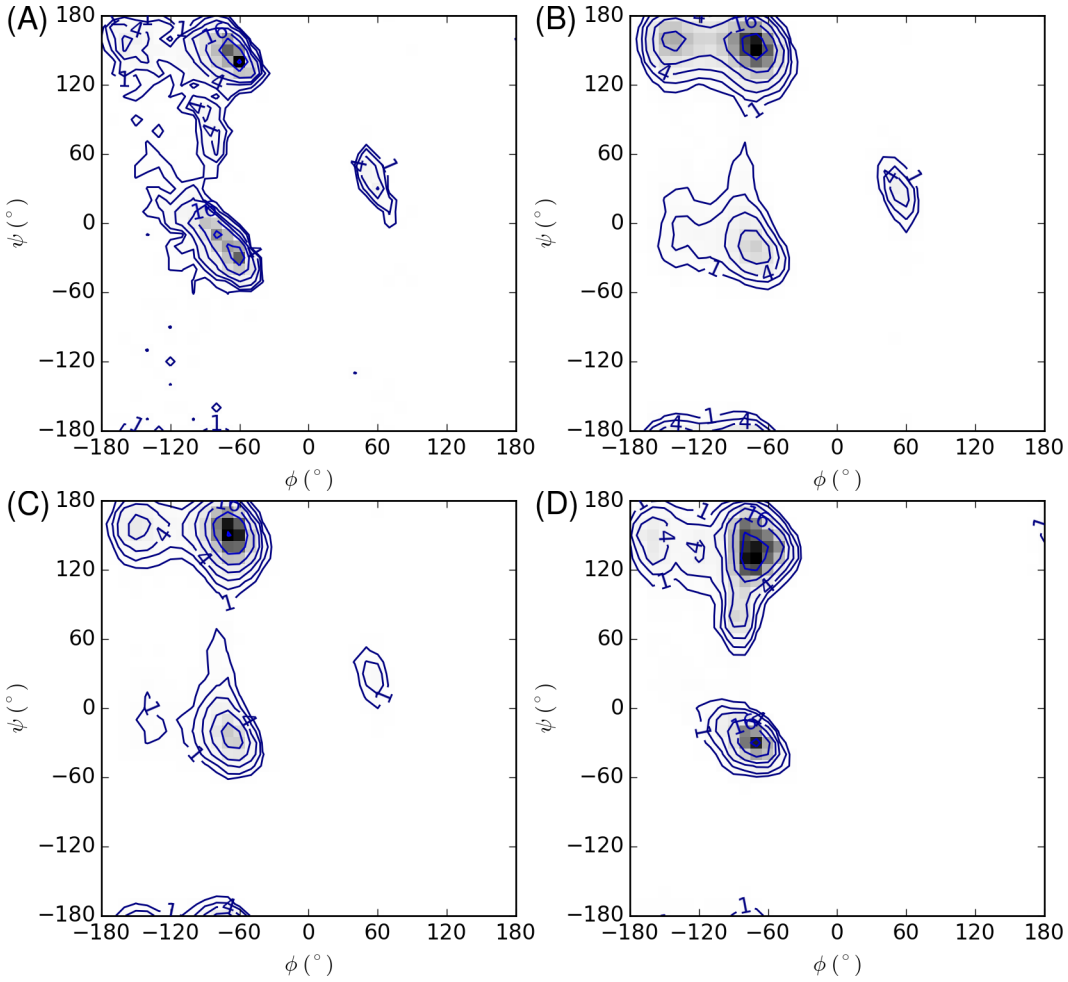

Figure 3.17: Histograms of alanine ϕ, ψ distributions based on (A) the PDB [Lovell et al., 2003], or the third residue of Ala_5 in simulations using (B) ff99SB, (C) ff14SB, or (D) A0 parameters.

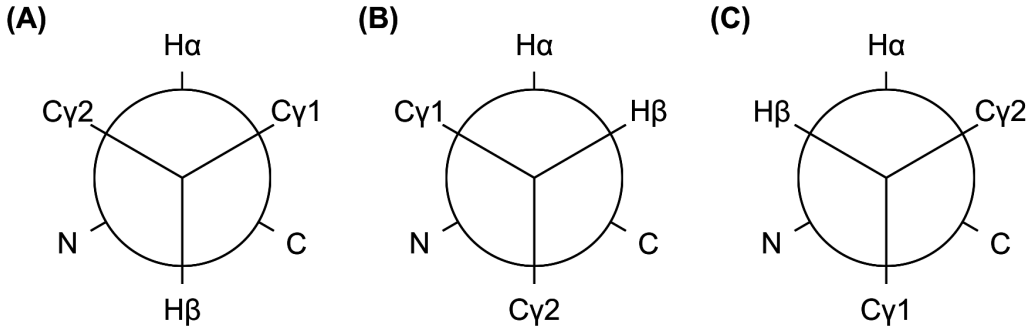

Figure 3.18: Newman projections of the side chain rotamers of L-valine, named for the dihedral angle between N and Cγ1: (A) trans (t); (B) gauche$^-$ (m); and (C) gauche$^+$ (p).

3.4 How do backbone parameters differ for β-branched valine?

As discussed above, an assumption from ff94 through ff99SB was that alanine is an appropriate model to train backbone parameters for all non-glycine amino acids that possess a β carbon. But scalar coupling results for ff99SB and ff14SB suggest that alanine and valine may need separate parameters. The differences implied by NMR scalar couplings for Ala_3 and Val_3 are corroborated by PDB distributions (Figure 3.1).

I thus refined backbone parameters against Val_1, applying the methodology developed for alanine. First, I generated a grid of dipeptide (V1) conformations, as was done for alanine. In this case, however, valine had not one grid across ϕ and ψ, but three—for the t (trans), m (gauche minus), and p (gauche plus) rotamers defined by Lovell et al. [2000], shown in Figure 3.18. The absolute QM and MM energies, as well as their differences, are plotted in Figure 3.19.

There are subtle differences among the QM − MM surfaces for different rotamers. It may be an issue that these backbone energy differences depend on side chain conformation, and thus cannot be reproduced fully in the context of all rotamers with a CMAP. To quantify the similarity of backbone errors across rotamers, the root mean square deviations (RMSD) in backbone energy

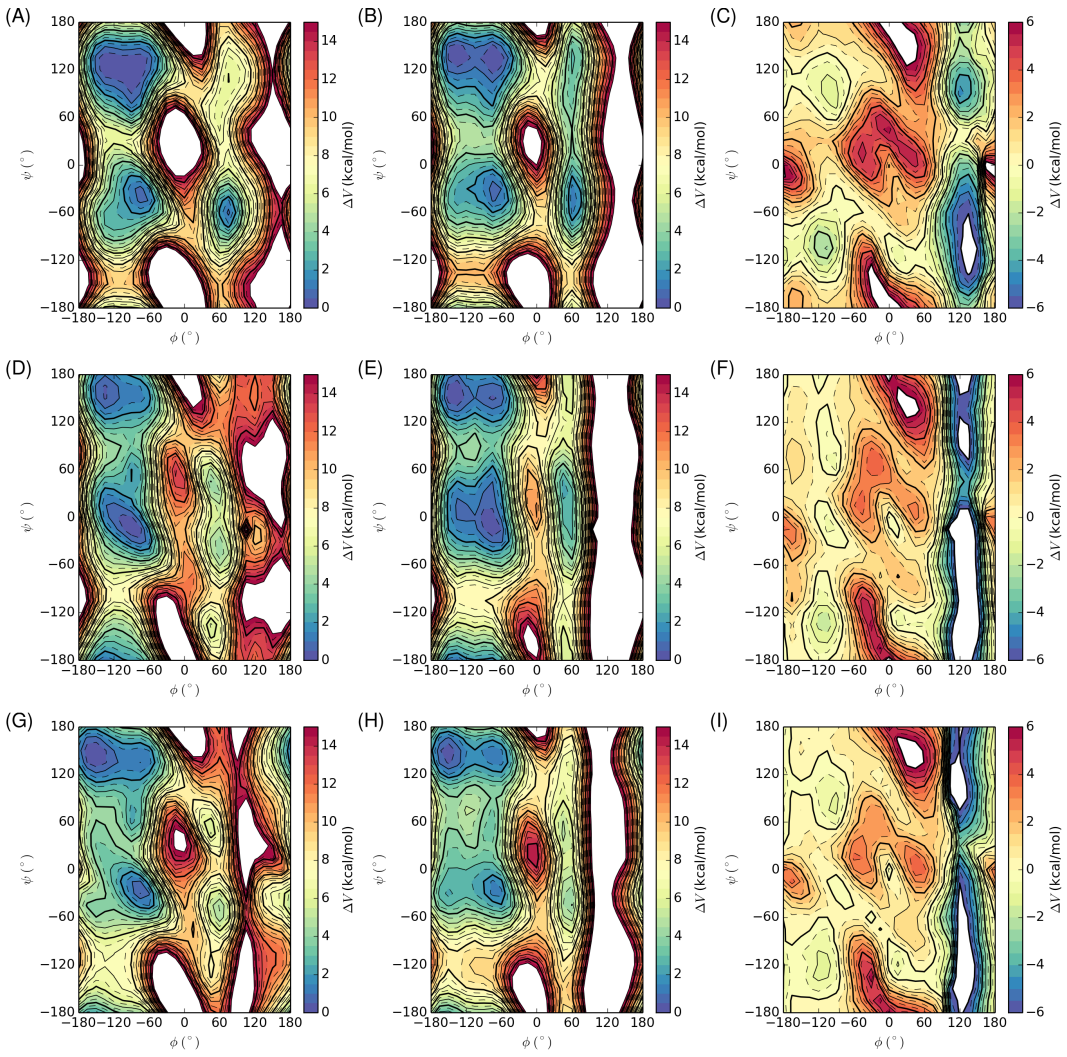

Figure 3.19: The V1 QM, MM, and QM-MM Ramachandran energy surfaces for the t (A–C), m (D–F), and p (G–I) rotamers as defined by Lovell et al. [2000].

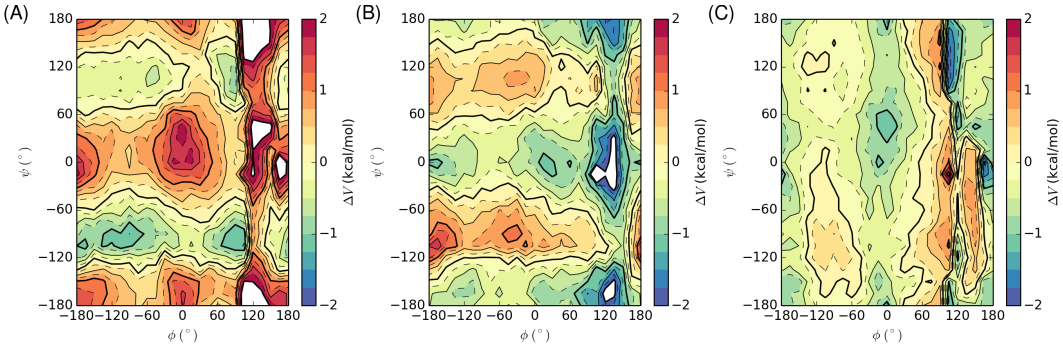

Figure 3.20: The V1 trans (A), gauche$^-$ (B), and gauche$^+$ (C) QM $-$ MM ϕ, ψ difference maps minus the average difference map.

differences were calculated between different side chain rotamers. This analysis revealed that the trans conformation is more different from the gauche conformations (RMSD of 2.0 kcal mol^{-1} to gauche$^-$ and 1.7 kcal mol^{-1} to gauche$^+$) than the gauche conformations are from each other (RMSD of 1.3 kcal mol^{-1}). The QM $-$ MM energy difference surfaces for each rotamer minus the average energy differences are plotted in Figure 3.20.

The trans and gauche$^-$ rotamers differ opposingly from the average, whereas the gauche$^+$ rotamer is closer to the average. In particular, whereas the trans rotamer requires more destabilization when ϕ or ψ are 0° or 0°, the gauche$^-$ rotamer exhibits the opposite trend. One difference could be that the trans rotamer has both side chain methyl groups gauche to the α-hydrogen, whereas the gauche$^-$ and gauche$^+$ rotamers have at least one methyl group gauche to both the backbone N and C (Figure 3.18). It may be that in the trans rotamer, the side chain methyls can't move away from the backbone as easily to avoid steric clashes, as moving away from the backbone by eclipsing the Hα would be accompanied by the two γ-methyls coming closer together. Additionally, the gauche$^-$ rotamer requires the least amount of additional stability at $\phi = -60°$ (Figure 3.20). As this rotamer has both side chain methyl groups gauche to the NH, it is possible that errors in steric clashes between the methyls and the backbone NH are responsible for this difference. If these factors are related to the side chain dependence in the backbone errors, then it

is possible that updating some van der Waals parameters could benefit transferability of a single backbone CMAP. Refitting non-bonded parameters may be important, but is beyond the goals established for this chapter. Which parameters might be beneficial to update is discussed in Section 3.6.

Despite the differences across rotamers, backbone energy comparisons against QM for all rotamers suggest that the demarcations between β and ppII are too well defined in ff14SB, consistent with the suggestions of scalar coupling and PDB comparisons. The average energies across ϕ and ψ for all three rotamers according to QM and MM are shown in Figure 3.21A-B. The QM − MM differences, shown in Figure 3.21C, include a higher β-ppII barrier in the MM profile, as well as different shaped basins. Notably, the energy differences suggest that α and ppII need to exhibit more rugged features, such as anticorrelation between ϕ and ψ in the α basin.

Applying a single valine correction is not as bad as the differences across rotamers would suggest, as the average valine energy difference surface is closer to the energy difference surfaces for each rotamer than the surfaces for each rotamer are to each other. The trans, gauche$^-$, and gauche$^+$ energy maps differ from the average map by RMSDs of 1.2 kcal mol^{-1}, 1.0 kcal mol^{-1}, and 0.7 kcal mol^{-1}, respectively. Thus a valine CMAP averaged over each rotamer, though not perfect, may still be a reasonable model for the valine energy surfaces.

To follow the alanine protocol of extrapolating to the monopeptide level, extrapolation to a V0 correction was the next goal. Adding the V1 CMAP and the A0 − A1 offset yielded the V0(A) CMAP, where the (A) signifies usage of the alanine zero-length offset. This CMAP is illustrated in Figure 3.21D. Here the alanine extrapolation is assumed to apply to valine.

A question that arises is whether the Ala_1 (A1) correction map also reproduces the V1 energy surfaces reasonably well. This was assessed by calculating RMSDs between the A1 and V1 difference maps. Intriguingly, the RMSDs between the A1 difference map and the V1 difference maps for each rotamer are comparable to the RMSDs between the V1 difference maps for different rotamers. The RMSDs from the A1 difference map to the V1 trans or gauche$^-$ maps are both 1.4 kcal mol^{-1}, whereas the RMSD from the A1 difference map

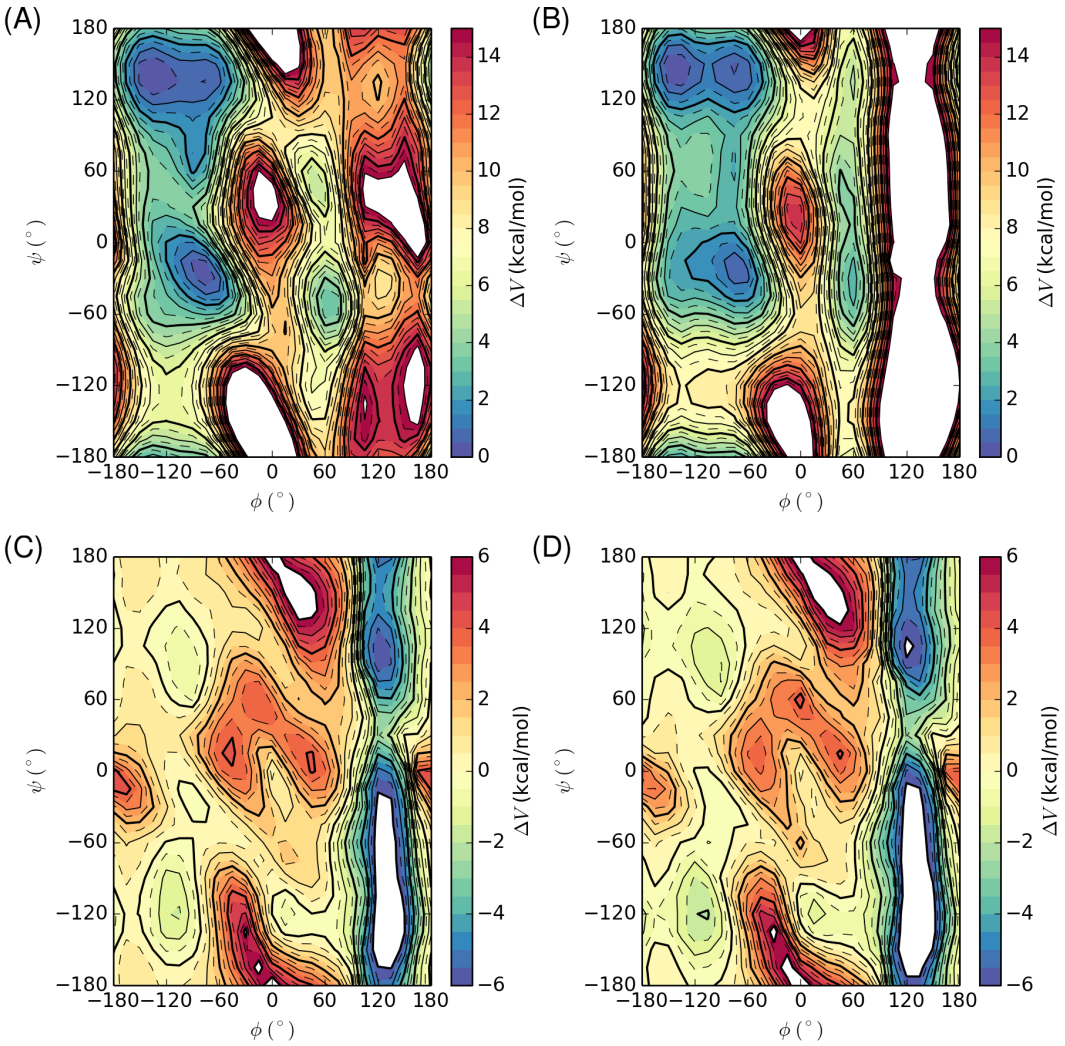

Figure 3.21: The average (A) QM and (B) MM energies, averaged for all three rotamers, of V1 across ϕ and ψ. (C) the difference between the average QM and MM profiles, and (D) the differences extrapolated to V0 using the A0 − A1 offsets.

to the gauche$^+$ map is 1.3 kcal mol^{-1}. Thus the differences between the energy map for alanine and the energy maps for valine rotamers are smaller than the differences between valine rotamers. It should be noted, however, that the average valine map is still closer to the different rotamer maps than is the alanine map.

The important question still remains of how the alanine and valine correction surfaces differ, and whether they are consistent with the sampling suggested by scalar couplings [Graf et al., 2007] and PDB distributions [Lovell et al., 2003]. The QM$-$MM energy differences (and thus CMAPs) for A1 were subtracted from the energy differences for V1, with the result graphed in Figure 3.22. There is a trend that conformations with $\phi \approx -120°$ are stabilized for valine more than for alanine. Specifically, the valine correction stabilizes $\phi = -120°$ relative to $\phi = -60°$ more than alanine by 0.5 ± 0.2 kcal mol^{-1}. This difference is contrary to the ff14SB increase in the β-ppII barrier for valine, and is consistent with the erroneous decrease in V3 $H_N H_\alpha$ scalar couplings accompanying ff14SB.

CMAPs for V1, V0(A), and A0 were used to simulate V3, along with ff99SB and ff14SB. The Ramachandran histogram of residue 2 for each force field is plotted in Figure 3.23. Notably, the shapes of the conformational basins look similar for simulations with the V1 and V0 CMAPs, but both look different from the shapes with ff99SB and ff14SB, which themselves look similar. Whereas ff99SB and ff14SB both have defined β and ppII populations with a barrier between them, the V1 and V0 force fields exhibit a continuous gradient of population from ppII, decreasing across toward β. Additionally, ff99SB and ff14SB both sample a helical basin that is wide across ϕ, sampling from $\sim -60°$ to $\sim -150°$, whereas the new force fields sample a more conservative helical basin terminating around $-120°$. Additionally, subtleties like the less rounded boundaries of the secondary structure population contours are consistent with the PDB distributions shown in Figure 3.1 and Figure 3.23F. Histograms of the PDB distribution [Lovell et al., 2003] and the second residue in Val_3 according to ff99SB, ff14SB, and V0(A) are brought together in Figure 3.24 to facilitate this comparison.

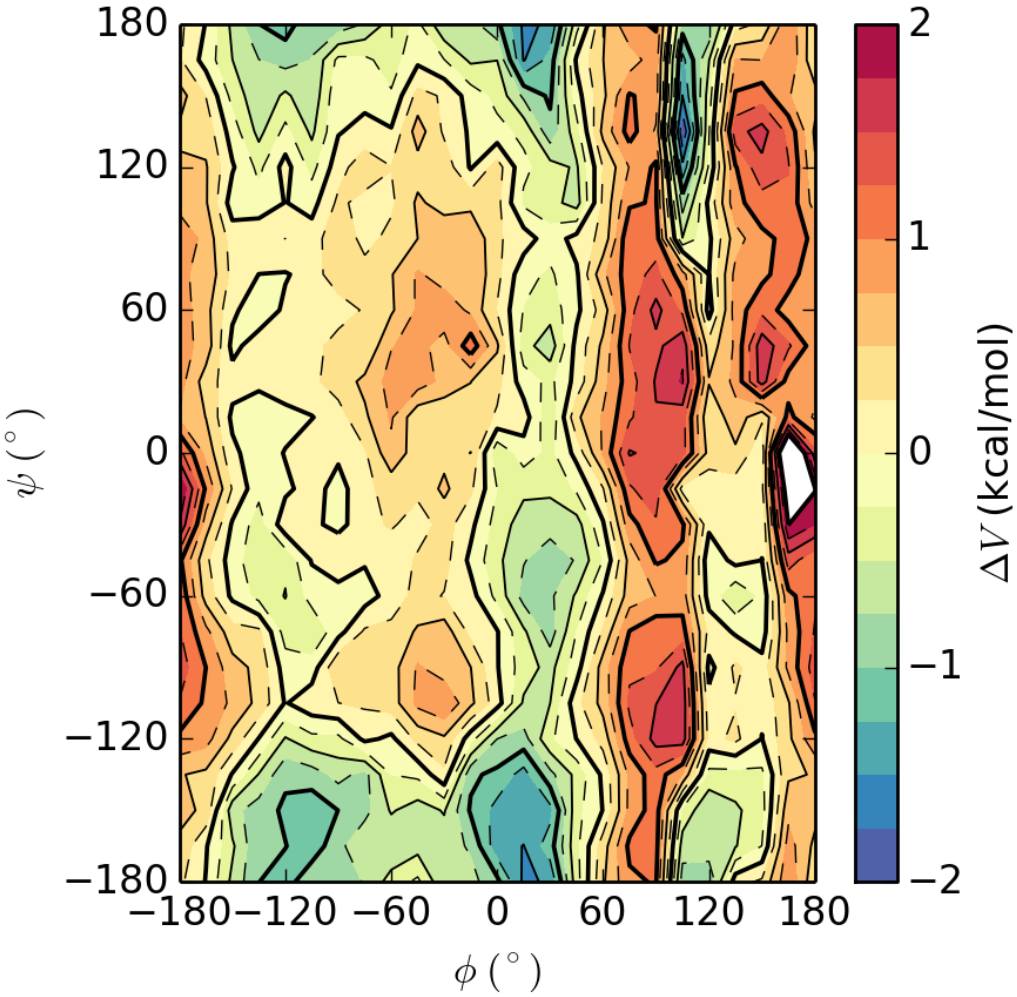

Figure 3.22: A map of the differences from the alanine CMAP to the valine CMAP.

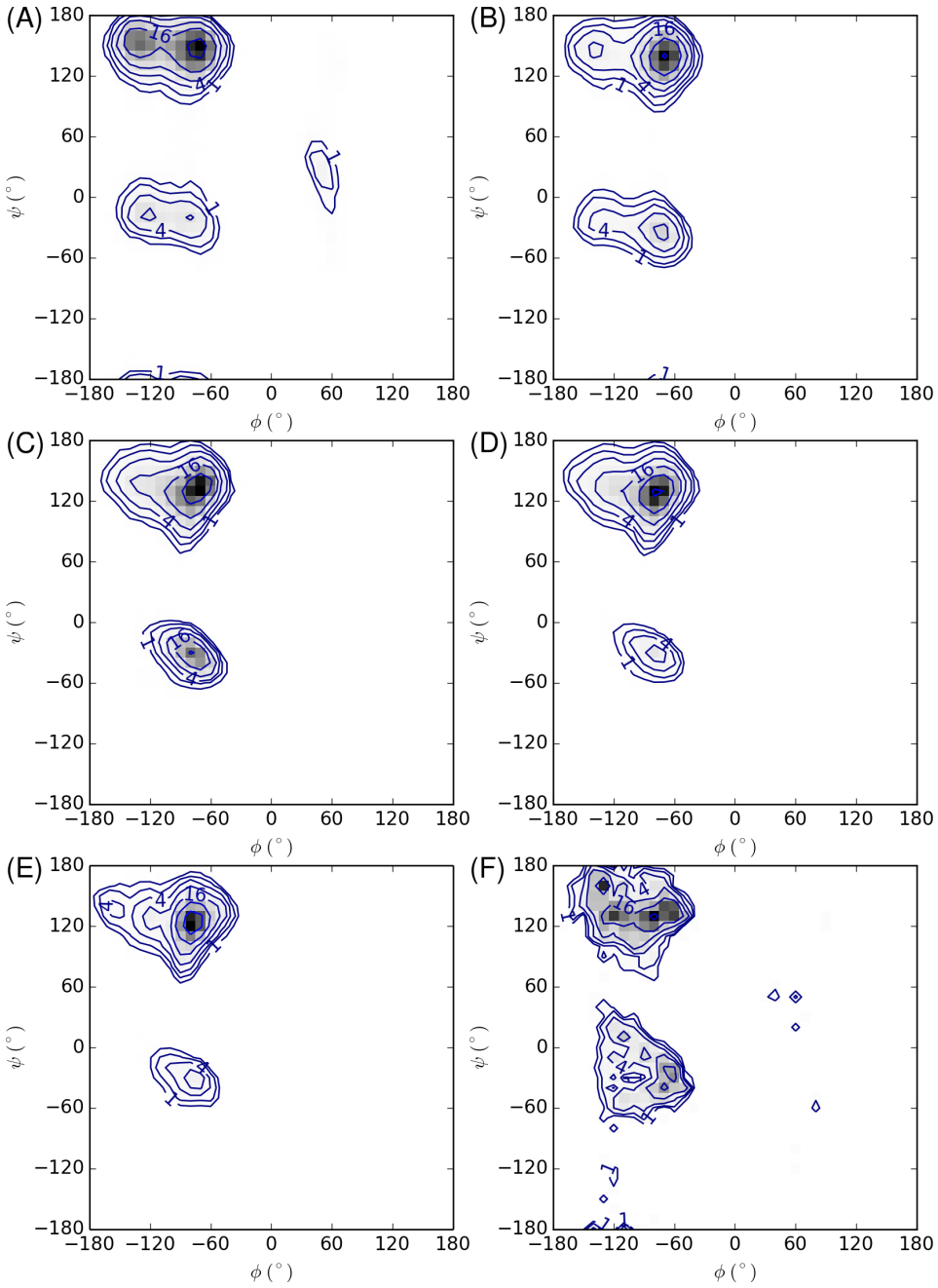

Figure 3.23: Ramachandran ϕ/ψ histogram of residue 2 of Val_3 with (A) ff99SB, (B) ff14SB, (C) V1-MM-water*, (D) V0(A)-MM-water*, and (E) A0 force fields, as well as (F) a histogram of valine conformations according to Lovell et al. [2003]. Note the lack of a β-ppII transition in the new valine-based force fields (panes C and D), whereas with the A0 parameters there is some energy separation between ppII and β.

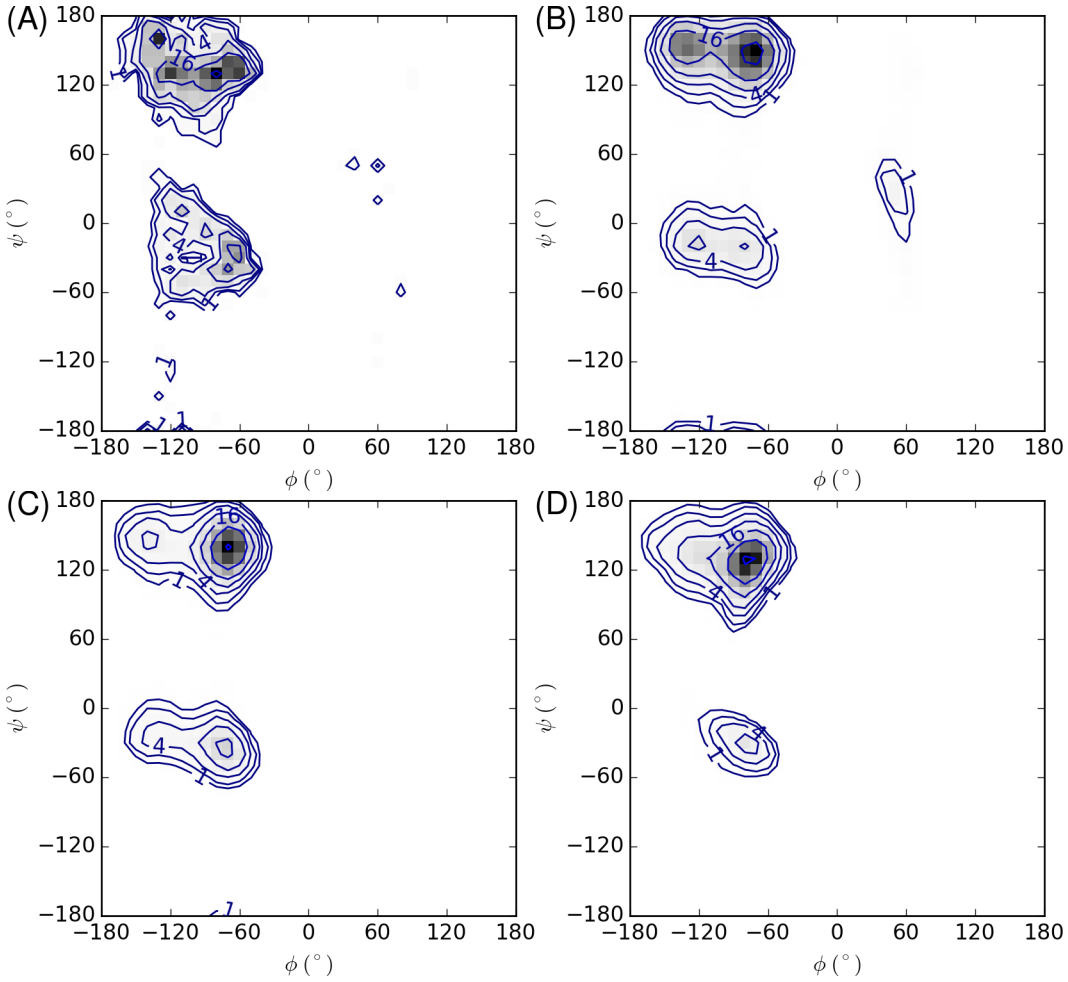

Figure 3.24: Histograms of valine ϕ, ψ distributions based on (A) the PDB [Lovell et al., 2003], or the second residue of Val_3 in simulations using (B) ff99SB, (C) ff14SB, or (D) V0(A) parameters.

Table 3.4: The χ^2 error for Val_3 scalar couplings in TIP3P solvent according to Orig, Dft1, and Dft2 Karplus parameters

Force field	Orig	Dft1	Dft2
ff99SB	1.08 ± 0.17	1.64 ± 0.13	1.25 ± 0.19
ff14SB	1.31 ± 0.03	4.48 ± 0.32	2.14 ± 0.12
A0	1.09 ± 0.16	4.19 ± 0.07	1.94 ± 0.13
V1	1.04 ± 0.01	3.90 ± 0.07	1.74 ± 0.04
V0(A)	0.97 ± 0.03	3.66 ± 0.08	1.60 ± 0.02

The new force fields don't only have a histogram that reflects the PDB distribution in overall shape, but also better reproduce V3 scalar couplings than ff14SB, with the V0(A) parameters performing comparably to ff99SB (Table 3.4). Notably, the ff99SB χ^2 with the Orig Karplus parameters was 1.08 ± 0.17, whereas that for V0(A) was 0.97 ± 0.03, lower on average but well within error bars. This result is important, as it suggests that a single protocol can achieve χ^2 of ~ 1 for both alanine and valine, as was not the case in ff99SB or ff14SB. The performance with the A0 parameters is also comparable to ff99SB, but with high uncertainties is yet difficult to compare against the V0(A) parameters. Further testing of these parameters will require more sampling of small systems like Val_3, as well as larger systems.

3.5 Conclusion

We demonstrated that quantum mechanics, with the appropriate manipulations, can generate energy surfaces for alanine and valine that fare well against small peptide scalar couplings. The conformational preferences displayed by these new force fields have a ϕ/ψ distribution more similar to those of the PDB [Lovell et al., 2003]. To achieve this agreement, implicit solvent in the training set energy calculations and extrapolation of parameters to those appropriate for a "zero-length" peptide were needed, as well as fitting by CMAPs. The performance in amino acids with diverse preferences like alanine and valine suggests a protocol that can be extended to other amino acids.

3.6 Future directions

The work of this chapter seeks to answer a very specific question: whether quantum mechanics, used in light of the insights gained in Chapter 2—that full conformational sampling (not just minima, with variations across multiple dihedrals) is important, that the details in the QM and MM calculations matter, and that additional parameters can be assigned for specific residues—can help further improve the backbone corrections employed in force fields like ff14SB. But it leaves many questions unanswered. The questions that the author feels are most important are discussed below.

Firstly, the COSMO versus PB comparison, which is at the heart of the positive results for A0 and V0(A) CMAPs, needs to be examined in more detail. In the preceding sections, I used all options that were recommended for each method. This included radii, that were thus allowed to differ between COSMO in the QM calculations, and PB in the MM calculations. This can result in artifacts between the MM and QM calculations that aren't even because of differences between PB and COSMO, but arise simply because the solvent cavity being treated is not the same in both models.

To examine this briefly, I recalculated the solvated MM training energies for Ala and Val, constructing the PB solvent cavity with the atomic radii employed by COSMO. The correction maps for the MM structures with implicit solvent presented before are again presented alongside the correction maps where COSMO radii are used in the QM and MM calculations, along with the difference between the two maps, for Ala in Figure 3.25 and for Val in Figure 3.26. The CMAPs derived using cosmo radii are labeled with the suffix '.cosmoradii,' hence these corrections are either A0.cosmoradii or V0(A).cosmoradii. The COSMO radii are larger than the bondi radii generally employed by PB; thus the helical hydrogen bond may be less shielded with the COSMO radii. Given the expected reduction in hydrogen bond shielding, it is not surprising that the energy surfaces derived from these new MM energies for Ala_3 do not require nearly the same helical stabilization relative to QM. Thus it is likely that the choice of radii, at least inasmuch as they are the same between both solvation models, is highly relevant.

Table 3.5: χ^2 deviations from experimental scalar couplings [Graf et al., 2007] for simulations of Ala_5 in TIP3P solvent, according to Orig, Dft1, and Dft2 sets of Karplus parameters (described in text).

Model	Orig	Dft1	Dft2
A0.cosmoradii	1.32 ± 0.05	3.08 ± 0.05	1.69 ± 0.06
A1.cosmoradii	1.59 ± 0.05	3.54 ± 0.16	2.03 ± 0.09
A3.cosmoradii	3.12 ± 0.04	5.71 ± 0.06	3.83 ± 0.05

Table 3.6: χ^2 deviations from experimental scalar couplings [Graf et al., 2007] for simulations of Ala_5 in OPC solvent, according to Orig, Dft1, and Dft2 sets of Karplus parameters (described in text).

Model	Orig	Dft1	Dft2
A0.cosmoradii	1.13 ± 0.06	2.80 ± 0.02	1.45 ± 0.04
A1.cosmoradii	1.19 ± 0.00	2.93 ± 0.07	1.52 ± 0.03
A3.cosmoradii	2.10 ± 0.48	4.35 ± 1.14	2.63 ± 0.61

The reduction in A3 χ^2 resulting from using the same radii is noteworthy. As shown in Table 3.5 for TIP3P solvent and in Table 3.6 for OPC solvent, the χ^2 was reduced roughly 40%. The χ^2 for A1.cosmoradii parameters, however, is comparable to that for A1 parameters. This would suggest that the limitation in the tetrapeptide could relate to the effects of different PB radii on the helical hydrogen bond. The results with A3.cosmoradii still do not achieve the same level of reproduction of scalar couplings as A1.cosmoradii or ff99SB, however. Also, A0.cosmoradii parameters perform better than A1.cosmoradii or A3.cosmoradii parameters in TIP3P solvent, although in OPC solvent the A0.cosmoradii and A1.cosmoradii parameter sets achieve similar χ^2 within error bars. The valine χ^2 with parameters derived using the COSMO radii (Table 3.7) are also similar to the parameters derived using different radii for PB and COSMO. Whether the monopeptide parameters derived from energies calculated using the COSMO radii are more reliable than monopeptide parameters using different radii is not clear from this test. But evaluating different radii in the training calculations could be beneficial.

A second issue is that these parameters have been tested only on small

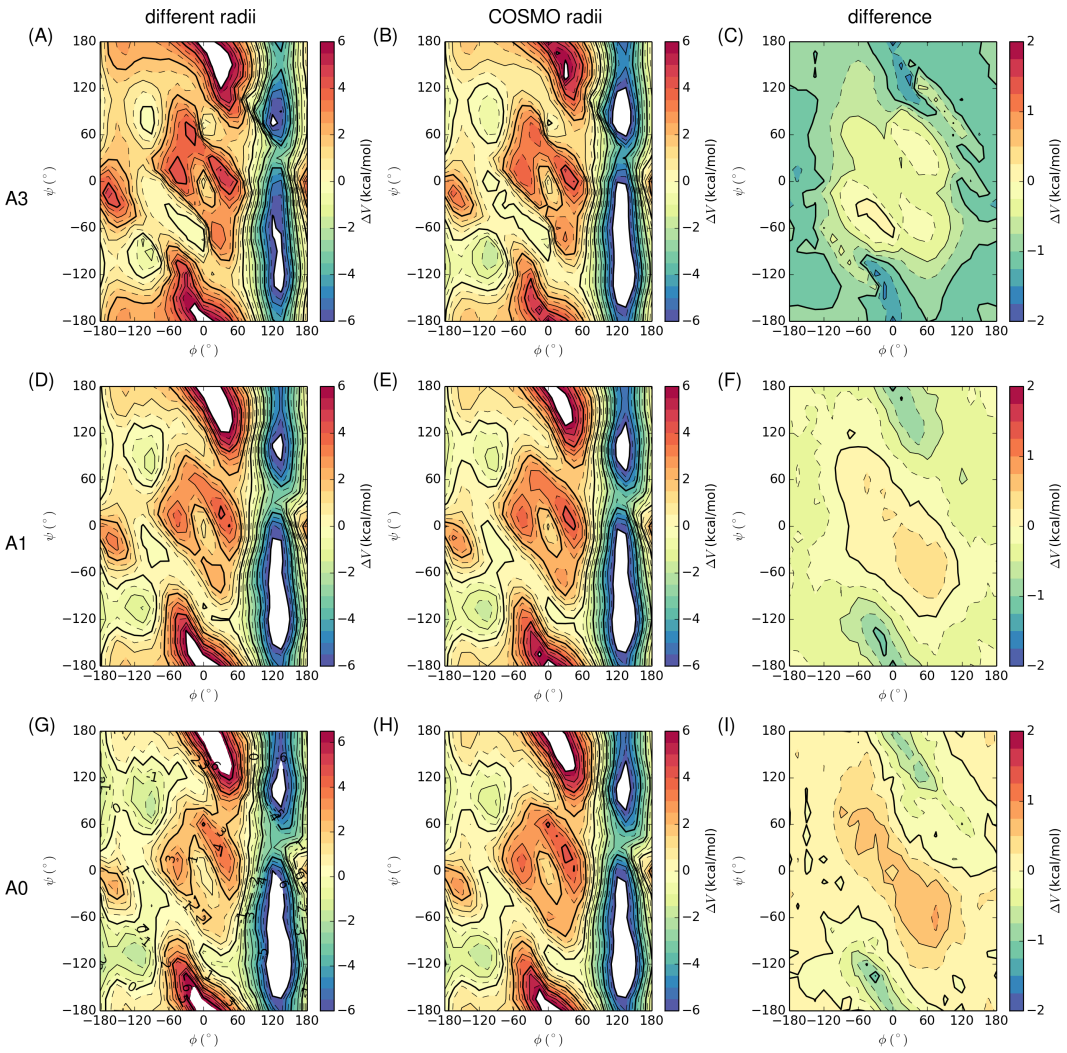

Figure 3.25: The correction maps for (rows) Ala_3, Ala_1, and Ala_0, according to (columns) calculations using recommended radii for each solvation model, calculations using the recommended radii for COSMO, and the difference from the correction using different radii to the correction using COSMO radii.

Table 3.7: The χ^2 error for Val_3 scalar couplings in TIP3P solvent according to Orig, Dft1, and Dft2 Karplus parameters

Force field	Orig	Dft1	Dft2
V1.cosmoradii	1.15 ± 0.14	3.65 ± 0.02	1.75 ± 0.14
V0(A).cosmoradii	0.95 ± 0.01	3.38 ± 0.02	1.55 ± 0.02

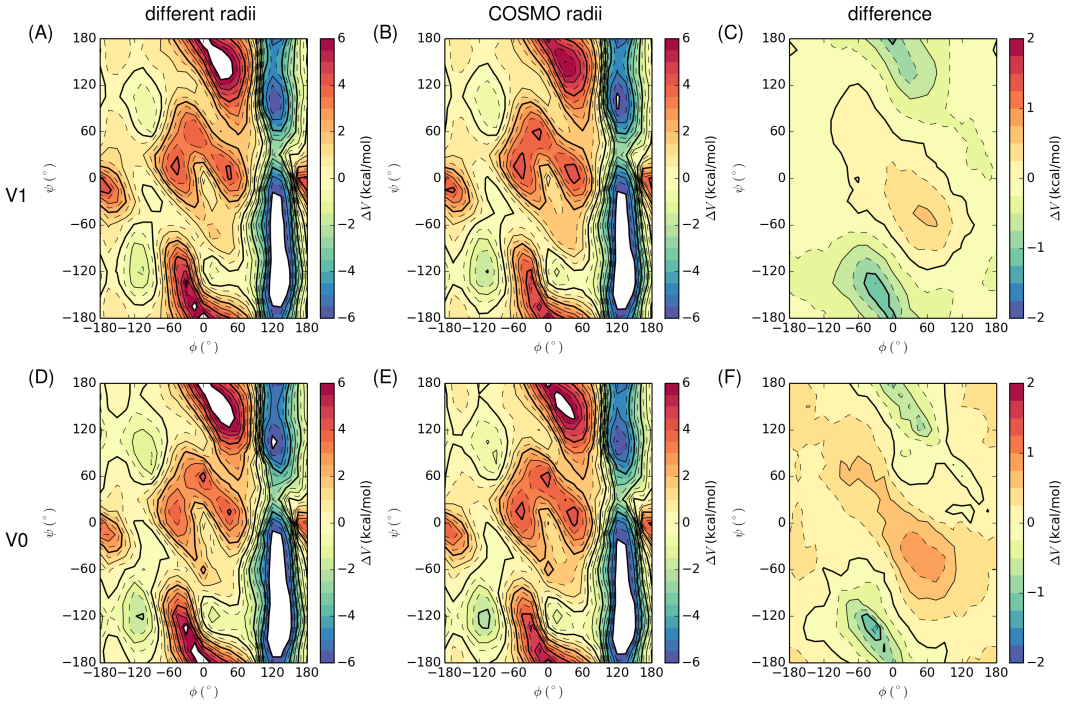

Figure 3.26: The correction maps for (rows) Val_1 and Val_0, according to (columns) calculations using recommended radii for each solvation model, calculations using the recommended radii for COSMO, and the difference from the correction using different radii to the correction using COSMO radii.

model systems. It of the utmost importance to perform more extensive validation on larger systems including full-length proteins, and especially to ascertain the effects on the stability of molecules like the ones used by Hai Nguyen and me previously [Nguyen et al., 2014]. A limited validation is presented here, with simulations of lysozyme with the new parameters being compared to simulations of lysozyme with ff14SB. Lysozyme had an issue previously as discussed in Chapter 2. The issue in the side chain sampling was most prominent in Q41 $\chi 1$ scalar couplings deviating from the experimental measurements. But there was also an issue with β-branched amino acids, particularly T43 and T69, having HNH_α scalar couplings that were too low (5.9 ± 0.2 Hz and 5.2 ± 0.3 Hz for T43 and T69, respectively, with ff14SB, compared to 9.3 Hz for both experimentally [Schwalbe et al., 2001]). This was concerning as the ff99SB scalar couplings were already lower than experiment (7.4 ± 0.4 Hz and 6.3 ± 0.4 Hz for T43 and T69, respectively), and the agreement was worsened with the updated parameters. One question, then, is whether the new backbone parameters might adjust the backbone sampling of these threonines, potentially increasing the HNH_α scalar couplings and improving Q41 $\chi 1$ scalar couplings, as well. Depending on which sets of parameters were used, T43 and T69 backbone scalar couplings are improved. With the A0 parameters applied to all non-glycine residues, the T43 and T69 HNH_α scalar couplings improve from ff14SB to 7.1 ± 1.4 Hz and 6.0 ± 0.2 Hz, respectively. With the A0.cosmoradii parameters, the HNH_α scalar couplings improve to 8.0 ± 0.3 Hz and 6.4 ± 0.3 Hz for T43 and T69, respectively. Meanwhile, Q41, which had an average normalized error (ANE, Equation (2.11)) of 0.17 ± 0.05 with ff99SB and 0.29 ± 0.06 with ff14SB, had an ANE of 0.19 ± 0.05 with ffA. With the ffA.cosmoradii parameters, however, the ANE remained at around 0.28 ± 0.10. These results show promise, but are not very well converged, and may depend on an issue briefly discussed in the next paragraph.

One important issue related to this analysis is whether the lysozyme crystal structure [Young et al., 1993] is a good starting point for dynamics simulations meant to be compared against NMR. In particular, the region of lysozyme from D66 to L83, which includes T69 and I87, has an RMSD between the NMR [Schwalbe et al., 2001] and crystal structure [Young et al., 1993] of

2.1 ± 0.2 Å (reported uncertainty is the standard deviation in the RMSD to the crystal structure across the 50 NMR structures), when the whole protein backbone (N, Cα, and C atoms) is aligned to the crystal structure. Some of the disagreement with NMR measurements may arise from the influence of the initial structure. This should be investigated further.

A third issue is that the backbone correction for valine depended on the side chain rotamer. This may arise from limitations in the ff14SB non-bonded parameters inherited from ff94. The author did a very simple analysis to investigate the van der Waals radii. Given the potential weaknesses observed in the steric interactions between the side chain methyl groups (CT and HC atom types) and the backbone O and H (with atom types O and H, respectively), the radii of each of the HC, CT, O, and H atoms was reduced by 1% or 10%, to examine whether lessening of steric clashes between the side chain γ-methyl groups and the backbone may be helpful. These reductions in van der Waals radii were evaluated by recalculating the energies of the valine conformations minimized with ff14SB. A better, but more expensive option, would be to re-minimize the conformations with the new van der Waals radii and calculate new QM energies based on the new set of conformations. This could be important because changing van der Waals interactions would likely affect the bonded degrees of freedom, such as angle bending, and should be done when this issue is investigated further.

To evaluate the limited tests that are performed here, simply measuring whether each change reduced the absolute error would not be adequate, as the highly trained dihedral parameters of ff14SB may be compensating for some weaknesses in nonbonded parameters. Thus, more accurate nonbonded energetics could potentially worsen agreement with QM. Instead, it is desirable to see that the backbone errors are similar across rotamers. This is measured by the side chain dependence (SCD) in the backbone errors. Much like the backbone dependence (BBD, Equation (2.7)) defined for side chain errors in the context of multiple backbone conformations, this is not a measure of the quality of the existing fit, but rather the potential ability to fit a single correction for all rotamers simultaneously. One important distinction from BBD is that SCD requires evaluation over more than two rotamers (rather than

α and β backbone contexts for BBD), so simply taking the differences in the relative errors between two rotamers won't suffice. Instead, the average difference in relative errors for all pairs of rotamers is computed. Thus the SCD was calculated as in Equation (3.14), where N_{sc} is the number of side chain rotamers, a and b are pairs of side chain rotamers, N_{bb} is the number of backbone conformations, i and j are pairs of backbone conformations, and $\text{REE}(i,j)_a = (E_{i,a}^{\text{MM}} - E_{j,a}^{\text{MM}}) - (E_{i,a}^{\text{QM}} - E_{j,a}^{\text{QM}})$ is the relative energy error (first defined for different side chain conformations in Equation (2.1)) between conformations i and j compared to QM, in the context of rotamer a.

$$\text{SCD} = \frac{2}{N_{\text{sc}}(N_{\text{sc}}-1)} \sum_{a}^{N_{\text{sc}}} \sum_{b<a} \frac{2}{N_{\text{bb}}(N_{\text{bb}}-1)} \sum_{i}^{N_{\text{bb}}} \sum_{j<i} |\text{REE}(i,j)_a - \text{REE}(i,j)_b| \tag{3.14}$$

The SCD was used to evaluate the "fittability" of the valine backbone parameters when each of the atom types, H or O in the backbone, and CT or HC in the side chain methyl, was reduced by 1% or 10%. As shown in Table 3.8, a 10% reduction in van der Waals radii is likely too extreme, causing SCD to increase in most cases, indicating less congruence between different rotamers. On the other hand, a 1% reduction in the van der Waals radius of any of the atom types reduced the SCD. Although the reduction is not huge (at most 6%, for reduction of the H van der Waals radius), this comparison does not exhaustively search for optimal van der Waals radii, and only considers one atom type at a time. A more thorough evaluation should involve more possible changes to the van der Waals radii, as well as combinations of changes to multiple atom types.

Ideally, once the above issues are resolved, one would apply the parameterization process to each amino acid individually. Having a separate set of parameters for each amino acid is an attractive option, as it would maximize the capacity to describe the sequence-dependent backbone preferences that are needed to connect primary to tertiary structure. The largest amino acids, and therefore the least conducive to ab initio QM calculations, have dozens of rotamers, however. Arginine, for example, has 34 side chain rotamers listed

Table 3.8: The side chain dependence (SCD, Equation (3.14)) with energies calculated using ff14SB (Atom type modified = None), or derivatives where van der Waals radii of atom types HC, O, CT, or H are reduced by 1% or 10%.

	SCD ($\mathrm{kcal\,mol^{-1}}$)	
Atom type modified	99% radii	90% radii
None	1.31	1.31
HC	1.25	1.37
O	1.25	1.37
CT	1.27	1.57
H	1.23	1.24

by Lovell et al. [2000]. Thus, backbone parameters for arginine would require 19 584 QM calculations to sample ϕ and ψ at each rotamer. Assuming sustained availability of 200 CPUs and that each calculation may take on the order of 10 h to complete, this set of calculations would take roughly 41 d. Considering arginine is just one of the twenty natural amino acids (albeit the most expensive one), it may be desirable to first group the amino acids based on common characteristics, before undertaking such a comprehensive and computationally expensive protocol as deriving unique parameters for each amino acid.

A possible arrangement of amino acids might be hypothesized by examining the PDB distributions of each [Lovell et al., 2003]. The ϕ distribution of each amino acid and H$-$Hα scalar couplings, calculated using the Hu and Bax [1997] Karplus equation evaluated for each ϕ angle contributing to the histogram, are depicted in Figure 3.27. While this scalar coupling calculation is across many different residues in various proteins and therefore nonphysical, it suggests that alanine should exhibit lower H$-$Hα scalar couplings ($6.3\,\mathrm{s^{-1}}$) than valine (7.8 $\mathrm{s^{-1}}$), and in fact, valine should have the second highest H$-$Hα scalar couplings. These values correlate quite well with NMR data for the second residues of $\mathrm{Ala_3}$ ($5.68 \pm 0.03\,\mathrm{s^{-1}}$) and $\mathrm{Val_3}$ ($7.94 \pm 0.02\,\mathrm{s^{-1}}$) [Graf et al., 2007]. Interestingly, the highest backcalculated H$-$Hα scalar couplings are for threonine at $7.9\,\mathrm{s^{-1}}$, with isoleucine nearby at $7.7\,\mathrm{s^{-1}}$. This similarity in sampling of all β-branched

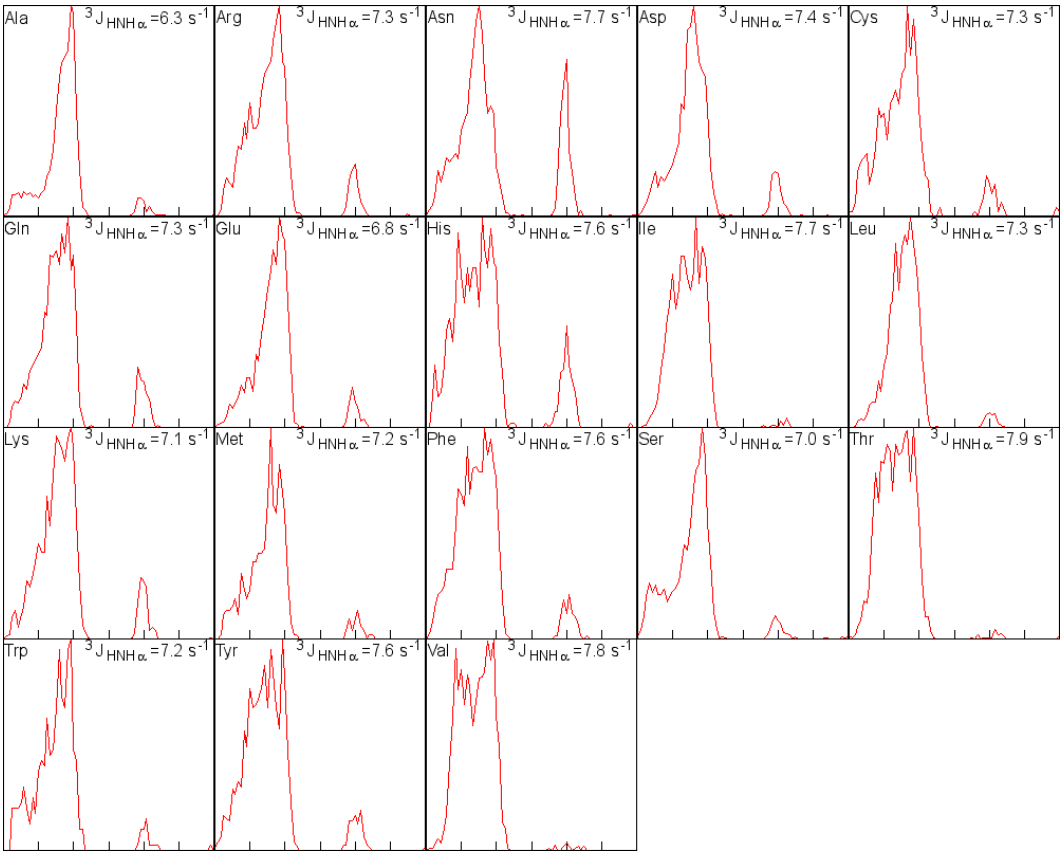

Figure 3.27: The ϕ histogram of each amino acid based on the conformations listed by Lovell et al. [2003]. The $^3J_{\mathrm{H-H}\alpha}$ scalar coupling in the upper right corner of each graph was calculated based on each distribution using the Hu and Bax Karplus parameters [Hu and Bax, 1997].

amino acids suggests that, based on the influence of the branch at the β-carbon, threonine and isoleucine may be grouped together with valine.

But alanine and valine (together with isoleucine and threonine) are conformational outliers. Many residues exhibit conformational preferences between those of alanine and valine (Figure 3.27). Additionally, whereas alanine only has a β-carbon, and therefore no χ side chain dihedrals, valine only has γ-carbons, and therefore only $\chi 1$. Meanwhile, many amino acids have longer side chains. It may be more appropriate to use a residue or set of residues other than alanine or valine to derive parameters for non-alanine, non-β-branched amino acids. The question becomes which of the amino acids with one heavy

Table 3.9: The possible composition of a future force field based on the method derived, applied to all amino acids in the left column, to yield parameters for the corresponding amino acids in the right column

AA derived from	AA(s) applied to
Ala	Ala
Val	Val · Ile · Thr
Pro	Pro
Gly	Gly
Leu	Leu
	Cys · Ser · Asn · Gln
	Phe · Tyr · Trp · His
	Asp · Glu
	Lys · Arg
	Met

atom at the side chain γ position, and perhaps at the δ position as well, would be suitable.

One residue that appears to be a well-suited model is leucine. Its conformational preferences are nearly halfway between alanine and valine, as measured by the $H_N H_\alpha$ scalar coupling. It is also uncharged, lessening potential concerns about non-bonded artifacts in QM/MM comparisons and improving the chances that the A0 extrapolation can be safely applied. Importantly, having only two side chain dihedrals, Lovell et al. have only listed five rotamers for leucine [Lovell et al., 2000]. Thus it is tractable for backbone corrections to be derived in the context of all five rotamers. As a first step, therefore, I propose a force field with the composition outlined in Table 3.9.

This landscape may still task leucine with too much responsibility, if the many amino acids assigned to it have residue-specific preferences that cannot be modeled adequately with the scheme in Table 3.9. If extensive test simulations suggest that certain residues have systematic deviations from experiment with leucine parameters, groups may be further separated. Each row for leucine (Leu) in Table 3.9 could comprise its own group, including leucine (by itself), polar amino acids, aromatics, acids, bases, and methionine (by itself, or perhaps grouped with leucine). This would be a large undertaking, and therefore

I recommend a stepwise protocol, making sure that each step is justified and only adding more fitting calculations as needed. It is still possible that the protocol derived here may encounter problems with some of the amino acids like the charged lysine, for example.

Ultimately, the necessity of these or other demarcations would need to be evaluated in more detail. To avoid prohibitive computation by reducing the load of costly QM calculations, appropriate residue groupings could be identified by comparing innate preferences of each amino acid in a small peptide to that residue's PDB distribution [Lovell et al., 2003]. It may be the case, for example, that Asp is better grouped with Asn than with Glu, or that neither grouping makes more sense than another. On the other hand, when training data can be obtained for all amino acids, it may be desirable to have an individual CMAP correction for each.

Once the appropriate parameter partition is identified, I feel that the protocol demonstrated for alanine and valine may be a useful means of obtaining quality training data using quantum mechanics. Compared to the ff14SB model partitioning amino acids into glycine and everything else, this should allow greater reproduction of sequence-dependent backbone preferences.

Chapter 4

Implications

I have shown that quantum mechanics can be successfully applied to enhancing the accuracy of molecular mechanics force fields based on ff99SB [Hornak et al., 2006]. As described in Chapter 2, comprehensive fitting of molecular mechanical side chain dihedral parameters against quantum mechanics (QM) energy profiles can enhance accuracy as measured by side chain scalar couplings, as well as improving secondary structure preferences. Furthermore, Chapter 3 presented a quantum mechanics-based fitting method that shows promise for improving the backbone potential energy surface beyond the many small tweaks [Best and Hummer, 2009, Li and Brüschweiler, 2011, Maier et al., 2015] applied to ff99SB. These accomplishments have implications for force field development as well as for protein folding and structure prediction.

A longstanding challenge to the biophysical community is to predict the folding of a protein's tertiary structure based solely on its primary sequence. The most successful efforts at structure prediction have included a blend of bioinformatic and physics-based approaches. For example, Rosetta employs a number of statistical rules to predict secondary structure motifs and then optimizes the arrangement of those motifs using physics [Simons et al., 1999]. Simulation-based methods that incorporate even a relatively sparse set of native contacts (e.g. from NOEs) have also been successfully applied to protein structure prediction [Perez et al., 2013].

Although information-based methods like Rosetta and Meld enjoy success at prediction of three-dimensional structures in competitions like CASP, successful folding to the correct three-dimensional structure using physics alone remains a difficult endeavor because of two simultaneous requirements. First, the employed physics model needs to be fast enough to predict the folding that typically occurs on the supra-microsecond, and even supra-millisecond, timescale. Second, a physics model of sufficient speed must also be accurate enough to select a single native protein structure of many possible alternatives that through a random search, as may be approached by a model that does not favor the native state, would not be found within the age of the universe [Levinthal, 1969].

The tertiary structure of a protein is encoded in its primary structure. As side chains are what differentiate the amino acids that comprise a protein's primary structure, I expect that the side chain parameter modifications presented here could have important implications for prediction of protein structure and stability.

For example, the GB-Neck2 implicit solvent model [Nguyen et al., 2013] that was recently developed in the Simmerling lab was paired with the updated side chain parameters presented in Chapter 2, and applied to the folding of seventeen proteins of varying topologies (α, β, or both) from 10 to 92 amino acids long [Nguyen et al., 2014]. Sixteen of seventeen proteins folded to the correct native structure. For fourteen of the seventeen proteins, this native structure was the preferred conformation. To our knowledge, this is the first time a physics-based force field has been so successfully paired with a physics-based implicit solvent model without explicitly depending on cancellation of error, and especially having been applied to a benchmark of proteins of such diversity.

What's noteworthy is that ff14SBonlysc was used partly because the native conformation [Freddolino et al., 2008] of one system, Fip35 [Liu et al., 2008], was not stable in simulations with ff99SB at 325 K, unfolding within hundreds of nanoseconds. To examine further, I simulated Fip35 starting from the native conformation for 4 runs each with ff99SB and ff14SBonlysc, extending each simulation to 18 µs, at 325 K, following all the same simulation parameters

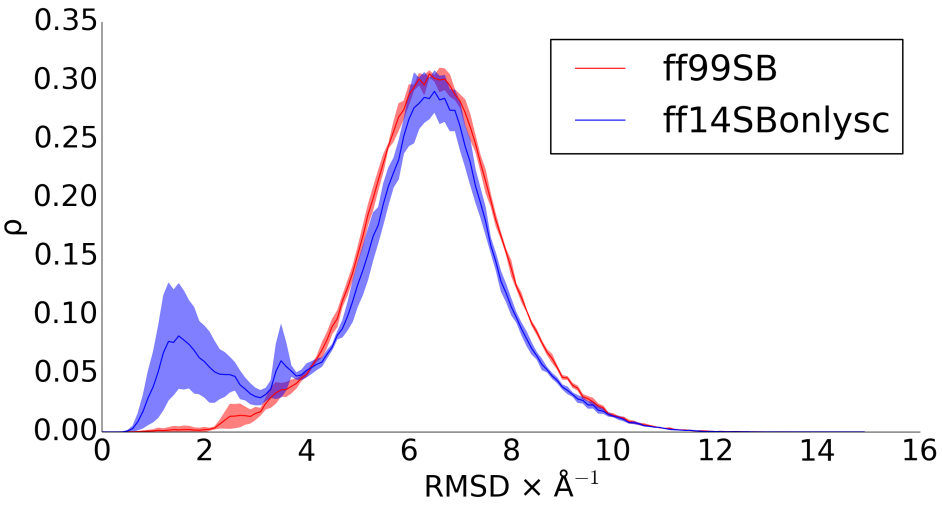

Figure 4.1: Histograms of the backbone (N, C_α, C) RMSD to the native structure of Fip35 for four simulations each with ff99SB (red) and ff14SBonlysc (blue).

used in the protein folding study [Nguyen et al., 2014], most notably GB-Neck2 implicit solvent with mbondi3 radii [Nguyen et al., 2013], 4 fs time step with hydrogen mass repartitioning [Feenstra et al., 2002], and the Langevin thermostat [Cies-gravela et al., 2001] with collision frequency $\gamma = 1.0\,\mathrm{ps}^{-1}$. To quantify the reproduction of the native structure, I calculated the RMSD of the backbone (N, C_α, C) of residues 10–28, including the two hairpins but not the long, flexible termini. These simulations confirmed the initial ff99SB test. Histograms of the RMSD (Figure 4.1), removing the first 6 µs that include the initial folded state, illustrate that ff99SB exhibits almost no conformations <2 Å RMSD during the latter 12 µs, whereas ff14SBonlysc exhibits a well-defined, if small, peak. To be clear, simulations with both force fields are predominantly unfolded. But the ff14SBonlysc simulations capture the native conformation in the structural ensemble, which did not happen with ff99SB. This suggests that the revised side chain parameters are able to improve protein stability dramatically, though there still may be room for improvement in, for example, the backbone parameters.

A truly accurate force field should not only distinguish the native fold from

unfolded conformations for varying protein sequences, but should also reproduce the sensitivity of structure to single amino acid substitutions in a family of related sequences. Koushik Kasavajhala has been studying TrpCage and several variants whose experimental folding characteristics have been shown to change with mutation of only a single residue. Accurate reproduction of this folding sensitivity would suggest reliable modeling at the level of individual amino acids. In his preliminary studies, he has found that ff14SB and ff14SBonlysc best reproduce the structural differences observed experimentally when specific residues are mutated. Thus, the updated side chain corrections common to both force fields impart structural preferences with enhanced sequence-specificity at fine granularity.

When it comes to training force field parameters, the question of what is a good target is fore. There is some debate over the continued utility of QM as a target as MM models mature and the required accuracy becomes higher. In CHARMM force field development, it is typical to employ quantum mechanics only in conjunction with empirical adjustments against, for example, crystal structures [MacKerell et al., 2004a, Best et al., 2012]. In Chapter 3, I examined whether training to match quantum mechanics energies might still be a viable option for improving backbone parameters, as it was in ff99SB [Hornak et al., 2006].

What I found, as for training side chain parameters, is that many details in how the QM and the MM calculations are performed can play pivotal roles in the efficacy of the resulting parameters. For example, training in vacuum against tetrapeptides can lead to parameters that strongly promote helical stability, compromising agreement with NMR scalar couplings [Graf et al., 2007]. But I was able to find combinations that yielded excellent agreement with the NMR data. Notably, these combinations involved the use of implicit solvent in the energy evaluations, as well as extrapolation of parameters to those appropriate for zero-length peptides, and use of coupled correction maps (CMAPs [MacKerell et al., 2004b]). I argue that the success of this quantum mechanical approach for alanine and valine suggests that quantum mechanics should continue to play a role in force field development, especially when moving to highly specific parameterizations where each amino acid may have

its own backbone parameters. Ultimately, Chapters 2 and 3 provide a means of enhancing the structural code in simulations. The rise of tertiary structure from primary sequence can thus be probed in silico in greater detail and with higher accuracy.

Bibliography

D. Adams. *The Hitchhiker's Guide to the Galaxy*. The Hitchhiker's Guide to the Galaxy. Pan Books, October 1979.

N. L. Allinger. Conformational analysis. 130. MM2. a hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society*, 99(25):8127–8134, 1977. doi: 10.1021/ja00467a001. URL `http://pubs.acs.org/doi/abs/10.1021/ja00467a001`.

N. L. Allinger, Y. H. Yuh, and J. H. Lii. Molecular mechanics. the MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society*, 111 (23):8551–8566, 1989. doi: 10.1021/ja00205a001. URL `http://pubs.acs.org/doi/abs/10.1021/ja00205a001`.

P. Arora. Personal Communication, 2015.

C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97 (40):10269–10280, 1993. doi: 10.1021/j100142a004. URL `http://dx.doi.org/10.1021/j100142a004`.

M. D. Beachy, D. Chasman, R. B. Murphy, T. A. Halgren, and R. A. Friesner. 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*, 119(25):5908–5920, 1997. doi: 10.1021/ja962310g. URL `http://dx.doi.org/10.1021/ja962310g`.

K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande. Are protein force fields getting better? a systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation*, 8(4):1409–1414, 2012. doi: 10.1021/ct2007814. URL `http://dx.doi.org/10.1021/ct2007814`. PMID: 22754404.

H. J. C. Berendsen, J. P. M. Postma, W. F. v. Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987. doi: 10.1021/j100308a038. URL `http://dx.doi.org/10.1021/j100308a038`.

K. D. Berndt, P. Güntert, L. P. Orbons, and K. Wüthrich. Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. *Journal of Molecular Biology*, 227(3):757 – 775, 1992. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/0022-2836(92)90222-6. URL `http://www.sciencedirect.com/science/article/pii/0022283692902226`.

R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *Journal of Physical Chemistry B*, 113(26):9004–9015, 2009.

R. B. Best, N.-V. Buchete, and G. Hummer. Are current molecular dynamics force fields too helical? *Biophysical Journal*, 95(1):L07–L09, 2008.

R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone , Ψ and side-chain χ1 and χ2 dihedral angles. *Journal of Chemical Theory and Computation*, 8 (9):3257–3273, 2012. doi: 10.1021/ct300400x. URL `http://pubs.acs.org/doi/abs/10.1021/ct300400x`.

A. Bondi. van der Waals volumes and radii. *The Journal of Physical Chemistry*, 68(3):441–451, 1964. doi: 10.1021/j100785a001. URL `http://pubs.acs.org/doi/abs/10.1021/j100785a001`.

D. Case, T. Darden, T. I. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. Goetz, I. Kolossvry, K. Wong, F. Paesani,

J. Vanicek, R. Wolf, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman. AMBER12, 2012.

D. Case, T. Darden, T. I. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. Goetz, I. Kolossvry, K. Wong, F. Paesani, J. Vanicek, R. Wolf, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman. AMBER14, 2014.

D. A. Case, C. Scheurer, and R. Brüschweiler. Static and dynamic effects on vicinal scalar J couplings in proteins and peptides: A MD/DFT analysis. *Journal of the American Chemical Society*, 122(42):10390–10397, 2000.

D. A. Case, T. E. Cheatham, H. Darden, R. Gohlke, R. Luo, K. M. Merz, A. Jr., Onufriev, C. Simmerling, B. Wang, and R. Woods. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26:1668–1688, 2005.

D. S. Cerutti, P. L. Freddolino, R. E. Duke, and D. A. Case. Simulations of a protein crystal with a high resolution X-ray structure: Evaluation of force fields and water models. *Journal of Physical Chemistry B*, 114(40): 12811–12824, 2010.

D. S. Cerutti, J. E. Rice, W. C. Swope, and D. A. Case. Derivation of fixed partial charges for amino acids accommodating a specific water model and implicit polarization. *The Journal of Physical Chemistry B*, 117(8):2328–2338, 2013. doi: 10.1021/jp311851r. URL `http://pubs.acs.org/doi/abs/10.1021/jp311851r`.

D. S. Cerutti, W. C. Swope, J. E. Rice, and D. A. Case. ff14ipq: A self-consistent force field for condensed-phase simulations of proteins. *Journal*

of Chemical Theory and Computation, 10(10):4515–4534, 2014. doi: 10.1021/ct500643c. URL http://dx.doi.org/10.1021/ct500643c.

R. N. Chapman, G. Dimartino, and P. S. Arora. A highly stable short α-helix constrained by a main-chain hydrogen-bond surrogate. *Journal of the American Chemical Society*, 126(39):12252–12253, 2004. doi: 10.1021/ja0466659. URL http://pubs.acs.org/doi/abs/10.1021/ja0466659. PMID: 15453743.

J. J. Chou, D. A. Case, and A. Bax. Insights into the mobility of methyl-bearing side chains in proteins from (3)J(CC) and (3)J(CN) couplings. *Journal of the American Chemical Society*, 125(29):8959–66, 2003.

M. Cies-gravela, S. P. Dias, L. Longa, and F. A. Oliveira. Synchronization induced by Langevin dynamics. *Physical Review E*, 63(6):065202, 2001.

W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollmann. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, 115(21):9620–9631, 1993. doi: 10.1021/ja00074a030. URL http://dx.doi.org/10.1021/ja00074a030.

W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N · log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

C. M. Davis, S. F. Xiao, D. P. Raeigh, and R. B. Dyer. Raising the speed limit for beta-hairpin formation. *Journal of the American Chemical Society*, 134 (35):14476–14482, 2012.

F. Ding, M. Layten, and C. Simmerling. Solution structure of HIV-1 protease flaps probed by comparison of molecular dynamics simulation ensembles and epr experiments. *Journal of the American Chemical Society*, 130(23):7184–7185, 2008. doi: 10.1021/ja800893d. URL http://dx.doi.org/10.1021/ja800893d. PMID: 18479129.

K. Y. Ding and A. M. Gronenborn. Protein backbone h-1(n)-c-13(alpha) and n-15-c-13(alpha) residual dipolar and j couplings: New constraints for nmr structure determination. *Journal of the American Chemical Society*, 126 (20):6232–6233, 2004.

Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003. ISSN 1096-987X. doi: 10.1002/jcc.10349. URL http://dx.doi.org/10.1002/jcc.10349.

R. L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins - application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–574, 1993.

K. Feenstra, C. Peter, R. Scheek, W. van Gunsteren, and A. Mark. A comparison of methods for calculating nmr cross-relaxation rates (noesy and roesy intensities) in small peptides. *Journal of Biomolecular NMR*, 23 (3):181–194, 2002. ISSN 0925-2738. doi: 10.1023/A:1019854626147. URL http://dx.doi.org/10.1023/A:1019854626147.

P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophys J*, 94(10):L75–L77, May 2008. ISSN 0006-3495. doi: 10.1529/biophysj.108.131565. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367204/. 131565[PII].

S. D. Fried, L.-P. Wang, S. G. Boxer, P. Ren, and V. S. Pande. Calculations of the electric fields in liquid solutions. *The Journal of Physical Chemistry*

B, 117(50):16236–16248, 2013. doi: 10.1021/jp410720y. URL http://pubs.acs.org/doi/abs/10.1021/jp410720y.

M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowsk, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, A. M. A. Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, H. M. Gordon, E. S. Replogle, and J. A. Pople. Gaussian 98, 1998.

P. Galison. Kuhn and the quantum controversy. *The British Journal for the Philosophy of Science*, 32(1):71–85, 1981. doi: 10.1093/bjps/32.1.71. URL http://bjps.oxfordjournals.org/content/32/1/71.short.

M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon. Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *The Journal of Physical Chemistry*, 97(14):3591–3600, 1993. doi: 10.1021/j100116a025. URL http://dx.doi.org/10.1021/j100116a025.

M. S. Gordon and M. W. Schmidt. Advances in electronic structure theory: Gamess a decade later. In C. Dykstra, G. Frenking, K. Kim, and G. Scuseria, editors, *Theory and Applications of Computational Chemistry: the first forty years*, pages 1167–1189. Elsevier, Amsterdam, 2005.

J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe. Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *Journal of the American Chemical Society*, 129(5):1179–1189, 2007. doi: 10.1021/ja0660406. URL http://dx.doi.org/10.1021/ja0660406. PMID: 17263399.

S. Grimshaw. *Novel approaches to characterizing native and denatured proteins by NMR.* Doctoral thesis, University of Oxford, 1999.

M. Hennig, W. Bermel, H. Schwalbe, and C. Griesinger. Determination of psi torsion angle restraints from (3)j(c-alpha,c-alpha) and 3j(c-alpha,h-n) coupling constants in proteins. *Journal of the American Chemical Society*, 122(26):6268–6277, 2000.

S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata. Crystal structure of a ten-amino acid protein. *Journal of the American Chemical Society*, 130 (46):15327–15331, 2008.

H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of Chemical Physics*, 120(20):9665–9678, 2004. doi: http://dx.doi.org/10.1063/1.1683075. URL `http://scitation.aip.org/content/aip/journal/jcp/120/20/10.1063/1.1683075`.

V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics*, 65(3):712–725, 2006.

J.-S. Hu and A. Bax. Determination of φ and $\chi 1$ angles in proteins from 13c13c three-bond j couplings measured by three-dimensional heteronuclear NMR. how planar is the peptide bond? *Journal of the American Chemical Society*, 119(27):6360–6368, 1997.

W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

S. Izadi, R. Anandakrishnan, and A. V. Onufriev. Building water models: A different approach. *The Journal of Physical Chemistry Letters*, 5(21):3863–3871, 2014. doi: 10.1021/jz501780a. URL `http://dx.doi.org/10.1021/jz501780a`.

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001-.

J. E. Jones. On the determination of molecular fields. II. from the equation of state of a gas. *Proceedings of the Royal Society of London. Series A*, 106(738):463–477, 1924. doi: 10.1098/rspa.1924.0082. URL `http://rspa.royalsocietypublishing.org/content/106/738/463.short`.

W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988. doi: 10.1021/ja00214a001. URL `http://pubs.acs.org/doi/abs/10.1021/ja00214a001`.

W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. doi: http://dx.doi.org/10.1063/1.445869. URL `http://scitation.aip.org/content/aip/journal/jcp/79/2/10.1063/1.445869`.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22 (12):2577–2637, 1983.

M. Karplus. Contact electron-spin coupling of nuclear magnetic moments. *The Journal of Chemical Physics*, 30(1):11–15, 1959.

M. Karplus. Vicinal proton coupling in nuclear magnetic resonance. *Journal of the American Chemical Society*, 85(18):2870–2871, 1963.

J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666, 1958. 10.1038/181662a0.

K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods. GLYCAM06: A generalizable

biomolecular force field. carbohydrates. *Journal of Computational Chemistry*, 29(4):622–655, 2008. ISSN 1096-987X. doi: 10.1002/jcc.20820. URL http://dx.doi.org/10.1002/jcc.20820.

A. Klamt and G. Schuurmann. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, pages 799–805, 1993. doi: 10.1039/P29930000799. URL http://dx.doi.org/10.1039/P29930000799.

P. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In W. van Gunsteren, P. Weiner, and A. Wilkinson, editors, *Computer Simulation of Biomolecular Systems*, volume 3 of *Computer Simulations of Biomolecular Systems*, pages 83–96. Springer Netherlands, 1997. ISBN 978-90-481-8528-3. doi: 10.1007/978-94-017-1120-3_2. URL http://dx.doi.org/10.1007/978-94-017-1120-3_2.

U. Kosinska Eriksson, G. Fischer, R. Friemann, G. Enkavi, E. Tajkhorshid, and R. Neutze. Subangstrom resolution x-ray structure details aquaporin-water interactions. *Science*, 340(6138):1346–1349, 2013. doi: 10.1126/science.1234306. URL http://www.sciencemag.org/content/340/6138/1346.abstract.

P. Kührová, A. D. Simone, M. Otyepka, and R. B. Best. Force-field dependence of chignolin folding and misfolding: Comparison with experiment and redesign. *Biophysical Journal*, 102(8):1897 – 1906, 2012. ISSN 0006-3495. doi: http://dx.doi.org/10.1016/j.bpj.2012.03.024. URL http://www.sciencedirect.com/science/article/pii/S0006349512003359.

O. F. Lange, D. van der Spoel, and B. L. de Groot. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophysical Journal*, 99(2):647–655, 2010.

C. Levinthal. How to Fold Graciously. In J. T. P. Debrunnder and E. Munck, editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a*

meeting held at Allerton House, Monticello, Illinois, pages 22–24. University of Illinois Press, 1969.

D.-W. Li and R. Brüschweiler. Certification of molecular dynamics trajectories with NMR chemical shifts. *Journal of Physical Chemistry Letters*, 1(1):246–248, 2009.

D.-W. Li and R. Brüschweiler. Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. *Journal of Chemical Theory and Computation*, 7(6):1773–1782, 2011. doi: 10.1021/ct200094b. URL `http://pubs.acs.org/doi/abs/10.1021/ct200094b`.

D.-W. Li and R. Brüschweiler. NMR-based protein potentials. *Angewandte Chemie*, 122(38):6930–6932, 2010. ISSN 1521-3757. doi: 10.1002/ange.201001898. URL `http://dx.doi.org/10.1002/ange.201001898`.

S. Lifson and A. Warshel. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and nalkane molecules. *The Journal of Chemical Physics*, 49(11):5116–5129, 1968. doi: http://dx.doi.org/10.1063/1.1670007. URL `http://scitation.aip.org/content/aip/journal/jcp/49/11/10.1063/1.1670007`.

K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins-Structure Function and Bioinformatics*, 78(8):1950–1958, 2010.

K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2), 2012. doi: 10.1371/journal.pone.0032131.

G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *Journal of the American Chemical Society*, 104(17):4546–4559, 1982. doi: 10.1021/ja00381a009. URL `http://dx.doi.org/10.1021/ja00381a009`.

F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proceedings of the National Academy of Sciences*, 105(7):2369–2374, 2008. doi: 10.1073/pnas.0711908105. URL `http://www.pnas.org/content/105/7/2369.abstract`.

S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins-Structure Function and Genetics*, 40(3): 389–408, 2000.

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

A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998. doi: 10.1021/jp973084f. URL `http://dx.doi.org/10.1021/jp973084f`. PMID: 24889800.

A. D. MacKerell, M. Feig, and C. L. Brooks. 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*, 25(11):1400–1415, 2004a. ISSN 1096-987X. doi: 10.1002/jcc.20065. URL `http://dx.doi.org/10.1002/jcc.20065`.

A. D. MacKerell, M. Feig, and C. L. Brooks. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical*

Society, 126(3):698–699, 2004b. doi: 10.1021/ja036959e. URL http://dx.doi.org/10.1021/ja036959e. PMID: 14733527.

J. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. Hauser, and C. Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 1 (1):1, 2015.

C. Martinez. *Enhanced Backbone Dihedral Parameters for Protein Simulations*. Doctoral thesis, Stony Brook University, 2014.

E. Miclet, J. Boisbouvier, and A. Bax. Measurement of eight scalar and dipolar couplings for methine-methylene pairs in proteins and nucleic acids. *Journal of Biomolecular Nmr*, 31(3):201–216, 2005.

F. Neese. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1):73–78, 2012. ISSN 1759-0884. doi: 10.1002/wcms.81. URL http://dx.doi.org/10.1002/wcms.81.

P. S. Nerenberg and T. Head-Gordon. Optimizing protein-solvent force fields to reproduce intrinsic conformational preferences of model peptides. *Journal of Chemical Theory and Computation*, 7(4):1220–1230, 2011.

H. Nguyen, D. R. Roe, and C. Simmerling. Improved Generalized Born solvent model parameters for protein simulations. *Journal of Chemical Theory and Computation*, 9(4):2020–2034, 2013.

H. Nguyen, J. Maier, and C. Simmerling. Proteins of diverse topologies can be folded atomistically in a matter of days. *JACS Letters*, 0(0):0, 2014.

M. Nilges. Calculation of protein structures with ambiguous distance restraints. automated assignment of ambiguous {NOE} crosspeaks and disulphide connectivities. *Journal of Molecular Biology*, 245(5):645 – 660, 1995. ISSN 0022-2836. doi: http://dx.doi.org/10.1006/jmbi.1994.0053. URL http://www.sciencedirect.com/science/article/pii/S0022283684700532.

A. Patgiri, A. L. Jochim, and P. S. Arora. A hydrogen bond surrogate approach for stabilization of short peptide sequences in α-helical conformation. *Accounts of Chemical Research*, 41(10):1289–1300, 2008. doi: 10.1021/ar700264k. URL http://pubs.acs.org/doi/abs/10.1021/ar700264k. PMID: 18630933.

R. PDB. Holdings report. http://www.rcsb.org/pdb/statistics/holdings.do. Accessed: 2014-11-12.

A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. C. III, C. A. Laughton, and M. Orozco. Refinement of the {AMBER} force field for nucleic acids: Improving the description of α/γ conformers. *Biophysical Journal*, 92(11):3817 – 3829, 2007. ISSN 0006-3495. doi: http://dx.doi.org/10.1529/biophysj.106.097782. URL http://www.sciencedirect.com/science/article/pii/S0006349507711827.

A. Perez, J. MacCallum, and K. A. Dill. Meld: Modeling peptide-protein interactions. *Biophysical Journal*, 104(2, Supplement 1):399a –, 2013. ISSN 0006-3495. doi: http://dx.doi.org/10.1016/j.bpj.2012.11.2224. URL http://www.sciencedirect.com/science/article/pii/S0006349512034704.

C. Perez, F. Lohr, H. Ruterjans, and J. M. Schmidt. Self-consistent karplus parametrization of (3)J couplings depending on the polypeptide side-chain torsion chi(1). *Journal of the American Chemical Society*, 123(29):7081–7093, 2001.

M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. Structure of haemoglobin: A three-dimensional fourier synthesis at 5.5-[angst]. resolution, obtained by X-ray analysis. *Nature*, 185 (4711):416–422, 1960. 10.1038/185416a0.

S. Piana, K. Lindorff-Larsen, and D. E. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47–L49, 3 2011. doi: 10.1016/j.bpj.2011.03.051. URL http://www.cell.com/biophysj/abstract/S0006-3495(11)00409-7.

S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences*, 110 (15):5915–5920, 2013. doi: 10.1073/pnas.1218321110. URL http://www.pnas.org/content/110/15/5915.abstract.

K. S. Pitzer. Potential energies for rotation about single bonds. *Discuss. Faraday Soc.*, 10:66–73, 1951. doi: 10.1039/DF9511000066. URL http://dx.doi.org/10.1039/DF9511000066.

J. J. Prompers and R. Brüschweiler. General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. *Journal of the American Chemical Society*, 124(16): 4522–4534, 2002. doi: 10.1021/ja012750u. URL http://pubs.acs.org/doi/abs/10.1021/ja012750u.

G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95 – 99, 1963. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/S0022-2836(63)80023-6. URL http://www.sciencedirect.com/science/article/pii/S0022283663800236.

P. Ren and J. W. Ponder. Polarizable atomic multipole water model for molecular mechanics simulation. *The Journal of Physical Chemistry B*, 107(24): 5933–5947, 2003. doi: 10.1021/jp027815+. URL http://pubs.acs.org/doi/abs/10.1021/jp027815+.

D. R. Roe and T. E. Cheatham. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095, 2013. doi: 10.1021/ct400341p. URL http://pubs.acs.org/doi/abs/10.1021/ct400341p.

J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013. doi: 10.1021/ct400314y. URL `http://dx.doi.org/10.1021/ct400314y`.

P. Salvador, I.-H. M. Tsai, and J. J. Dannenberg. J-coupling constants for a trialanine peptide as a function of dihedral angles calculated by density functional theory over the full Ramachandran space. *Phys. Chem. Chem. Phys.*, 13:17484–17493, 2011. doi: 10.1039/C1CP20520J. URL `http://dx.doi.org/10.1039/C1CP20520J`.

F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics*, 2(6):110–4, 1946.

M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery. General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11):1347–1363, 1993. ISSN 1096-987X. doi: 10.1002/jcc.540141112. URL `http://dx.doi.org/10.1002/jcc.540141112`.

H. Schwalbe, S. B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C. M. Dobson, C. Redfield, and L. J. Smith. A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Science*, 10(4): 677–688, 2001.

S. A. Showalter and R. Bruschweiler. Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *Journal of Chemical Theory and Computation*, 3(3):961–975, 2007.

K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins-Structure Function and Bioinformatics*, 37(3):171–176, 1999.

L. J. Smith, M. J. Sutcliffe, C. Redfield, and C. M. Dobson. Analysis of phi and chi-1 torsion angles for hen lysozyme in solution from h-1-NMR spin spin coupling-constants. *Biochemistry*, 30(4):986–996, 1991.

W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982. doi: http://dx.doi.org/10.1063/1.442716. URL `http://scitation.aip.org/content/aip/journal/jcp/76/1/10.1063/1.442716`.

E. J. Thompson, A. J. DePaul, S. S. Patel, and E. J. Sorin. Evaluating molecular mechanical potentials for helical peptides and proteins. *PLoS ONE*, 5 (4):e10056, 2010.

E. O. Travis. Python for scientific computing. *Computing in Science and Engineering*, 9(3):10–20, 2007.

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

E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36(suppl 1):D402–D408, 2008. doi: 10.1093/nar/gkm957. URL `http://nar.oxfordjournals.org/content/36/suppl_1/D402.abstract`.

O. Vahtras, J. Almlöf, and M. Feyereisen. Integral approximations for lcao-scf calculations. *Chemical Physics Letters*, 213(5–6):514 – 518, 1993. ISSN 0009-2614. doi: http://dx.doi.org/10.1016/0009-2614(93)89151-7. URL `http://www.sciencedirect.com/science/article/pii/0009261493891517`.

N. Vajpai, M. Gentner, J.-r. Huang, M. Blackledge, and S. Grzesiek. Side-chain $\chi 1$ conformations in urea-denatured ubiquitin and protein G from 3J

coupling constants and residual dipolar couplings. *Journal of the American Chemical Society*, 132(9):3196–3203, 2010. doi: 10.1021/ja910331t. URL http://dx.doi.org/10.1021/ja910331t. PMID: 20155903.

S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. Structure of ubiquitin refined at 1.8 a resolution. *Journal of Molecular Biology*, 194(3):531–544, 1987.

M. Wall. GAlib: A C++ library of genetic algorithm components. *Mechanical Engineering Department, Massachusetts Institute of Technology*, 1996.

D. Wang, K. Chen, J. L. Kulp, and P. S. Arora. Evaluation of biologically relevant short alpha-helices stabilized by a main-chain hydrogen-bond surrogate. *Journal of the American Chemical Society*, 128(28):9248–9256, 2006.

J. Wang and P. A. Kollman. Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry*, 22(12):1219–1228, 2001. ISSN 1096-987X. doi: 10.1002/jcc.1079. URL http://dx.doi.org/10.1002/jcc.1079.

J. Wang, Q. Cai, Y. Xiang, and R. Luo. Reducing grid dependence in finite-difference Poisson–Boltzmann calculations. *Journal of Chemical Theory and Computation*, 8(8):2741–2751, 2012. doi: 10.1021/ct300341d. URL http://dx.doi.org/10.1021/ct300341d.

J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984. doi: 10.1021/ja00315a051. URL http://pubs.acs.org/doi/abs/10.1021/ja00315a051.

S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force-field for simulations of proteins and nucleic-acids. *Journal of Computational Chemistry*, 7(2):230–252, 1986.

B. L. Welch. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

L. Wickstrom, A. Okur, and C. Simmerling. Evaluating the performance of the ff99SB force field based on NMR scalarcoupling data. *Biophysical Journal*, 97(3):853–856, 2009.

J. Wirmer and H. Schwalbe. Angular dependence of (1)J(N-i,C-alpha i) and (2)J(N-i,C alpha(i-1)) coupling constants measured in J-modulated HSQCs. *Journal of Biomolecular Nmr*, 23(1):47–55, 2002.

A. Wlodawer, J. Walter, R. Huber, and L. Sjölin. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. *Journal of Molecular Biology*, 180(2):301 – 329, 1984. ISSN 0022-2836. doi: http://dx.doi.org/10.1016/S0022-2836(84)80006-6. URL `http://www.sciencedirect.com/science/article/pii/S0022283684800066`.

X. Xiao, T. Zhu, C. G. Ji, and J. Z. H. Zhang. Development of an effective polarizable bond method for biomolecular simulation. *The Journal of Physical Chemistry B*, 117(48):14885–14893, 2013. doi: 10.1021/jp4080866. URL `http://pubs.acs.org/doi/abs/10.1021/jp4080866`.

A. C. M. Young, J. C. Dewan, C. Nave, and R. F. Tilton. Comparison of radiation-induced decay and structure refinement from X-ray data collected from lysozyme crystals at low and ambient-temperatures. *Journal of Applied Crystallography*, 26:309–319, 1993.

M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham, and P. Jurecka. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of Chemical Theory and Computation*, 7(9):2886–2902, 2011.

M. Zgarbova, F. J. Luque, J. Sponer, T. E. Cheatham, M. Otyepka, and P. Jurecka. Toward improved description of DNA backbone: Revisiting epsilon and zeta torsion force field parameters. *Journal of Chemical Theory and Computation*, 9(5):2339–2354, 2013.

W. Zhang, T. Hou, C. Schafmeister, W. S. Ross, and D. A. Case. LEaP, 2010.

C.-Y. Zhou, F. Jiang, and Y.-D. Wu. Residue-specific force field based on protein coil library. RSFF2: Modification of AMBER ff99SB. *The Journal of Physical Chemistry B*, 119(3):1035–1047, 2015. doi: 10.1021/jp5064676. URL `http://dx.doi.org/10.1021/jp5064676`. PMID: 25358113.

Appendix A

Additional ff14SB analysis

This appendix shows the sampling of ff99SB and ff14SB in terms of ϕ and ψ 2d histograms (Figures A.1 to A.5). Most ϕ/ψ distributions are roughly equivalent. Only occasionally ff14SB drives the sampling toward $\phi = -60°$.

Also shown are RMSDs and RMSD histograms in terms of probability density function, and cumulative distribution function (Figures A.6 to A.9). Although RMSD histograms are generally comparable, within error bars, there are some slight qualitative differences. Most noticeably, lysozyme backbone RMSDs fluctuate for every force field but ff14SB, even ff14SBonlysc. But for ff14SB, as depicted in the histograms, the RMSD remains quite low, below 1 Å except for one excursion to a max of about 2 Å for less than 10 ns in run 3. Due to large uncertainties for other force fields, it is unclear how significant this difference is quantitatively. Additionally, it is unclear whether ff14SB might be trapped in the crystal structure, whereas other force fields may have found a conformation that is sampled in solution. Still, all force fields sample RMSDs predominantly below 2 Å for all systems.

A.1 Protein Ramachandran histograms

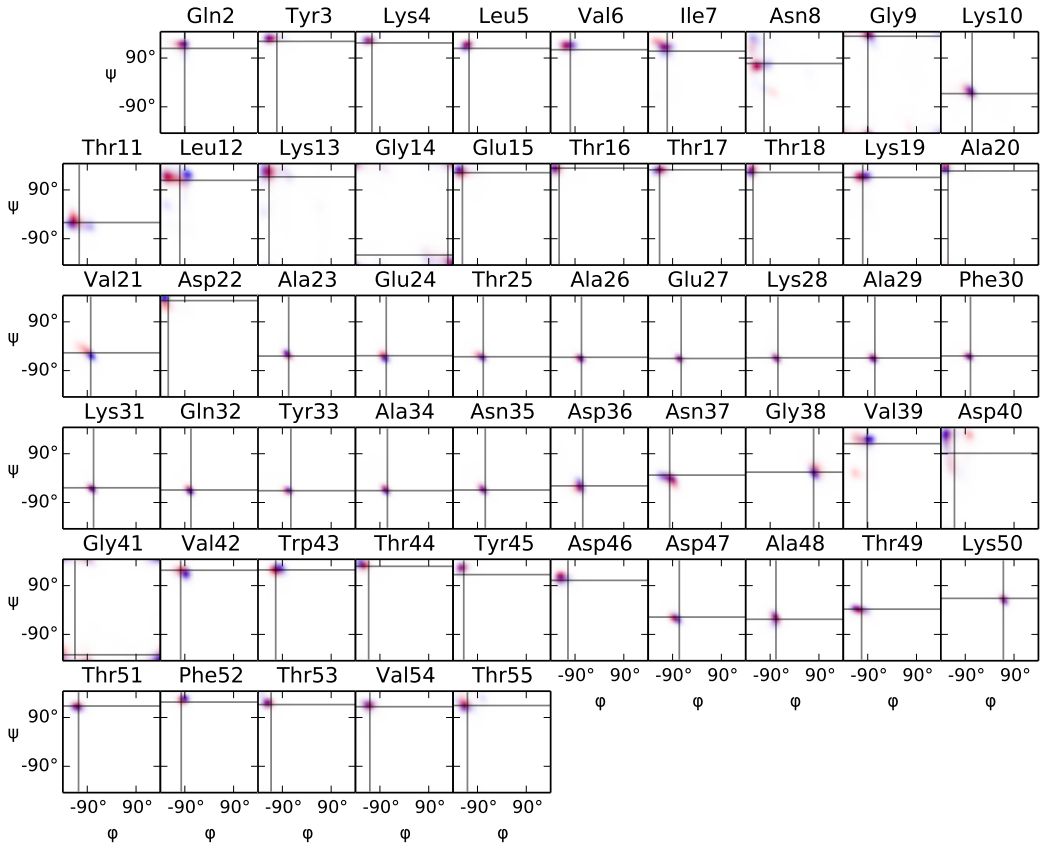

Figure A.1: Ramachandran histograms of each residue in GB3 from four simulations each with ff99SB (red) and ff14SB (blue). Vertical and horizontal lines indicate the experimental ϕ and ψ, respectively.

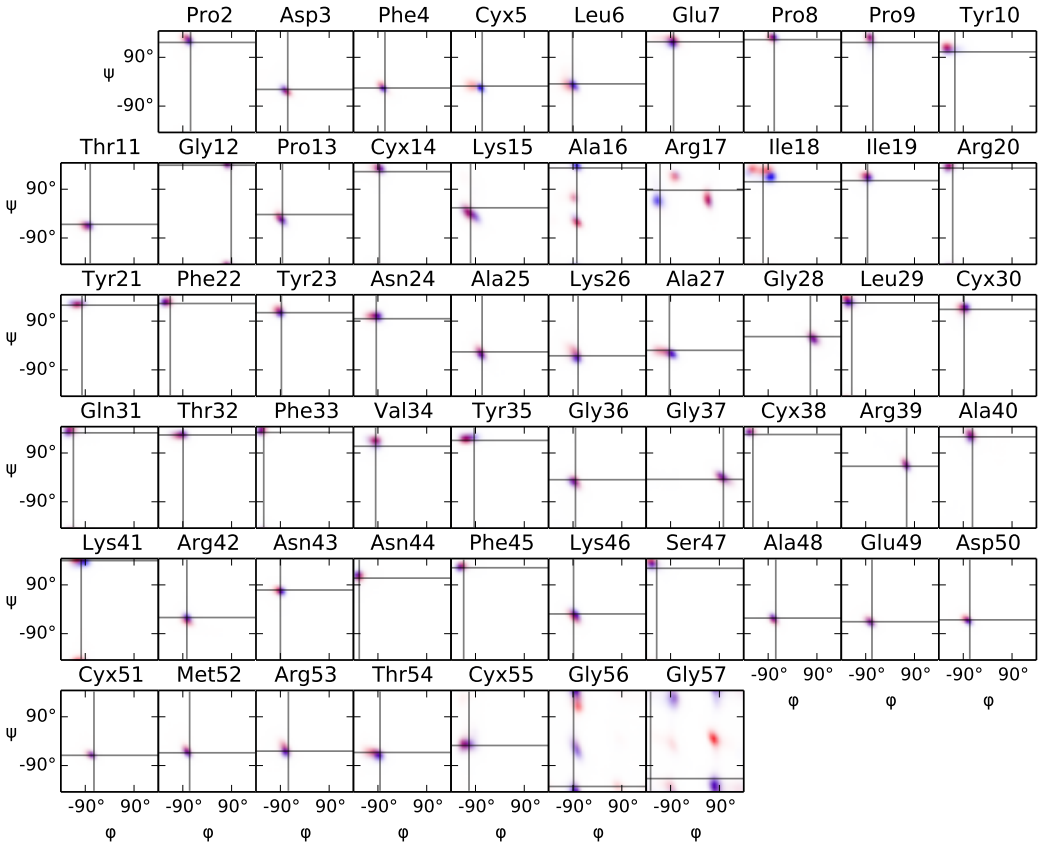

Figure A.2: Ramachandran histograms of each residue in bovine pancreatic trypsin inhibitor from four simulations each with ff99SB (red) and ff14SB (blue). Vertical and horizontal lines indicate the experimental ϕ and ψ, respectively.

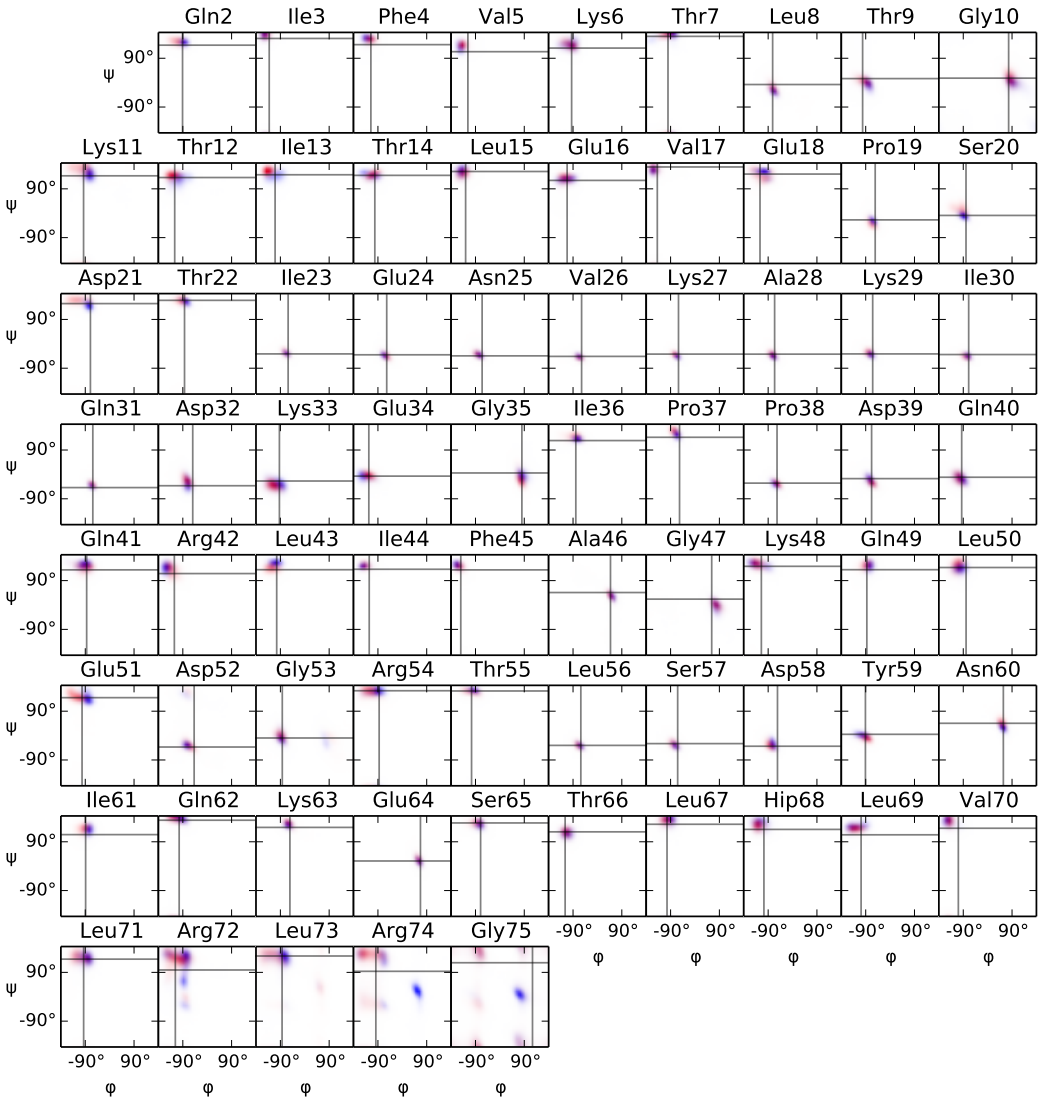

Figure A.3: Ramachandran histograms of each residue in ubiquitin from four simulations each with ff99SB (red) and ff14SB (blue). Vertical and horizontal lines indicate the experimental ϕ and ψ, respectively.

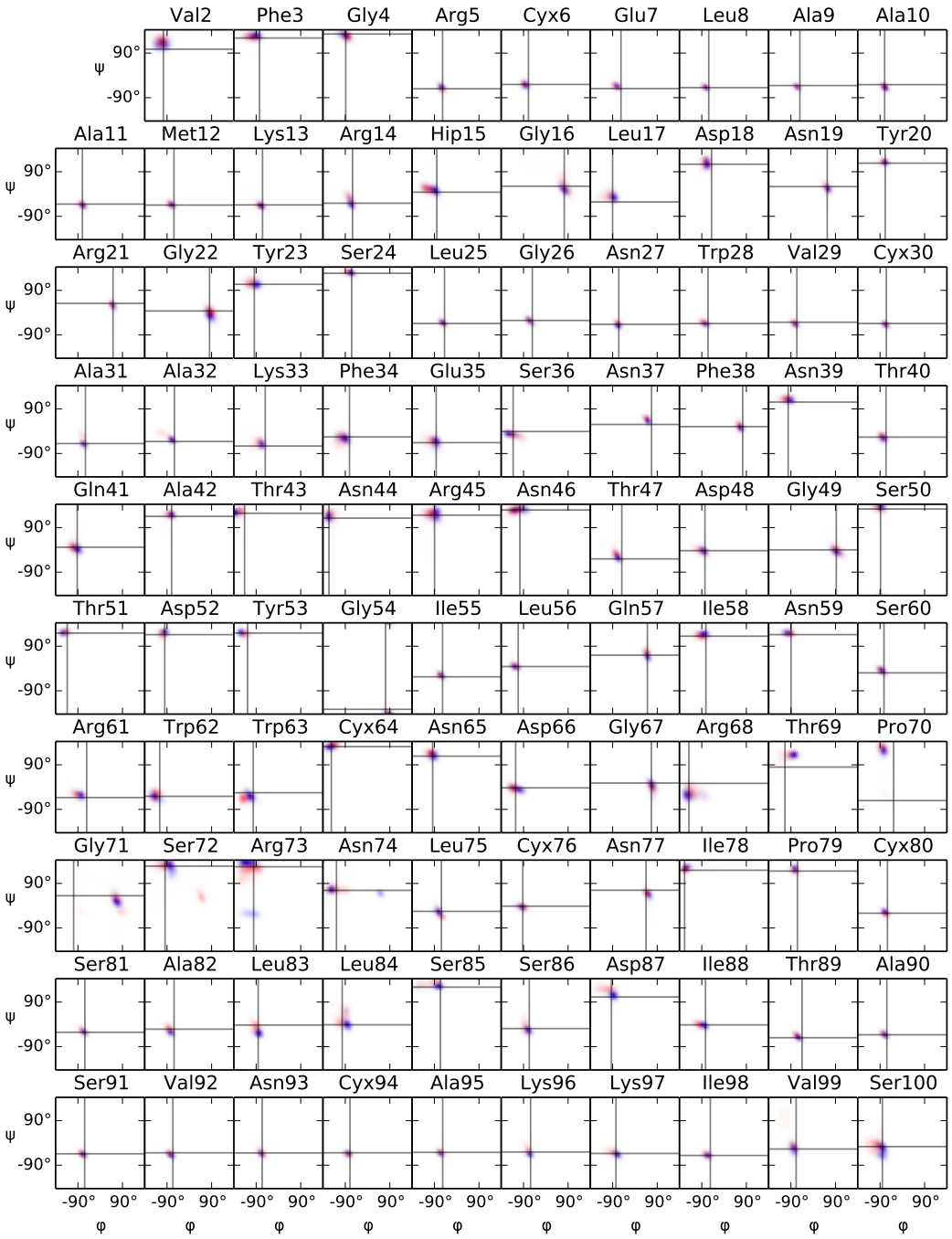

Figure A.4: Ramachandran histograms of each residue in lysozyme from four simulations each with ff99SB (red) and ff14SB (blue), for residues 2 to 100. Vertical and horizontal lines indicate the experimental ϕ and ψ, respectively.

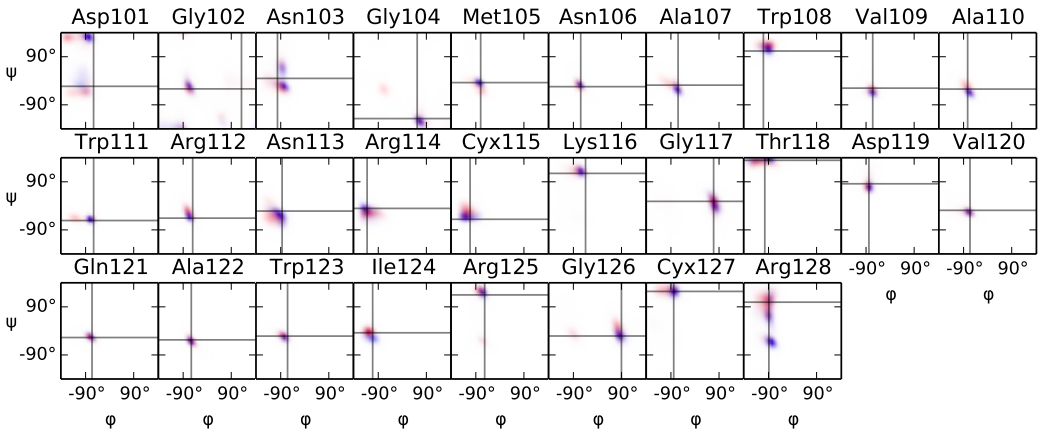

Figure A.5: Ramachandran histograms of each residue in lysozyme from four simulations each with ff99SB (red) and ff14SB (blue), for residues 101 to 128. Vertical and horizontal lines indicate the experimental ϕ and ψ, respectively.

A.2 Protein backbone RMSDs

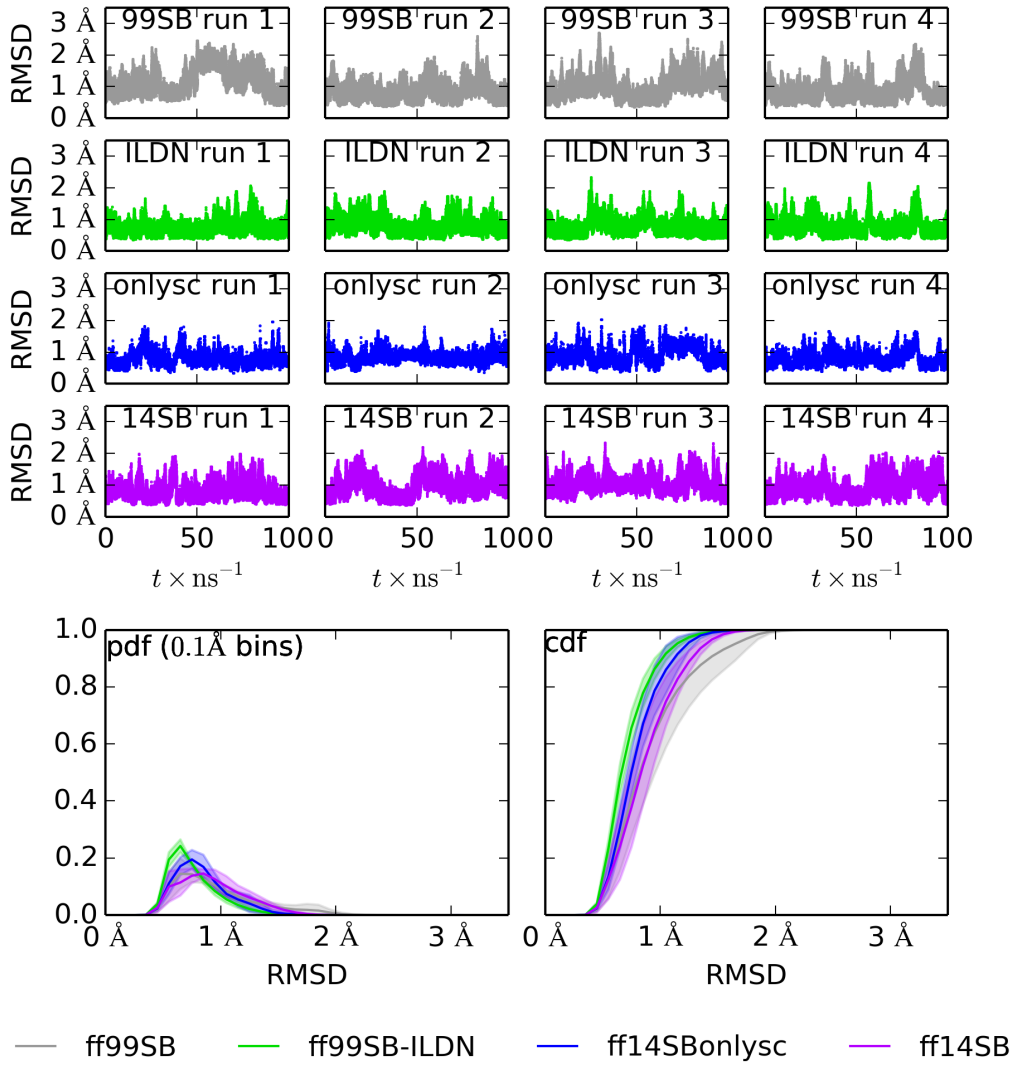

Figure A.6: GB3 backbone (N, Cα, C) RMSD to 1P7E [Ulmer et al., 2003] for four runs of ff99SB (top row, gray), ff99SB-ILDN (second row, green), ff14SBonlysc (third row, blue), and ff14SB (fourth row, purple). The probability density function (pdf) and cumulative distribution function (cdf) are plotted for each force field in the bottom row.

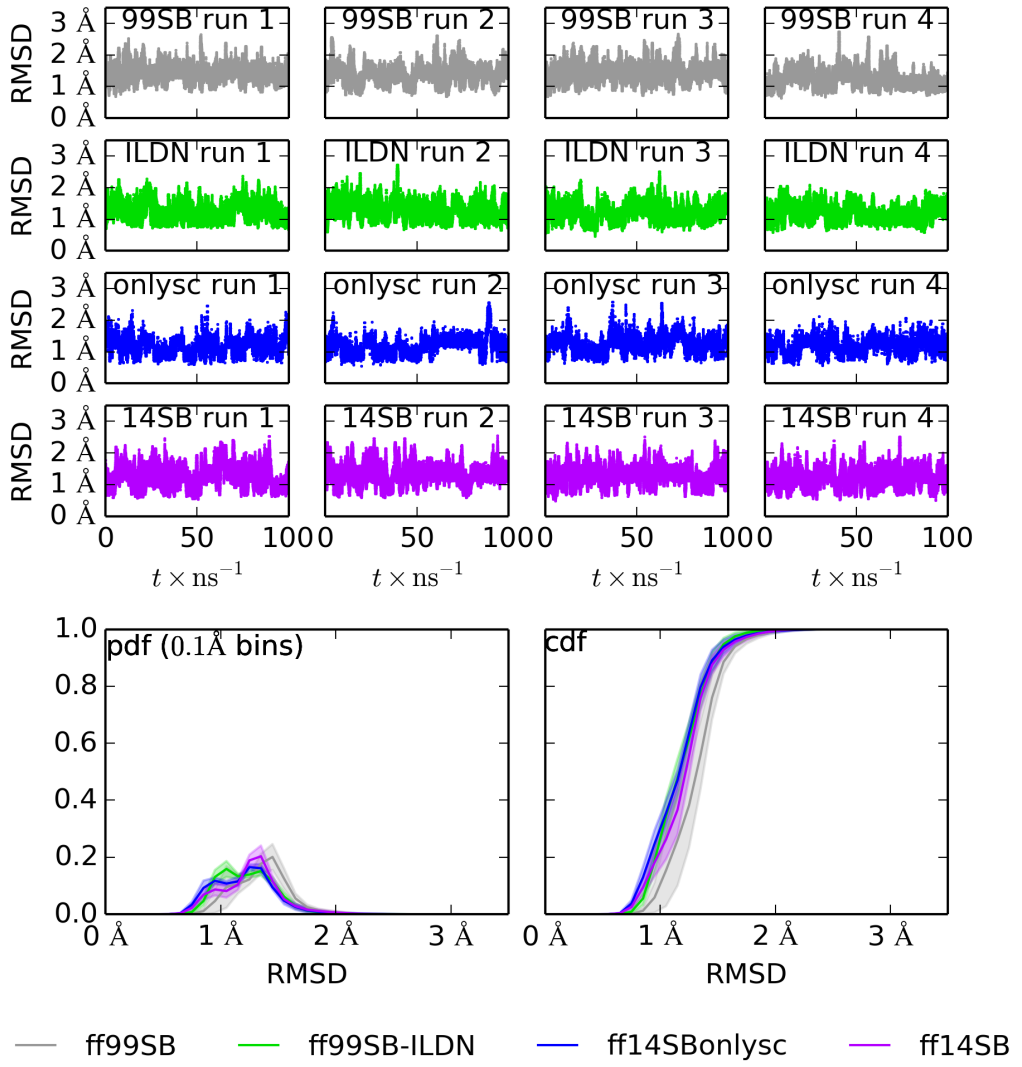

Figure A.7: BPTI backbone (N, Cα, C) RMSD to 5PTI [Wlodawer et al., 1984] for four runs of ff99SB (top row, gray), ff99SB-ILDN (second row, green), ff14SBonlysc (third row, blue), and ff14SB (fourth row, purple). The probability density function (pdf) and cumulative distribution function (cdf) are plotted for each force field in the bottom row.

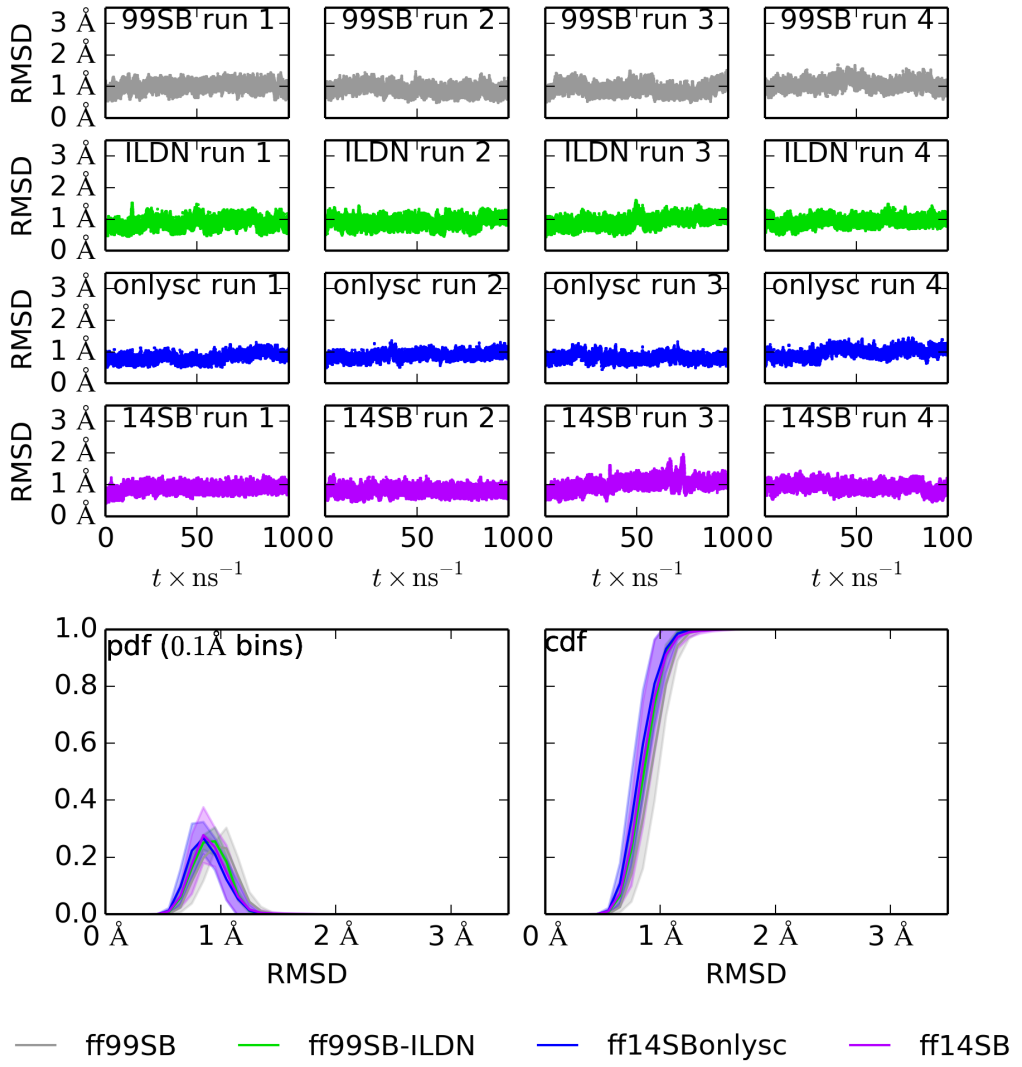

Figure A.8: Ubiquitin backbone (N, Cα, C) RMSD to 1UBQ [Vijay-Kumar et al., 1987] for four runs of ff99SB (top row, gray), ff99SB-ILDN (second row, green), ff14SBonlysc (third row, blue), and ff14SB (fourth row, purple). The probability density function (pdf) and cumulative distribution function (cdf) are plotted for each force field in the bottom row.

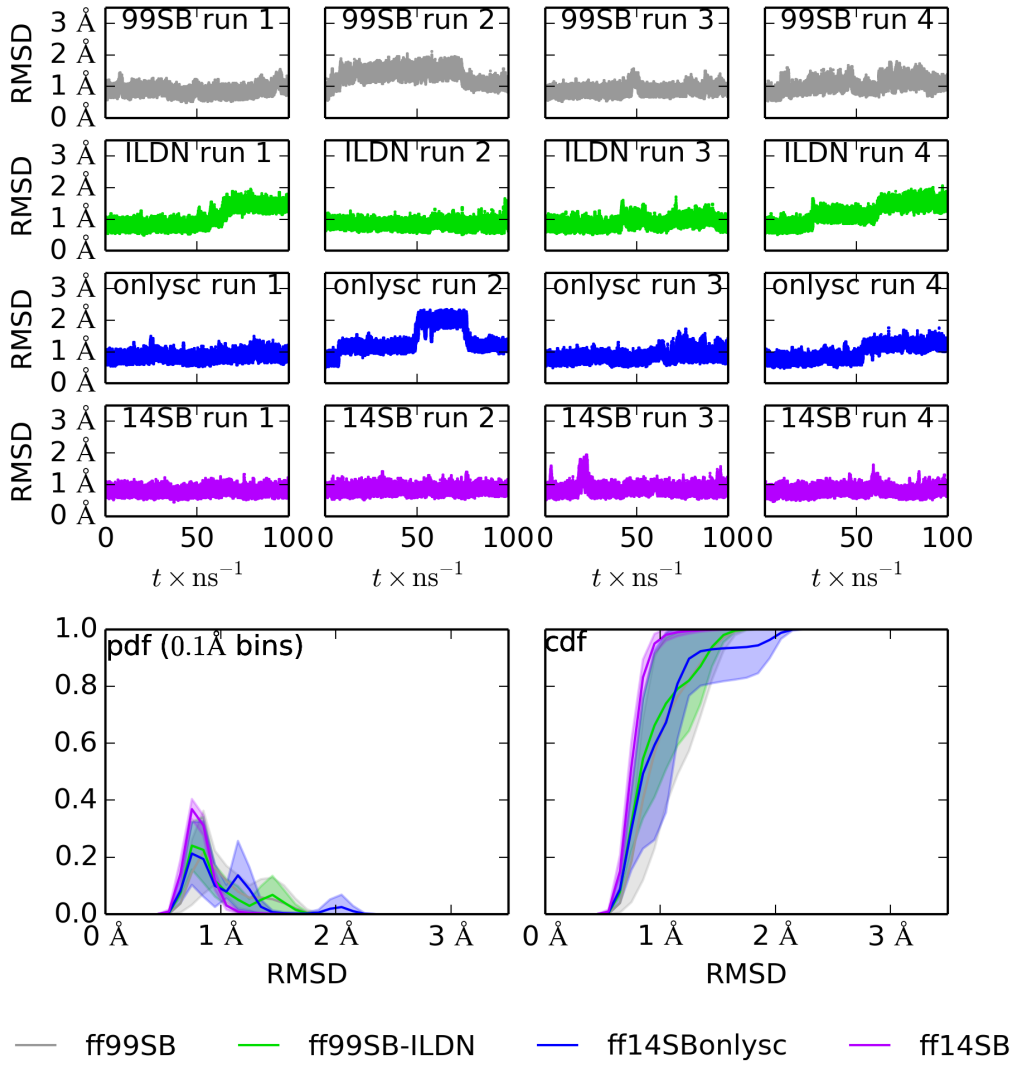

Figure A.9: Lysozyme backbone (N, Cα, C) RMSD to 6LYT [Young et al., 1993] for four runs of ff99SB (top row, gray), ff99SB-ILDN (second row, green), ff14SBonlysc (third row, blue), and ff14SB (fourth row, purple). The probability density function (pdf) and cumulative distribution function (cdf) are plotted for each force field in the bottom row.

Appendix B

ff12SB versus ff14SB

The ff12SB that was bundled with AMBER12 [Case et al., 2005, 2012] differs from ff14SB presented here in three ways. First, in an effort to reduce the degree of in vacuo non-bonded interactions in the ff12SB training set—particularly between side chain and backbone, as might arise when polar atoms deviate from planarity—all possible four-atom dihedrals were frozen. This had the unfortunate effect of locking angles centered on branching atoms, actually worsening steric clashes and limiting agreement with quantum mechanics energies. By restraining only one dihedral per rotable side chain bond in the ff14SB training set, we achieved a finer fit to quantum energies and better agreement with experimental scalar couplings.

Second, in ff12SB corrections with periodicities corresponding to their offsets around a bond were allowed to differ. As corrections with such a relation have the ability to cancel, many of these corrections had large and potentially arbitrary magnitudes. In very small peptides, $\chi 1$ corrections were observed to alter the relative positions of the main chain N and C, affecting backbone dynamics in ways that are difficult to predict. Therefore, corrections that are in phase were forced to be identical in ff14SB.

Third, ff12SB side chain corrections allowed phase shifts other than 0° or 180°. While this permitted better agreement with quantum energies, it prohibits use of the same parameters as molecules change chirality. Therefore, ff14SB corrections only employ 0° or 180° phase shifts.

The difference in ff12SB and ff14SB training sets is significant for a few residues. In Supplementary Figure 1, we show the errors of ff12SB and ff14SB against the sets of energies used to train each. Naturally, errors are lowest for

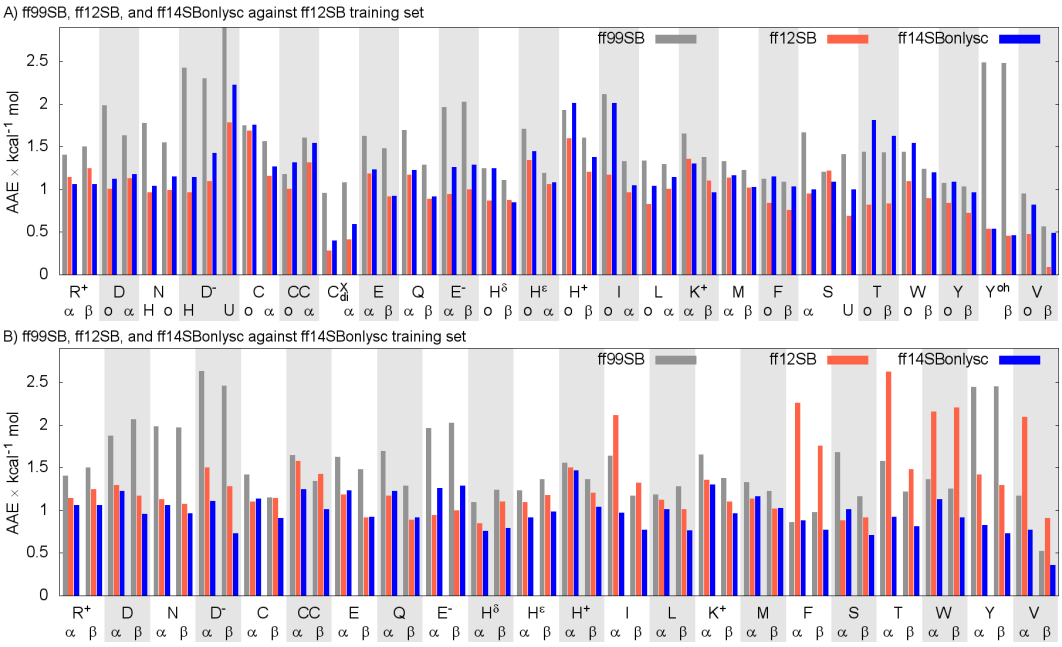

Figure B.1: The errors of ff12SB and ff14SB against the quantum mechanics energies used to train ff12SB (A) and ff14SB (B).

each force field against its own training set. But ff12SB reproduces the energies of several residues in the ff14SB training set notably poorly. Whereas ff14SB is at most 1.2 ± 0.1 times the error of ff99SB (for threonine) against the ff12SB training set, ff12SB has 2.2 ± 0.4 times the error of ff99SB against ff14SB phenylalanine targets. After that, ff12SB has 1.8 ± 0.0 times the error for valine, 1.7 ± 0.1 times for tryptophan, and 1.4 ± 0.2 times for threonine. The ff12SB training generated parameters less appropriate for the ff14SB target energies of several residues than ff99SB.

What's important is how such differences in the training affect accuracy of simulations in real systems. In Supplementary Figure 2, we illustrate some of these differences in GB3, ubiquitin, and lysozyme in terms of normalized error versus $\chi 1$ scalar couplings [Lindorff-Larsen et al., 2010, Berndt et al., 1992, Grimshaw, 1999, Hu and Bax, 1997, Chou et al., 2003, Miclet et al., 2005, Schwalbe et al., 2001, Smith et al., 1991]. In fact, some residues improved dramatically, with aspartate and threonine normalized scalar coupling errors being $39 \pm 7\,\%$ and $34 \pm 8\,\%$ better with ff14SB, respectively. Lysine, arginine, isoleucine, serine, valine, leucine, and tyrosine also improve by $29 \pm 7\,\%$, $22 \pm 15\,\%$, $20 \pm 11\,\%$, $19 \pm 8\,\%$, $16 \pm 12\,\%$, $14 \pm 12\,\%$, and $8 \pm 2\,\%$, although some of these differences approach insignificance. Phenylalanine and trypophan are worse by $17 \pm 2\,\%$ and $5 \pm 2\,\%$, respectively, though the normalized

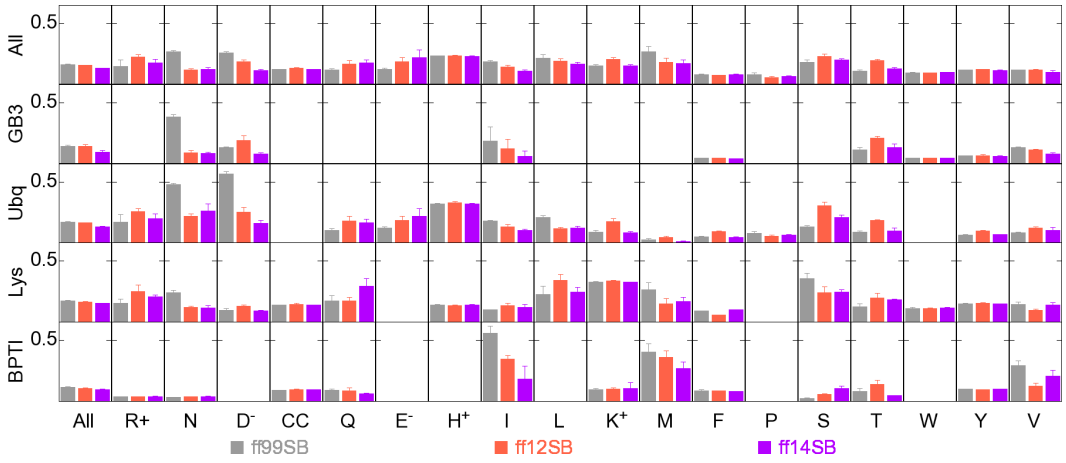

Figure B.2: The mean absolute errors of simulations with ff99SB (gray), ff12SB (tomato), and ff14SB (purple) compared to experimental J couplings, normalized by Karplus curve range, of all residues and each residue horizontally in all and each of GB3, ubiquitin (Ubq), and lysozyme (Lys) vertically. Error bars represent the standard error in the mean absolute errors of each independent run.

errors in ff14SB are still quite small—0.08 ± 0.00 and 0.10 ± 0.00 for phenylalanine and tryptophan, respectively. Altogether, ff14SB better reproduces side chain scalar couplings by $17 \pm 3\,\%$ with a normalized error of 0.13 ± 0.00, compared with ff12SB's normalized error of 0.16 ± 0.00.

Appendix C

Analytical CMAP fitting

The CMAP correction utilizes a bicubic interpolation scheme to evaluate any arbitrary point on a surface from a fixed grid of data points. In this scheme, there are actually many bicubic interpolations for each square on the grid defined by four neighboring points. To ensure the bicubic interpolated surface is smooth and continuous, the values, derivatives, and mixed double derivative are ensured to be identical at the interfaces of any two squares.

First, cubic splines are fit to each row and each column of points on the grid, to allow one to obtain derivatives. Typically, each cubic spline is fit such that the values, first, and second derivatives match at each point in one dimension. Through some algebraic manipulation, this amounts to solving the system of equations:

$$\begin{bmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{bmatrix} \vec{D} = \begin{pmatrix} 3(y_2 - y_N) \\ 3(y_3 - y_1) \\ \vdots \\ 3(y_N - y_{N-2}) \\ 3(y_1 - y_{N-1}) \end{pmatrix} \tag{C.1}$$

This means that $\vec{D}$ can be summarized as:

$$\vec{D} = \mathbf{M}_{141}^{-1} \begin{pmatrix} 3(y_2 - y_N) \\ 3(y_3 - y_1) \\ \vdots \\ 3(y_N - y_{N-2}) \\ 3(y_1 - y_{N-1}) \end{pmatrix} \tag{C.2}$$

The bicubic interpolation takes the form:

$$f(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \tag{C.3}$$

One can build a sixteen-element vector out of the a_{ij} values, which we will call $\vec{a}$. $f(x,y)$ can then be seen as the dot product between $\vec{a}$ and another vector containing the appropriate $x^i y^j$, called $\vec{x}$.

$$\vec{a} = \begin{pmatrix} a_{00} & a_{10} & a_{20} & a_{30} & a_{01} & a_{11} & \cdots & a_{23} & a_{33} \end{pmatrix}^\top \tag{C.4}$$

$$\vec{x} = \begin{pmatrix} 1 & x & x^2 & x^3 & y & xy & \cdots & x^2y^3 & x^3y^3 \end{pmatrix}^\intercal \tag{C.5}$$

Additionally, one can begin to establish some identities between the a_{ij} values and grid points defining square boundaries, by evaluating the function and its derivatives at each point. For instance,

$$f(0,0) = a_{00} \tag{C.6}$$

$$f(1,0) = a_{00} + a_{10} + a_{20} + a_{30} \tag{C.7}$$

$$f(0,1) = a_{00} + a_{01} + a_{02} + a_{03} \tag{C.8}$$

$$f(1,1) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij} \tag{C.9}$$

One can similarly evaluate the derivatives with respect to x ($f_x(x,y)$) or with respect to y ($f_y(x,y)$).

$$f_x(x,y) = \sum_{i=1}^{3}\sum_{j=0}^{3} i x^{i-1} y^j \tag{C.10}$$

$$f_y(x,y) = \sum_{i=0}^{3}\sum_{j=1}^{3} j x^i y^{j-1} \tag{C.11}$$

As there are sixteen possible a_{ij}, there must be sixteen equations to define $\vec{a}$. For $f(x,y)$ and its single derivatives, we can establish four identities each—a total of twelve. The remaining four can be established from the cross derivative

$\frac{\partial^2 f}{\partial x \partial y}$.

$$f_{xy}(x,y) = \sum_{i=1}^{3}\sum_{j=1}^{3} ija_{ij}x^{i-1}y^{i-1} \tag{C.12}$$

If one considers each of these functions evaluated at $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$, one can set up a relationship between $\vec{a}$ and the values at the grid points by the matrix $\mathbf{A}$ (Equation C.13) (Equation C.14).

$$\mathbf{A} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 0 & 2 & 4 & 6 & 0 & 3 & 6 & 9
\end{bmatrix} \tag{C.13}$$

$$\mathbf{A}\vec{a} = \begin{pmatrix} f(0,0) \\ f(1,0) \\ f(0,1) \\ f(1,1) \\ f_x(0,0) \\ f_x(1,0) \\ f_x(0,1) \\ f_x(1,1) \\ f_y(0,0) \\ f_y(1,0) \\ f_y(0,1) \\ f_y(1,1) \\ f_{xy}(0,0) \\ f_{xy}(1,0) \\ f_{xy}(0,1) \\ f_{xy}(1,1) \end{pmatrix} \tag{C.14}$$

In the application of a bicubic interpolation, one knows $f(x, y)$ at the boundaries of each square, and can determine the derivatives from cubic splines fit through each point. In this way, the derivatives depend on the values of $f(x, y)$. Each evaluation only requires the dot product of $\vec{a}$ and $\vec{x}$, as established above. Thus one typically would be interested in solving for $\vec{a}$, which turns out to be a rather simple solution: $\vec{a}$ is $\mathbf{A}^{-1}$ times the vector of $f(x, y)$ and its derivatives.

If, however, one wishes to solve for a grid that will provide an $\vec{a}$ that optimizes agreement with an arbitrary set of data, one cannot simply solve for $\vec{a}$, as this vector has certain properties. Nor can one simply solve for $f(x, y)$ and its derivatives; the derivatives must naturally follow from $f(x, y)$. Thus, if one wants to solve for a set of values defining a bicubic interpolation for the CMAP correction of protein conformations, one must state the problem of $f(x, y)$ in terms of the grid points alone.

The single derivatives can be expressed in terms of $f(x, y)$ using Equation C.2. For example, the set of derivatives with respect to ϕ along a given ψ may be expressed as

$$\begin{pmatrix} f_x(0,0)_{-180^\circ,\psi} \\ f_x(0,0)_{-165^\circ,\psi} \\ \vdots \\ f_x(0,0)_{165^\circ,\psi} \end{pmatrix} = \mathbf{M}_{141}^{-1} \begin{pmatrix} 3(f(0,0)_{-165^\circ,\psi} - f(0,0)_{165^\circ,\psi}) \\ 3(f(0,0)_{-150^\circ,\psi} - f(0,0)_{-180^\circ,\psi}) \\ \vdots \\ 3(f(0,0)_{180^\circ,\psi} - f(0,0)_{150^\circ,}) \end{pmatrix} \tag{C.15}$$

Thus, anywhere one must multiply by $f_x(0,0)_{\phi,\psi}$, one may express that by adding the appropriate row of Equation C.15.

$$f(\phi,\psi) = \begin{pmatrix}\vec{x}_{aa} & \vec{x}_{ba} & \cdots & \vec{x}_{xx}\end{pmatrix} \begin{pmatrix}\vec{a}_{aa}\\ \vec{a}_{ba}\\ \vdots\\ \vec{a}_{xx}\end{pmatrix} \tag{C.16}$$

$$= \begin{pmatrix}\vec{x}_{aa} & \vec{x}_{ba} & \cdots & \vec{x}_{xx}\end{pmatrix} \begin{bmatrix}\mathbf{A}^{-1} & & & \\ & \mathbf{A}^{-1} & & \\ & & \ddots & \\ & & & \mathbf{A}^{-1}\end{bmatrix} \begin{pmatrix}\vec{f}_{aa}\\ \vec{f}_{ba}\\ \cdots\\ \vec{f}_{xx}\end{pmatrix} \tag{C.17}$$

$$= \begin{pmatrix}\vec{x}_{aa} & \vec{x}_{ba} & \cdots & \vec{x}_{xx}\end{pmatrix} \begin{bmatrix}\mathbf{A}^{-1} & & \\ & \ddots & \end{bmatrix} \mathbf{M_F} \begin{pmatrix}\vec{F}\\ \vec{F}_x\\ \vec{F}_y\\ \vec{F}_{xy}\end{pmatrix} \tag{C.18}$$

Above, $\mathbf{M_F}$ is composed of N_{grid} successive matrix *slices* $\mathbf{m_F}$, where each *slice* $\mathbf{m_F}$ has components $\mathbf{m'_F}$ that are the $\mathbf{m'_F}$ of the former slice, except with the final column moved to the beginning.

$$\mathbf{m_F} = \begin{bmatrix}\mathbf{m'_F} & 0 & 0 & 0\\ 0 & \mathbf{m'_F} & 0 & 0\\ 0 & 0 & \mathbf{m'_F} & 0\\ 0 & 0 & 0 & \mathbf{m'_F}\end{bmatrix} \tag{C.19}$$

The $\mathbf{m'_F}$ for the first slice is shown below.

$$\mathbf{m'_F} = \begin{bmatrix}1 & 0 & 0(\times N_{\text{rows}}-2) & 0 & 0 & 0(\times N_{\text{rows}}-2)\\ 0 & 0 & 0(\times N_{\text{rows}}-2) & 1 & 0 & 0(\times N_{\text{rows}}-2)\\ 0 & 1 & 0(\times N_{\text{rows}}-2) & 0 & 0 & 0(\times N_{\text{rows}}-2)\\ 0 & 0 & 0(\times N_{\text{rows}}-2) & 0 & 1 & 0(\times N_{\text{rows}}-2)\end{bmatrix} \tag{C.20}$$

As $\vec{F}$ represents the energies at the grid points of the CMAP correction, it is most sensible to arrange this in the order input to molecular dynamics engines in CHARMM and AMBER. This order samples all values of ψ before incrementing ϕ, and then sampling all values of ψ for the next ϕ, and so on.

Thus, we define $\vec{F}$ as in Equation C.21.

$$\vec{F} = \begin{pmatrix} f(0,0)_{-180°,-180°} \\ f(0,0)_{-180°,-165°} \\ f(0,0)_{-180°,-150°} \\ \vdots \\ f(0,0)_{165°,165°} \end{pmatrix} \tag{C.21}$$

Then, as the derivatives of $\vec{F}$ depend on $\vec{F}$, we must define $\begin{pmatrix} \vec{F}^\intercal & \vec{F_x}^\intercal & \vec{F_y}^\intercal & \vec{F_{xy}}^\intercal \end{pmatrix}^\intercal$ in terms of $\vec{F}$.

$$\begin{pmatrix} \vec{F} \\ \vec{F_x} \\ \vec{F_y} \\ \vec{F_{xy}} \end{pmatrix} = \begin{bmatrix} \mathbf{1} \\ \mathbf{M_x} \\ \mathbf{M_y} \\ \mathbf{M_x M_y} \end{bmatrix} \vec{F} \tag{C.22}$$

Thus we need only define the derivative matrices $\mathbf{M_x}$ and $\mathbf{M_y}$, starting with $\mathbf{M_y}$:

$$\mathbf{M_y} = \begin{bmatrix} \mathbf{M_{141}}^{-1}\mathbf{M_{-303}} & & \\ & \mathbf{M_{141}}^{-1}\mathbf{M_{-303}} & \\ & & \ddots \end{bmatrix} \tag{C.23}$$

where $\mathbf{M_{-303}}$ is defined, analogously to Equation C.15, as:

$$\mathbf{M_{-303}} = \begin{bmatrix} 3 & & & \cdots & -3 \\ -3 & 3 & & \ddots & \\ & -3 & 3 & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \\ & & -3 & & 3 \\ 3 & & & -3 & \end{bmatrix} \tag{C.24}$$

Knowing $\mathbf{M_y}$, we can apply this matrix to evaluate the derivatives with respect to x by simply rearranging the values in $\vec{F}$ by an adapter matrix $\mathbf{C}$. This can be accomplished by setting to 1 the column equal to the modulus of the row with respect to the dimension, times the dimension, plus the floor of the row divided by the dimension, for each row, leaving all other values 0. An

example for a 3×3 matrix in vector form is shown below.

$$\mathbf{C_{3\times 3}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{C.25}$$

Such a matrix could be used to switch the columns and rows when calculating derivatives across y, and then to switch back the columns and rows. Thus, $\mathbf{M_x}$ depends on $\mathbf{M_y}$.

$$\mathbf{M_x} = \mathbf{C}\mathbf{M_y}\mathbf{C} \tag{C.26}$$

Thus, the final equation is:

$$f(\phi, \psi) = \begin{pmatrix} \vec{x}_{aa} & \vec{x}_{ba} & \cdots & \vec{x}_{xx} \end{pmatrix} \begin{bmatrix} \mathbf{A}^{-1} & & \\ & \ddots & \end{bmatrix} \mathbf{M_F} \begin{bmatrix} \mathbf{1} \\ \mathbf{M_x} \\ \mathbf{C}\mathbf{M_x}\mathbf{C} \\ \mathbf{M_x}\mathbf{C}\mathbf{M_x}\mathbf{C} \end{bmatrix} \vec{F} \tag{C.27}$$

As all but $\vec{F}$ can be pre-computed into a single row-vector, this becomes a problem amenable to linear-least squares analytical solution by composing a target vector containing all $f(\phi, \psi)$ and a data matrix containing all the rows as specified in Equation C.27. Then $\vec{F}$ is the vector of unknowns readily obtained from a singular value decomposition or other solver method.