THE UNIVERSITY OF CHICAGO

NEW METHODS USING RIGOROUS MACHINE LEARNING FOR COARSE-GRAINED PROTEIN FOLDING AND DYNAMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
JOHN M. JUMPER

CHICAGO, ILLINOIS
MARCH 2017

To Sarah, Everett, Carolyn, and my grandfather Fred House

“All models are wrong but some are useful” – George Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis would be impossible without my family. My wife Carolyn is unbelievably wise and supportive. She is my best counsel and greatest strength. Sarah and Everett are absolutely wonderful, and they gently remind me of the important of work-life balance. When I spend too much time on research, Sarah closes my laptop and reminds me, “Daddy, you can do that later.” I also want to thank my parents for putting my education first when I was young, giving me a great foundation on which I am still building.

I owe a great intellectual debt to my advisors, Tobin Sosnick and Karl Freed. It is essential for theoretical work to have a close connection to experiment evidence, and Tobin is excellent at interpreting between these two cultures of science. He further provided me with enormous latitude to address protein modelling using an idiosyncratic mix of physics, statistics, and computer science techniques, even when such approaches little resemble to familiar techniques. Karl is a consummate theorist and is an inspiration to a lifetime of tackling hard problems with insightful techniques. I also want to thank my lab mates for vigorous discussion and a willingness to argue forcefully in and out of group meetings.

Finally, I want to thank my friends in physical chemistry for making Hyde Park a great place to spend my time.

CHAPTER 1
INTRODUCTION

Proteins are challenging physical systems to model because of an interplay between statistical mechanics that explores conformational space compatible with a protein's sequence and evolution that explores sequence space compatible with a protein's structure and function. While some aspects of protein structure are amenable to deduction from simple principles, such as Linus Paulings' prediction of helical and sheet structure, much of our knowledge of protein physics comes from a great number of generally weak energetic preferences rather than hard rules, requiring a more flexible approach to understanding structure and dynamics. Since the pioneering work of Karplus on molecular dynamics, scientists have attempted to characterize the wiggling and jiggling of protein atoms by watching them do so on their computer in order to transfer computational observations to experimental hypotheses. This thesis will explore new, rigorous methods to characterize both the motions of proteins and the parameters that control them.

Sampling a conformation from the Boltzmann ensemble of proteins is extremely challenging owing to three separate factors. The first challenge is to represent the complicated potential energy surface of the protein. The cooperative folding behavior of proteins[2] indicates that proteins are often finely balanced between widely-separated states, the native state and the unfolded state. We must represent the energetics of these states carefully in order to reproduce this behavior. The tremendous number of potential conformations requires us to separate the native state from all other conformations by a large free energy gap in order to observe the high fraction of the native state observed in the experimental ensemble.

The second challenge is that a realistic protein potential energy will be subject to the separation of time scales[7] that typifies protein dynamics. It is plausible that one can create a unrealistic potential energy function that folds nearly downhill to the native state for even large proteins with great accuracy; this often is the basis of structure prediction methodologies that make no attempt to represent the Boltzmann ensemble of a protein. For models that seek to represent the totality of protein physics, we likely must accept that folding is often cooperative, with the attendent extended conformational search.

This thesis and the Upside model described herein tackles the potential and sampling challenges simultaneously. By carefully designed to be physical, this procedure should mostly extract physical information from statistical regularities (deviations from randomness) of the training data. In some cases, however, the potential energy will lack the necessary flexibility in its functional form to model the underlying physics. In these cases, unphysical interaction energies of the terms present in the potential may be required to best match experimental data. While such unphysical interaction energies provide the best prediction of training data by definition, they may limit the ability of the resulting models to generalize to related prediction (e.g. while a potential energy may provide the best native-state structure predictions, it may not generalized to correctly predict the properties of extended states).

A similar situation occurs in other potential functions. Three-molecule interactions are very important to the energetics of water clusters and water interfaces, yet are absent in many common water models like TIP3P[3]. To obtain the impressive agreement of water models to bulk water observables[5], three-molecule interactions must be incorporated into two-molecule interactions in an averaged sense. This approximation works for many properties of interest but can give flawed predictions at interfaces. The general solution is to identify the deficiencies and propose modified potentials to reduce the impact of these errors. The solution for protein statistical potentials would be similar.

1.1 Computational trends that favor statistical potentials

Modern computational resources are typically characterized by an abundance of floating point performance divided among many cores. These individual cores run in parallel, and synchronous communication between the cores is typically quite costly. This description fits both the multicore CPUs and GPUs within a computer, as well as the numerous but weakly coupled computers found in a typical supercomputing cluster. The weakly-coupled nature of computational elements makes it very difficult to use the entirety of a cluster to quickly simulate a single physical system, though it can be achieved with enormous engineering effort[6]. Instead, the natural path forward for utilizing modern computational resources is to express protein dynamics problems using collections of loosely coupled simulations.

From a traditional protein dynamics points of view, the Folding@Home project addresses this computational challenge using a large collection of very short simulations to characterize the dynamic process of proteins using volunteer resources[4]. Interpreting the results of these simulations is challenging, and often requires an assumption of Markovianity of the dynamics on a chosen, finite partition of the observed state space of the protein[1]. It is beyond the scope of this thesis to explore the advantages and limitations of such a method.

Approaches similar to Folding@Home are likely unnecessary for coarse-grained models of single protein dynamics, because the coarse-graining both reduces the computational load and increases the simulation decorrelation rate. At the same time, utilizing modern, large-scale computational resources is hampered by the inability to use more than a few cores to accelerate the small computation per step of coarse-grained models. This benefits experimentalists and other less-specialized consumers of computational models, who are often not well-versed in the operation of supercomputing clusters and hence benefit from the development of simulation methodology that runs easily on commonly accessible resources.

Instead of focusing enormous computational resources on simulating a single protein, a natural path forward is to direct the coarse-grained simulations to simultaneously run hundreds or thousands of distinct trajectories. The main approach discussed in this thesis is to use the simulation trajectories of 10^2–10^3 different proteins to identify weaknesses of the simulation model, and then perturb the potential energy function to increase accuracy. The costs from such a procedure, if carried out across a large fraction of the PDB, may be quite large but this significant computation only needs to be performed once. This expensive training creates an inexpensive-to-simulate model of proteins that enables even non-specialists to use easily. The trends in computational resources make this approach increasingly viable, so long as we may equilibrate of single proteins on a small number of computational cores in reasonable time. Our newly-created Upside software and parameterized potential described in this thesis are expressly intended to enable sufficiently-fast PDB-scale training on current computational resources.

The resulting Upside model of protein physics enables a great variety of protein studies, from folding to conformational change to binding to predictions of protein structure using Upside as a strong Bayesian prior on protein structures. The focus of this thesis has been to develop techniques that allow researchers to easily obtain a sample from the Boltzmann distribution for proteins of small-to-moderate size. Most of the dividends of this research program are still to be reaped and a selection of potential applications are discussed in chapter 5.

1.2 Contributions of this thesis

This thesis makes contributions to both the sampling and parameterization problems for coarse-grained protein potentials.

In chapter 2, we solve a major limitation of protein simulations, the inability to have detailed side chain interactions without the difficult equilibration of very rough energy surface. We define and approximate a novel scheme for instantaneously-equilibrated side chain free ensembles. This side chain free energy allows rapid sampling in the smoother potential energy surface of the protein backbone conformations for molecular dynamics while retaining the detailed energetics of the side chain conformations. Importantly, the free energy is inherently many-body, as is typical for interacting systems with pair interactions, capturing the non-additivity of side chain interactions due to the inability to simultaneously adopt multiple rotamer conformations simultaneously. As an associated benefit, direct optimization of the rotamer prediction accuracy of the side chain model yields a side chain rotamer predictor competitive to the state of the art in accuracy while running two orders of magnitude faster.

In chapter 3, we address the parameterization challenge by using the contrastive divergence method from machine learning to parameterize a detailed potential for protein folding. We show that only optimizing near-crystal conformations is sufficient to define a potential capable of *de novo* folding of small proteins to high accuracy. Furthermore, we examine the ability of our model to represent temperature-denatured states of proteins in preparation for future work on folding pathways. It should be noted that chapters 2 and 3 are intended for publication, so that they contain redundancies with other parts of the thesis.

Chapter 4 describes the details of the Upside model. In particular, the subtleties of working with an existing torsional potential for the Ramachandran angles is discussed, as well as the need for careful treatment of side chain-backbone interactions. We also discuss a number of decisions in the design of a coarse-grained potential necessary to ensure effective and rapid sampling.

Integral to the theoretical work described above, we have developed an entirely new simulation package for the simulation and training of coarse-grained physics models, including our semi-implicit side chain model. This simulation packages applies lessons from the machine learning community that it is essential to have a single framework for training and testing statistical models so the same implementation used for protein dynamics can be used to compute the parameter derivatives needed to train the model. This simulation package is also extensible so that arbitrarily sophisticated potential energy functions or virtual coordinates may be created, which is intended to be an extensible platform to integrate sophisticated new sources of information, such as that available from statistical predictions of secondary or tertiary structure. This software is publically available under an open-source license at `https://psd-repo.uchicago.edu/freed-and-sosnick-lab/upside-md`.

Finally, chapter 5 concludes with future work that builds on the methods and perspectives developed in this work. Both future applications to protein dynamics and enhanced parameterization techniques are presented.

1.3 References

[1] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, 2014.

[2] Hüseyin Kaya and Hue Sun Chan. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Structure, Function, and Bioinformatics*, 40(4):637–661, 2000.

[3] R Kumar and JL Skinner. Water simulation model with explicit three-molecule interactions. *The Journal of Physical Chemistry B*, 112(28):8311–8318, 2008.

[4] Stefan M Larson, Christopher D Snow, Michael Shirts, and Vijay S Pande. Folding@home and genome@home: Using distributed computing to tackle previously intractable problems in computational biology. *arXiv preprint arXiv:0901.0866*, 2009.

[5] Pekka Mark and Lennart Nilsson. Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *The Journal of Physical Chemistry A*, 105(43):9954–9960, 2001.

[6] David E Shaw, JP Grossman, Joseph A Bank, Brannon Batson, J Adam Butts, Jack C Chao, Martin M Deneroff, Ron O Dror, Amos Even, Christopher H Fenton, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53. IEEE Press, 2014.

[7] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

CHAPTER 2

MAXIMUM LIKELIHOOD SIDE CHAIN PACKING

Work in this chapter has appeared in [10].

2.1 Introduction

Two major challenges must be overcome in order to confront the difficulties of accurately simulating protein dynamics. The first is the necessity of balancing the large and competing sources of energy and entropy whose total determines both the thermodynamics and the native conformation of the protein. The second challenge involves the intensive sampling required to obtain a Boltzmann ensemble of conformations. The sampling challenge is addressed here by integrating out the side chain free energy to produce a coarse-grained configuration defined just in terms of the backbone N, C_α, and C atoms. Consequently, backbone motions evolve on a smoother coarse-grained free energy surface with greatly reduced side chain rattling (molecular friction) compared to that for standard all-atom molecular dynamics simulations (see article on long side chain timescales).

The uncertainty in the position of coarse-grain interactions introduces the difficulty of accurately parameterizing a coarse-grained model to represent the physical interactions. Moreover, all-atom force fields produce conformations that deviate from experiment, especially for unfolded proteins[20]. Hence, rather than following the customary process of matching the energies of the coarse-grained model to approximate the already inexact energies of atomistic force fields or trying to interpret raw statistics for the distribution of interatomic distances in the Protein Data Bank (PDB)[1] and defining the correct reference state appropriate to the statistical potential[7], our side chain energies are determined as those that best reproduce the side chain conformations observed in the PDB given fixed backbone configurations. This maximum likelihood approach has key advantages: (1) it directly provides an interpretation of the structural information as a sample from the statistical mechanical ensemble of side chain packing, and (2) it can be evaluated quickly since we will show that approximating the Boltzmann distribution for the side chains in a fixed backbone configuration does not require laborious discrete sampling of the χ angles. This work also presents a computationally extremely inexpensive, coarse-grained approximation for describing side chain packing, thereby allowing the model to be directly used as the free energy function for molecular dynamics simulations. Our method enables rapidly equilibrating coarse-grained simulation that can nonetheless contain significant molecular detail.

2.2 Constructing and evaluating the energy of the side chains

The positions of the N, C_α, and C atoms constitute the backbone trace. The native backbone trace determines the fold of the protein, and the free energy of the trace is likely to preserve the major barriers that determine the slow degrees of freedom in the protein. The strategy in our method, called *Upside*, is to perform dynamics simulations of the backbone trace, while still including sufficient structural details (side chain structures and free energies, etc.) necessary to compute realistic forces on the three atoms of the backbone trace. This strategy yields the advantage that the inclusion of the side chain free energy, rather than the side chains themselves, greatly smooths the potential governing the dynamics of the backbone trace, especially because of the reduction of steric rattling attributable to the side chains as they try to sample multiple substates in the condensed state.

First consider a representation of the protein configurations in terms of the coordinates $(\{b_i\}, \{\chi_i\})$ where b_i represents the positions of the backbone N, C_α, and C atoms on the i-residue and χ_i represents the side chain χ-angles on the i-th residue. Since bond lengths and angles are approximately constant for proteins, the positions of the protein atoms can be reconstructed with high accuracy from the $(\{b_i\}, \{\chi_i\})$ coordinates. Given a potential

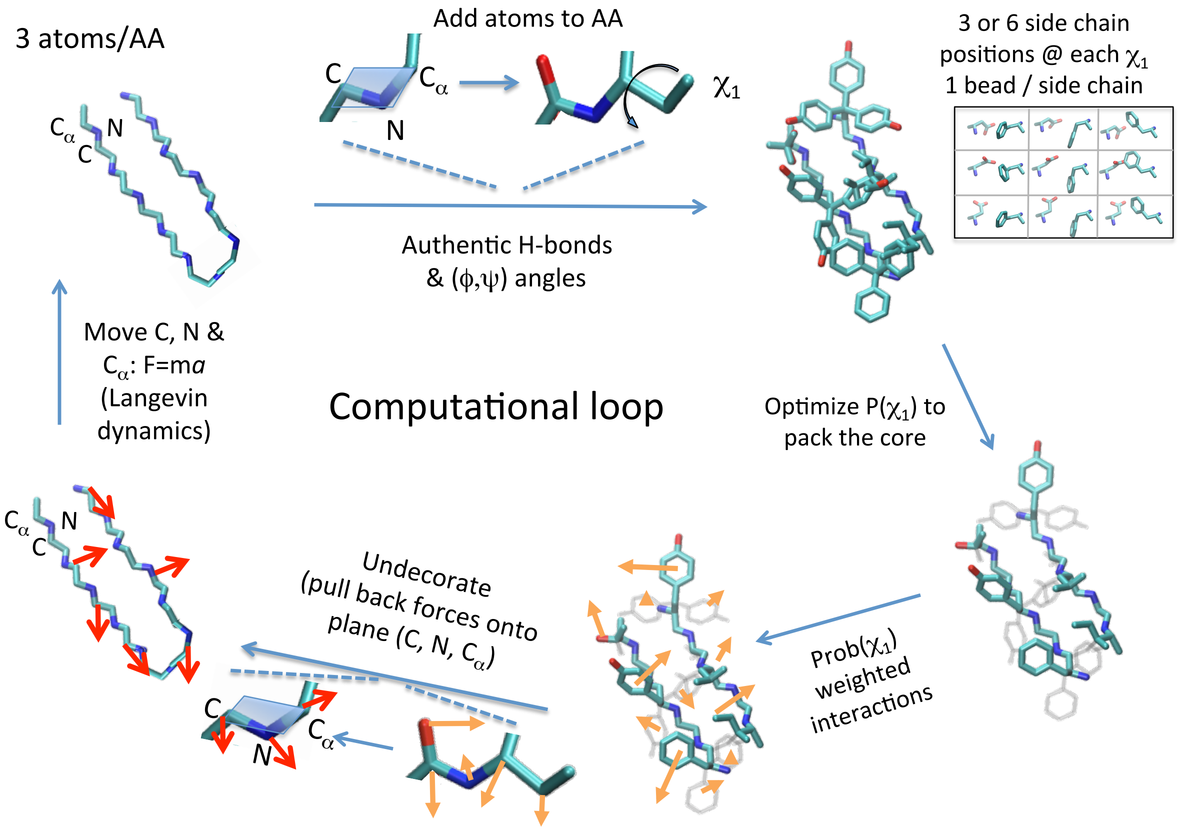

3 atoms/AA
C_α
N
C
Add atoms to AA
C
C_α
N
χ_1
3 or 6 side chain positions @ each χ_1
1 bead / side chain
Authentic H-bonds & (ϕ,ψ) angles
Move C, N & C_α: F=m*a* (Langevin dynamics)
Computational loop
Optimize $P(\chi_1)$ to pack the core
C_α
N
C
Undecorate (pull back forces onto plane (C, N, C_α)
C
C_α
N
Prob(χ_1) weighted interactions

Figure 2.1: Inner loop of Upside calculation. Executing this computational loop in the present simulations takes between 0.25 CPU milliseconds (BBA) and 0.82 CPU milliseconds (ubiquitin) per iteration (each simulation replica is run on single 2.60 GHz processor core). The side chain potential enters into the integration step simply as a complicated, many-body energy function that may be treated with standard techniques of molecular simulations.

energy $V(\{b_i\}, \{\chi_i\})$, we define the free energy as a function of the backbone configuration,

$$\bar{V}(\{b_i\}) = -\log \int d\chi_1 \cdots \chi_N \, e^{-V(\{b_i\},\{\chi_i\})}. \tag{2.1}$$

Natural energy units are used so that $k_{\mathrm{B}}T = 1$. An intermediate step of this derivation requires the introduction of a discrete approximation $\{\tilde{\chi}_i\}$ for our χ-angles and a discrete approximation $\bar{V}(\{b_i\}, \{\tilde{\chi}_i\})$ for the potential.

Rather than directly calculate (2.1), we define an intermediate discrete approximation to $\bar{V}$ that is amenable to approximation techniques. Consider a discrete coarse-graining function g so that $\tilde{\chi}_i = g(\chi_i)$, where $\tilde{\chi}_i$ is a small integer ($\tilde{\chi}_i \in \{1, \ldots, 6\}$ in this work). The coarse-grain potential $\tilde{V}$ is defined so that

$$e^{-\tilde{V}(\{b_i\},\{\tilde{\chi}_i\})} \approx \int d\chi_1 \cdots \chi_N \left(\prod_i \delta_{\tilde{\chi}_i f(\chi_i)} \right) e^{-V(\{b_i\},\{\chi_i\})}. \tag{2.2}$$

In principle, any coarse-graining function for the side chains may be used; however the discrete approximation $\tilde{V}$ to the potential provides a more accurate approximation whenever the distribution of χ-angles is sharply peaked (in the true potential V) within each discrete state $\tilde{\chi}$. See Figure 2.3 for an example of a coarse-graining function and see section 2.4 where an optimized coarse-graining function f is derived. We make the following assumptions on the form of $\tilde{V}$. First, we assume there is an explicit function $y_i(b_i, \tilde{\chi}_i)$ for the side chain coordinates based only on the backbone coordinates and side chain state for residue i. We may relax the requirements to depend on only a single residue's backbone position, but the requirement that y_i depend on only a single side chain state $\tilde{\chi}_i$ is firm. These directed coordinates are approximately side chain centers of mass with direction given by the C$_\beta$–C$_\gamma$ bond vector direction. However, a parameterization of these side chain position functions separately for each amino acid type enables the maximization of the accuracy of the approximation in equation (2.2). A further assumption is that $\tilde{V}$ can be expressed in the form

$$\tilde{V}(\{b_i\},\{\tilde{\chi}_i\}) = V^{\text{backbone}}(\{b_k\})+ \tag{2.3}$$

$$\sum_i V_i^{(1)}(\{b_k\},\tilde{\chi}_i, y_i(b_i,\tilde{\chi}_i))+ \tag{2.4}$$

$$\sum_{i,j} V_{ij}^{(2)}(y_i(b_i,\tilde{\chi}_i), y_j(b_j,\tilde{\chi}_j))), \tag{2.5}$$

where the pair interaction $V_{ij}^{(2)}(y_i, y_j) = 0$ for the side chain is taken to vanish beyond a cutoff R_{cutoff}. Notice that the dependence of the potential on the backbone is completely general except the potential is assumed to contain at most a pairwise dependence on the discrete rotamer states $\tilde{\chi}_i$. Explicit parameterizations for y_i and $\tilde{V}$ are defined in section 2.5 using the principle of maximum likelihood.

One can simulate the Boltzmann ensemble for $\tilde{V}$ using molecular dynamics for the backbone $\{b_i\}$ and Monte Carlo moves for the side chain states $\{\tilde{\chi}_i\}$, but the strong steric interactions lead to a slow equilibration and dynamics for both the side chains and backbone. Since we are predominantly interested in backbone motions, we return to the free energy $\bar{V}$ in (2.1), now summing over discrete side chain states instead of integrating over continuous side chain angles,

$$e^{-\bar{V}(\{b_i\})} \approx \sum_{\tilde{\chi}_1,\ldots,\tilde{\chi}_N} e^{-\tilde{V}(\{b_i\},\{\tilde{\chi}_i\})}. \tag{2.6}$$

The potential $\bar{V}$ represents a further coarse-graining of the system by completely replacing the influence of the side group with a potential describing the adiabatic free energy of the side chains for a given fixed backbone conformation. Because $\bar{V}$ depends only on the (continuous) backbone coordinates, this choice of $\bar{V}$ enables running standard molecular dynamics simulations instead of a hybrid of Monte Carlo and molecular dynamics. The potential $\bar{V}$ is a much smoother function of the backbone coordinates than the original $V(\{b_i\},\{\chi_i\})$ because the replacement of the side chain degrees of freedom with the approximate free energy of the side chains greatly reduces steric rattling and molecular friction. The reduction of the ruggedness of the energy landscape enhances diffusion within conformational basins but preserves the overall structure and barriers of the conformational ensemble.

2.3 Approximating the discrete free energies of the side chains

The benefits of running dynamics with the coarse grained $\bar{V}$ enter at great cost because using even three coarse-grained states per side chain implies a summation over 3^N $\tilde{\chi}$-states in equation (2.6). Furthermore, the vast majority of those 3^N states have steric clashes or other large energies and, therefore, contribute little to the free energy of the side groups.

To approximate the free energy of the side chains $\bar{V}$, it is convenient to express our problem in the language of Ising models so that we can apply standard techniques developed in that context. For a fixed backbone configuration $\{b_i\}$,

$$\begin{aligned}\tilde{V}(\{b_i\},\{\tilde{\chi}_i\}) &= \bar{v}(\{\tilde{\chi}_i\}) \\ &= \sum_i v_i^{(1)}(\tilde{\chi}_i) + \sum_{\substack{i,j \\ \text{neighbors}}} v_{ij}^{(2)}(\tilde{\chi}_i,\tilde{\chi}_j), \end{aligned} \tag{2.7}$$

where the potentials $\bar{v}$ are written in lowercase to indicate suppression of the dependence on the fixed backbone coordinates $\{b_i\}$ in order to focus on the side chain contribution. Notice that with the backbone positions fixed, each single-residue potential $v_i^{(1)}$ is simply a vector with as many components as the number of possible states for $\tilde{\chi}_i$ (e.g. length-6 vectors). Similarly, each of the pair potentials $v_{ij}^{(2)}$ is a small 6x6 matrix of potential energies to cover the 36 possibilities. These single and pair potentials are calculated only once before evaluating the free energy as described in section 2.5. Moreover, the pair summation in equation (2.7) only applies for residues pairs i and j that are neighbors spatially. A pair of residues (i,j) are neighbors if inter-residue distance $|y_i(\tilde{\chi}_i) - y_j(\tilde{\chi}_j)|$ is less than a cutoff R_{cutoff} for any of their possible discrete states $(\tilde{\chi}_i,\tilde{\chi}_j)$. In this work, we use $R_{\text{cutoff}} = 7\text{Å}$

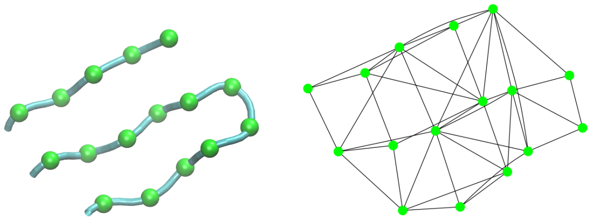


Figure 2.2: Fragment of protein G with associated interaction graph ($R_{\text{cutoff}} = 7\text{Å}$). A pair of residues has a connection whenever their side chain beads are within R_{cutoff} for any side chain states.

for side chain-side chain interactions and $R_{\text{cutoff}} = 5\text{Å}$ for side chain-backbone interactions.

The potential $\tilde{V}$ may be visualized as an energy function on a graph with one discrete site per amino acid. The graph has a connection between any two residues that are within the cutoff separation R_{cutoff} as defined above. This graph is illustrated in Figure 2.2 for a model protein configuration. The structure of this graph varies dynamically over the course of a simulation because the definition of neighboring residues depends on the backbone configuration $\{b_i\}$. The potential varies smoothly as the backbone moves so long as the pairwise potential functions are continuous in the backbone coordinates. The potential $\tilde{V}$ is continuous despite the changing connections of the graph, because the strength of the potential for each interaction approaches zero at R_{cutoff} just before the connection is eliminated from the graph. Problems such as this, with discrete potentials on an arbitrary graph, are extensively studied in both statistical mechanics (as variants of the Ising model) and machine learning (as undirected graphical models or Markov random fields)[24]. Below we adopt some well studied approximations from these fields to provide accurate and tractable methods for computing our coarse-grain potential $\bar{V}$.

Two approximations (see [24]) are invoked to compute the free energy from

$$\bar{V} = G^{\text{SC}} = -\log \sum_{\tilde{\chi}_1,\cdots,\tilde{\chi}_N} e^{-v(\{\tilde{\chi}_i\})}. \tag{2.8}$$

The first approximation is to express the free energy G^{SC} in terms of the entropy and average energy of the Boltzmann ensemble where the entropy has been replaced by an approximation,

$$\begin{aligned} G^{\mathrm{SC}} &= \langle \bar{v} \rangle - S \\ &\approx \langle \bar{v} \rangle - S^{\mathrm{approx}}, \end{aligned} \tag{2.9}$$

where $\langle \bar{v} \rangle$ and S^{approx} are defined below. Both the average energy and a mutual information approximation to the entropy may be expressed using the single-residue probabilities $p_i(\tilde{\chi}_i)$ that residue i is in state $\tilde{\chi}_i$ in the Boltzmann ensemble of $\bar{v}$ and similarly for the joint probabilities $p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)$. Using p_i and p_{ij}, the approximate energy and entropy are

$$\begin{aligned} \langle \bar{v} \rangle =& \sum_i \sum_{\tilde{\chi}_i} p_i(\tilde{\chi}_i) v_i^{(1)}(\tilde{\chi}_i) + \\ & \sum_{\substack{i,j \\ \text{neighbors}}} \sum_{\tilde{\chi}_i, \tilde{\chi}_j} p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) v_{ij}^{(2)}(\tilde{\chi}_i, \tilde{\chi}_j) \end{aligned} \tag{2.10}$$

$$\begin{aligned} S^{\mathrm{approx}} =& -\sum_i \sum_{\tilde{\chi}_i} p_i(\tilde{\chi}_i) \log p_i(\tilde{\chi}_i) + \\ & - \sum_{\substack{i,j \\ \text{neighbors}}} \sum_{\tilde{\chi}_i, \tilde{\chi}_j} p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) \log \frac{p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)}{p_i(\tilde{\chi}_i), p_j(\tilde{\chi}_j)}. \end{aligned} \tag{2.11}$$

The mutual information approximation to the entropy ignores contributions from three-residue and higher correlations. We intend to minimize the approximate free energy (2.9) over all putative Boltzmann probability distributions for the side chain states $\{\tilde{\chi}_i\}$. Notice that only the 1-side chain probabilities p_i and 2-side chain probabilities p_{ij} are required to compute the average energy and approximate entropy; we do not need the more complicated full joint probability distribution of the $\{\tilde{\chi}_i\}$ states for all side chains. In addition to the mutual information approximation of the entropy, we assume that any pair probability p_{ij} represents possible pair probabilities from a Boltzmann distribution, so that the only task is to minimize the free energy with respect to the pair probabilities. The only constraints imposed are that they must satisfy the obvious consistency conditions for probabilities,

$$p_i(\tilde{\chi}_i) = \sum_j p_{ij}(\tilde{\chi}_j) \tag{2.12}$$

$$\sum_{\tilde{\chi}_i, \tilde{\chi}_j} p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) = 1 \tag{2.13}$$

$$\sum_{\tilde{\chi}_i} p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) = \sum_{\tilde{\chi}_k} p_{jk}(\tilde{\chi}_j, \tilde{\chi}_k) \tag{2.14}$$

$$p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) = p_{ji}(\tilde{\chi}_j, \tilde{\chi}_i). \tag{2.15}$$

However, use of only the conditions (2.12)-(2.15) is insufficient to ensure that a joint probability distribution exists for all the variables consistent the with the choices of p_i and p_{ij}. As an explicit example,

$$p_{12} = p_{23} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \tag{2.16}$$

$$p_{13} = \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix} \tag{2.17}$$

obeys conditions (2.12)-(2.15) but is not representable by any probability distribution on the three residues. This is clear because residue 1 is completely correlated to residue 2, and residue 2 is completely correlated to residue 3, but residues 1 and 3 are independent, which is impossible.

Accepting the two approximations for entropy and representability, the free energy becomes

$$G^{\mathrm{SC}} \approx \min_{\{p_i\},\{p_{ij}\}} (\langle \bar{v} \rangle - S^{\mathrm{approx}}). \tag{2.18}$$

Thus, we now have a tractable approximation to free energy of the side chain. We can minimize that free energy using a self-consistent iteration technique called belief propagation; see appendix 2.11 for details. The iteration typically converges rapidly, often in 10-20 steps.

Molecular dynamics simulations require calculations of the forces on the backbone coordinates, $-\frac{d\bar{V}}{db_i}$. The derivatives can be computed simply using the chain rule, noting that several terms will be zero because the pair probabilities minimize the free energy,

$$\begin{aligned}\frac{dG^{\mathrm{SC}}}{db_k} &= \frac{\partial G^{\mathrm{SC}}}{\partial b_k} + \sum_i \frac{\partial G^{\mathrm{SC}}}{\partial p_i}\frac{\partial p_i}{\partial b_k} + \sum_{\substack{i,j \\ \text{neighbors}}} \frac{\partial G^{\mathrm{SC}}}{\partial p_{ij}}\frac{\partial p_{ij}}{\partial b_k} \\ &= \frac{\partial G^{\mathrm{SC}}}{\partial b_k} = \frac{\partial \langle \bar{v} \rangle}{\partial b_k} = \left\langle \frac{\partial \bar{v}}{\partial b_k} \right\rangle \\ &= \sum_i \sum_{\tilde{\chi}_i} p_i(\tilde{\chi}_i) \frac{\partial v_i^{(1)}}{\partial b_k}(\tilde{\chi}_i) + \\ &\quad \sum_{\substack{i,j \\ \text{neighbors}}} \sum_{\tilde{\chi}_i,\tilde{\chi}_j} p_{ij}(\tilde{\chi}_i,\tilde{\chi}_j) \frac{\partial v_{ij}^{(2)}}{\partial b_k}(\tilde{\chi}_i,\tilde{\chi}_j) \end{aligned} \tag{2.19}$$

where $\frac{\partial G^{\mathrm{SC}}}{\partial p_i} = \frac{\partial G^{\mathrm{SC}}}{\partial p_{ij}} = 0$ because p_i and p_{ij} are chosen to minimize G^{SC}. The remaining simplifications occur because S^{approx} is independent of the backbone coordinates. While the underlying side chain interactions are pairwise additive and vanish outside the cutoff radius R_{cutoff}, the free energy (2.9) is a many-body potential that can interact over arbitrary distances.

Since the approximate free energy due to the side chains is not a convex function of the probabilities, local minima may arise and impair the the self-consistent iteration from finding the global minimum. To reduce the danger posed by the presence of local minima, calculations are begun from a carefully initialized state, as detailed in appendix 2.11. Other self-consistent approximations exist for the side group free energy, such as tree-reweighted belief propagation[23], that are typically less accurate but always converge to the global minimum of their approximate free energy. Another limitation of the present approximation

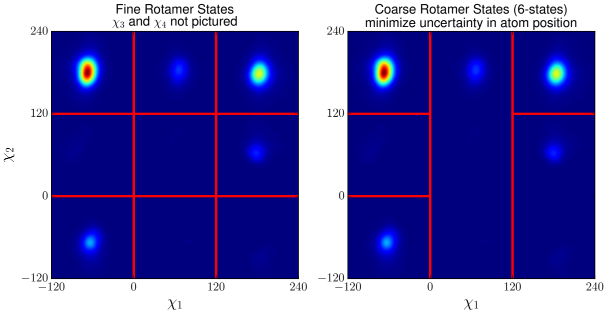


Figure 2.3: Example of optimized coarse states for arginine overlaid on the PDB distribution of the rotamer angles χ_1 and χ_2. Each of the six coarse states contains only a single fine state that has high probability, so that the variance of dihedral angles within each coarse state is small.

scheme arises when a bi-stable or multi-stable energy landscape is possible for the rotamer states. If well-separated and equally important minima are present for a single backbone configuration in the rotamer free energy surface, the probabilities only converge to a single minumum and thus underestimate the entropy of the side chains. While this does not appear to occur near the native well, we have not extensively searched for special backbone configurations that would result in bi-stable rotamer energies. The characterization of such problematic configurations, likely near free energy barriers, is left to future work.

2.4 Optimized mapping to coarse states

The χ-angles for the side chains are partitioned into discrete states in an optimized manner. The NDRD rotamer library[18] provides a set of approximate discrete states for each residue type according to their frequencies of occurrence in a non-redundant set of high resolution protein structures in the PDB. However, the number of rotamer states in the NDRD library can be quite large. For instance, naively using all 81 rotamers for each arginine, means that computing the pair interaction $v_{i,j}$ for two arginines would require computing $81^2 = 6561$ energy values. Consequently, instead of using all possible rotamer states, several NDRD rotamer states are combined into 3–6 coarse-grained rotamer states for the sake of manageable computational cost.

We choose aggregate the rotamer states of the side chain to minimize the positional uncertainty of side chain atoms in each state. A search over all possible aggregations is conducted to find the aggregation that provides the smallest possible error. More formally, the NDRD rotamer library[18] is used to define the atomic positions $x_{ij}^{f}(\phi, \psi)$, where i is the atom (such as C_β), j is the coordinate (x, y, or z), and f is the fine-grained rotamer state. Each rotamer state has a probability $p^f(\phi, \psi)$ specified in the NDRD library from frequencies in the PDB for each fine-grained rotamer state as a function of the backbone dihedral angles (ϕ, ψ). Each fine-grained state f may belong to exactly one coarse-grained state c (i.e. the c states form a partition of the f states). Given the choice of a coarse-grained state c, an average is performed over the fine-grained atomic positions, and sum is taken over the probabilities of all fine-grained states f grouped into c according to the prescription,

$$q^c(\phi, \psi) = \sum_{s \in t} p^f(\phi, \psi) \tag{2.20}$$

$$y_{ij}^{c}(\phi, \psi) = \frac{1}{q^c(\phi, \psi)} \sum_{s \in t} p_{ij}^{f}(\phi, \psi)\, x_{ij}^{f}(\phi, \psi), \tag{2.21}$$

where q^c is the coarse-grained probability and y_{ij}^{c} is the coarse-grained atomic position.

The error incurred by coarse-graining is defined as the variance of the atom positions within each coarse-grained state, weighted by the frequency of occurrence of the coarse-grained state in the PDB. Specifically, the error $\sigma^2(\phi, \psi)$ is defined as,

$$\sigma^2(\phi, \psi) = \sum_{s} \frac{p^f(\phi, \psi)}{N_{\text{atom}}} \sum_{ij} (x_{ij}^{f}(\phi, \psi) - y_{ij}^{c(f)}(\phi, \psi))^2, \tag{2.22}$$

where N_{atom} is the number of atoms in the side chain and $c(f)$ is the coarse-grained state c that contains the fine-grained state f. The error depends implicitly on the state decomposition $c(f)$ and measures the deviation of the atoms within each state. This error favors the fine-grained states f that have higher frequency of occurrences in the PDB.

The division of fine-grained states into coarse-grained states is restricted for simplicity to be independent of the Ramachandran angles for the residue,

$$\sigma^2 = \int p^{\mathrm{Rama}}(\phi,\psi)\,\sigma^2(\phi,\psi)\,d\phi\,d\psi, \tag{2.23}$$

where $p^{\mathrm{Rama}}(\phi,\psi)$ is the frequency of each Ramachandran angle taken from the PDB coil library. Note that this error term depends implicitly on the decomposition $c(f)$ and weights for the (ϕ,ψ) pairs according to the frequency in the coil library.

An optimal coarse-grained representation of the side chain rotamer states is obtained by minimizing σ^2 for each residue type over all partitions $c(f)$. We force the coarse-graining $c(f)$ to obey a few conditions, essentially to make sure that $c(f)$ is easily interpretable in terms of χ_1 and χ_2 as well as limiting the number of possibilities that must be checked by the brute-force minimization. In particular, the mapping from coarse-states back to χ_1 rotamer states is unambiguous because no single coarse state contains two different χ_1 rotamer states. We impose the following conditions,

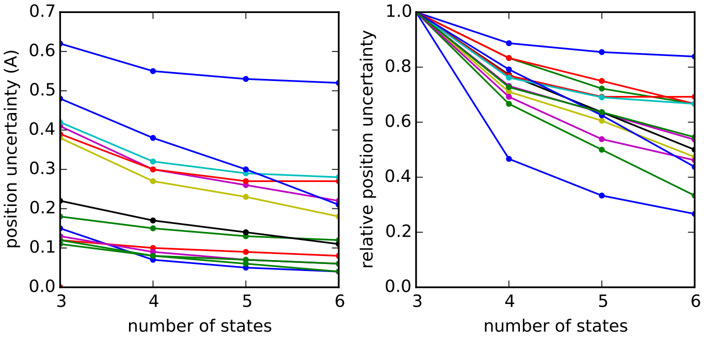
1. $c(f)$ depends only on the χ_1 and χ_2 rotamer states of f (i.e. if f_1 and f_2 states differ only in their χ_3 or χ_4 states, then $c(f_1) = c(f_2)$)

2. Each coarse state c must contain only a single χ_1 state (i.e. if f_1 and f_2 have different χ_1 states, then $c(f_1) \neq c(f_2)$)

3. Each coarse state c must contain a contiguous range of χ_2 values. This greatly reduces the number of possible coarse-grainings for residues with non-rotameric χ_2 angles like asparagine.

Optimizing the decomposition of the coarse-grained state $c(f)$ proceeds by completely enumerating all possible decompositions into coarse-grained states that contain no more than six fine-grained states and by imposing the three conditions.

We optimize the decomposition of the coarse-grained state $c(f)$ by completely enumerating all possible decompositions into coarse-grained states that satisfy the three conditions



Restype	States	Restype	States
ALA	1	LEU	3
ARG	6	LYS	3
ASN	6	MET	6
ASP	6	PHE	6
CYS	3	PRO	3
GLN	6	SER	3
GLU	6	THR	3
GLY	1	TRP	6
HIS	6	TYR	6
ILE	3	VAL	3

Figure 2.4: Error in the decomposition of rotamer states into coarse-grained states as a function of the number of side chain states. The position uncertainty is σ. The relative uncertainty is the position uncertainty for each number of states divided by the accuracy at 3 states. For residues without a rotable χ_2, such as valine, it is not possible to coarse-grain. One, three or six rotamer states are used, depending on the residue type. The computational time to compute the pairwise interactions and solve for the free energy scales roughly as the number of coarse rotamer states squared, so there is an incentive to use as few coarse states as possible. The table (above) summarizes the number of states chosen for each amino acid type.

above and contain no more than six coarse states.

2.5 Parametric bead locations and interactions

Paralleling the necessity of coarse-graining the rotamer states, side chain atoms also require coarse-graining in order to obtain an inexpensive side chain model. This reduction in the number of degrees of freedom may further be justified since the atomic positions of the side chains are uncertain due to the discretization and aggregation of the rotamer states, meaning that there is little value in assigning precise positions for all atoms. We instead use a single oriented bead (location and direction coordinates) to represent each side chain (note that the direction is independent of the side chain, e.g. in aromatic residues it may be the ring normal unit vector). The locations and directions of the side chain beads are changed by the optimizer during the optimization of the potential. The improvement in prediction accuracy from using optimized side chain positions rather than the static positions is substantial, and the accuracy of predicting crystallographic rotamer states rose from 48% to 60% as a result of using optimized positions.

While the side chain belief propagation can handle essentially any pairwise residue potential, considerable care is required to design an interaction form that complex enough to represent a wide range of protein physics and adequately represent side chain packing, yet simple enough to be trained by gradient descent. To do so, we gather key requirements to develop a simple protein model.

The basic component is a radial pair interaction between side chain beads. This term can reproduce an effective side chain excluded volume but it is less obvious that such a model is appropriate to describe attractive interactions. The attractive protein interactions are often directional, arising from the dipole moments of polar residues and the ring stacking of aromatic residues. For this reason, we add a separate directional interaction that depends on the cosine of the angle between the bead direction and the inter-bead separation vector. The bead direction is chosen by the optimizer, and the starting point of the optimization is for the bead direction to be the C_β–C_γ bond vector. This interaction form is sufficient to capture simple angular-dependent interactions of side chains, providing a large boost to the ability of the model to handle side chain hydrogen bonding and other dipolar interactions. Concretely, each interaction pair is described by positions y_1 and y_2 and directions n_1 and n_2. From this the distance $r_{12} = |y_1 - y_2|$ and displacement unit vector $n_{12} = (y_1 - y_2)/r_{12}$ are calculated. The form of the interaction is given by

$$\begin{aligned} V = \kappa(&\,\mathrm{unif}(r_{12}) + \\ &\mathrm{ang}_1(-n_1 \cdot n_{12})\,\mathrm{ang}_2(n_2 \cdot n_{12})\,\mathrm{dir}(r_{12})), \end{aligned} \tag{2.24}$$

where unif, ang_1, ang_2, and dir are smooth curves. The smooth curves are represented by cubic splines. The prefactor κ is 1 for interactions between two side chain beads.

For side chain-backbone interactions, however, it is advantageous to have the prefactor κ depend on the hydrogen bonding state of the backbone residue because the presence of

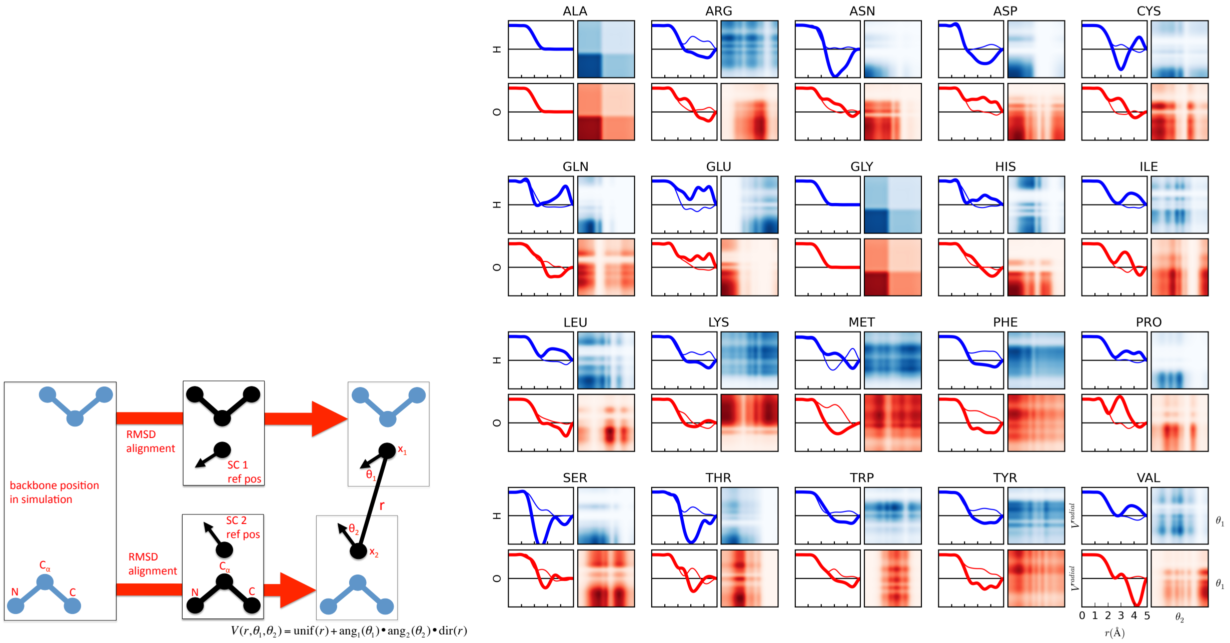


Figure 2.5: Left panel depicts side chain interaction coordinates for equation (2.24). Right panel depicts side chain interactions with backbone hydrogen (blue) and oxygen (red) sites; axes labels indicated on lower right plot. Within each group of 4 plots, the curves on the left hand side present the radial components of the interaction. The thin lines provide the unif(r) interaction and the thick line is the directional interaction unif$(r)+$dir(r). The right plots are heat maps of the angular interactions $\text{ang}_1(\theta_{\text{H/O}})\,\text{ang}_2(\theta_{\text{SC}})$, where the lower left corner depicts the interaction where $\theta_1 = \theta_2 = 0°$. Within each group of 4 plots, the upper plots represent side chain interactions with hydrogen and the bottom plots represent side chain interactions with oxygen.

one hydrogen bond inhibits forming another. Specifically, the interaction between a backbone hydrogen or oxygen is given a hydrogen bond confidence score f, a number that is typically close to 0 for non-Hbonded and 1 for Hbonded residues. We set $\kappa = 1 - f$ so that the interaction is only turned on for hydrogens or oxygens that are not participating in a backbone-backbone hydrogen bond. The physical motivation is that the directional interaction primarily describes the effects of the dipole interactions, and in a hydrogen bond the C=O and N–H dipoles approximately cancel each other. While it is theoretically possible for the algorithm to learn carefully balanced hydrogen and oxygen interactions that themselves cancel out on hydrogen bonded pairs, it is much easier to achieve a physically-reasonable model if we enforce the zeroing of directional interactions with already hydrogen-bonded pairs.

The side chain-backbone interactions are needed to describe helix capping. We have observed that a proper description of these capping effects is required to avoid helix fraying. Furthermore, Harper and Rose[8] have observed that N-terminal capping of a helix by side chains is more likely to be observed than is C-terminal capping of the side chain. This finding is consistent with our maximum likelihood training (below), where side chain-amide hydrogen interactions are fit with stronger (i.e. higher confidence) potentials than side chain-oxygen interactions. Harper and Rose also note that hydrophobic residues play a strong role in helix capping by covering exposed protein backbone at the ends of helices. To provide our model with the freedom to describe this effect, an additional side chain-backbone interaction is added with three beads representing the hydrophobic portion of the backbone. The location of the three beads are initialized from the reference position of N, C_α, and C and are optimized with the rest of the parameters. For this interaction, $\kappa = 1$.

2.6 Maximum likelihood training

2.6.1 Training objective function

The side chain model is trained by the maximum likelihood principle. Specifically, we determine the set of parameters that maximizes the log probability of the true side chain states $\tilde{\chi}_p$ in the Boltzmann ensemble of all possible side chain states $\tilde{\chi}$ for the fixed backbone positions X_p for each protein p.

$$\begin{aligned} p(\tilde{\chi}_p) &= \frac{e^{-V(\tilde{\chi}_p)}}{\sum_{\tilde{\chi}} e^{-V(\tilde{\chi})}} && (2.25) \\ -\log p(\tilde{\chi}_p) &= V(\tilde{\chi}_p) + \log\left(\sum_{\tilde{\chi}} e^{-V(\tilde{\chi})}\right) && (2.26) \\ &= V(\tilde{\chi}_p) - G^{\mathrm{SC}} && (2.27) \\ &= E_{\mathrm{gap}}. && (2.28) \end{aligned}$$

The evaluation of E_{gap} requires the evaluation of the free energy of the side chains, a quantity that is intractable to calculate exactly. Fortunately, our side chain energy (2.18) approximates the true side chain free energy G^{SC} that appears in (2.27). Furthermore, the expression for the parametric derivative (2.19) allows for gradient descent optimization to minimize the average gap energy.

The side chain packing interaction is trained using a large, non-redundant collection of crystal structures from the PDB with 50–500 residues and resolution less than 2.2Å. From a training set of protein structures, we extract the sequences s_p, backbone trace positions X_p, and true coarse-grained side chain states the $\tilde{\chi}_p$ for each protein p. The proteins are further filtered using PISCES[25] so that all pairs of proteins have sequence similarity less than 30%. Non-globular structures in the dataset are removed, as we suspect that the side chain packing of these structures are more strongly influenced by other chains in the crystal structures. We define non-globular structures as outliers in the linear relationship between $\log(N_{\text{res}})$ and $\log(R_g)$; the outliers are identified using the RANSAC algorithm[6]. After filtering, 6255 chains remained, containing approximately 1.4 million residues.

2.6.2 Regularization

Since there are 210 types of amino acid pairs and the potential for each pair has 10s of associated parameters, the model contains more than 10000 parameters just for side chain pair interactions. When maximizing the likelihood for a system with such a large numbers of parameters, it is often beneficial to add a penalty term, called a regularizer or a maximum a posteriori prior, that favors simpler models, greatly reduces overfitting, as well as encourages models that will better generalize to molecular dynamics simulations. Two types of regularization penalties are contrasted. The first penalty simply encourages lower energies except at the repulsive core of the interaction.

The second penalty is a lower bound penalty for the energy that very strongly discourages energies below a certain threshold, where the lower bound is chosen to be as strict as possible without significantly reducing accuracy. The term is needed empirically as the model sometimes learns very low energies for certain side chain-backbone hydrogen bonding for residues such as aspartic acid, on the order of -15 kT. We suspect that these large energies reflect systematic differences between crystallographic structures and the Boltzmann ensemble of proteins in physiological conditions. If an aspartic acid can hydrogen bond to the backbone, the electron density of rotamers with unbound states are likely to be quite small and the crystallographer refining the structure will likely not register the minor alternative rotamer (additionally, we do not use multiple states for a single residue). This effect is likely exacerbated by the low temperature of crystallization. As a result, an unrealistically favorable energy is learned for side chain-backbone hydrogen bonding. This over-estimation is allowable if the only objective is accurate predictions of side chain rotamers, but such large energies (greater than the backbone-backbone hydrogen bond energy by a large amount) are problematic when running dynamics. Enforcing a lower bound on the energies suppresses these large energies.

$$\begin{aligned} E_{\text{reg}} =& k_r \int f\left(V(r,\theta_1,\theta_2) - \frac{5kT}{1. + e^{r/(0.2\text{Å})-10}}\right) dr\ \cos(\theta_1)d\theta_1\ \cos(\theta_2)d\theta_2 + \\ & 2\int (V(r,\theta_1,\theta_2) - V_{\text{lb}})_-^2\ dr\ \cos(\theta_1)d\theta_1\ \cos(\theta_2)d\theta_2 \end{aligned} \tag{2.29}$$

$(x)_-$ is zero whenever x is positive and x otherwise. The function f is the logarithm of the student-t distribution, which is convenient for regularizing while still allowing large energies when necessary. The hyperparameters k_r and V_{lb} are chosen below based on the results of the maximum likelihood training.

2.6.3 Optimization and validation

To check for overfitting, 20% of the proteins are randomly chosen to be left out of the optimization as a validation set. Decisions on the functional form of the interactions and the regularization parameters k_r and V_{lb} are made based on data that is not used in the gradient descent optimization.

The Adam optimizer[12] is used to minimize the energy gap. This optimizer is convenient because it automatically adjusts the gradient descent step size for each parameter according to the typical scale of the gradient in that dimension. This rescaling is important because spline coefficients at large radii tend to have much larger gradient magnitudes than parameters at small radii. For full details of the optimization including initialization, please see section 2.10.

2.6.4 Training results

The accuracy of the results is represented in Figure 2.6 as

$$\text{acc} = e^{-E_{\text{gap}}^{\text{total}}/N_{\text{res}}^{\text{total}}}. \tag{2.30}$$

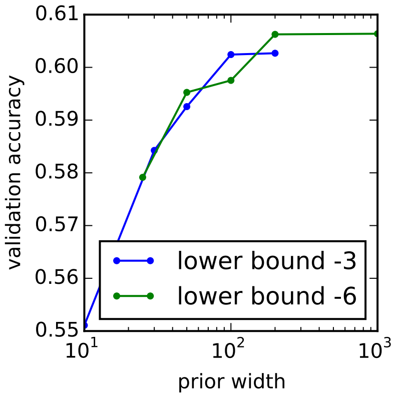


Figure 2.6: Accuracy of the model (probability assigned to true $\tilde{\chi}$ state) versus regularization strength (essentially the inverse of k_r). Regularization as employed in this work does not appear to improve the accuracy of the side chain packing significantly. The regularization, especially the lower bound to the energy, may have a strong effect on the accuracy of protein simulation, even if does not increase packing accuracy.

This represents roughly the geometric mean over residues in the test set of the probability assigned to the true side chain state.

Figure 2.5 depicts the optimized side chain interactions with backbone hydrogen and oxygen. Most of the trends in the figure can be explained by helix capping motifs described in [8]. The prominent role of side chain-hydrogen interactions is consistent with the observation that side chain-backbone hydrogen bonding is more common on the N-terminus than the C-terminus. The strong valine-oxygen interaction is more puzzling, since we do not expect strong, favorable interactions of oxygen atoms with hydrophobic residues. The strong valine interaction may be caused by statistical correlations of valine with the geometry of oxygen atoms that are associated with the *hydrophobic* capping of helices by valine. The valine interaction represents a likely weakness in the training, as the statistical training that maximizes side chain packing accuracy may not capture the physics that drives protein backbone dynamics. The physical reasonableness of the other terms, however, is encouraging.

To compare to state of the art side chain prediction methods, we compare to SCWRL4[14] on its training and validation set of side chains conformations. As per SCWRL's validation procedure, the side chains with less than 25th percentile electron density are excluded. To

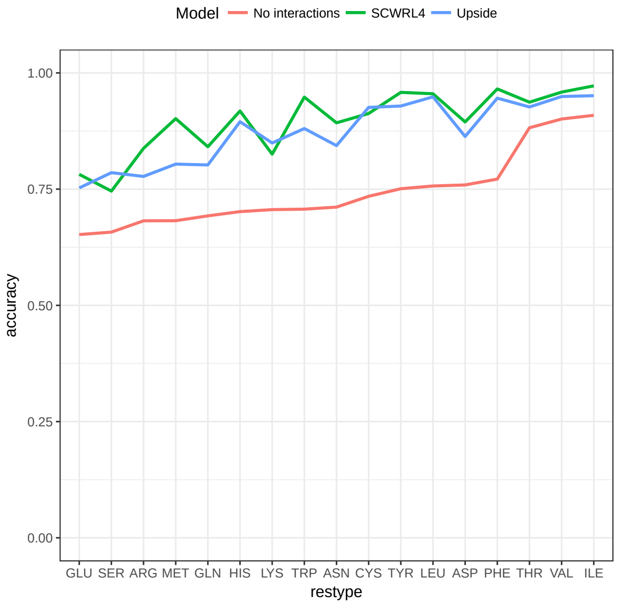


Figure 2.7: Comparison of χ_1 prediction accuracy for Upside and SCWRL4. The “No interactions” line represents the accuracy of the only the NDRD rotamer library without any interactions; this library is used in both Upside and SCWRL. Note that Upside is approximately 150x faster than SCWRL at side chain prediction, in addition to return a probability distribution instead of a single answer.

avoid biasing the comparison toward Upside, the SCWRL set of proteins is split so that 20% of the proteins are withheld for validation, while the rest are used for maximum likelihood training of Upside. The accuracy metric chosen is to calculate the fraction of side chains, excluding glycine, alanine, and proline, for which the Upside or SCWRL predicted χ_1 angle agrees with the crystallographic conformation. This accuracy metric is typically used to assess side chain prediction and is always larger than the geometric mean accuracy used for Upside training for a properly calibrated method.

As seen in Fig 2.7, Upside is very accurate, predicting the correct χ_1 conformation 87.6% of the time. SCWRL4 is slightly more accurate, achieving 89.4% correct predictions, but with a number of limitations. First, Upside is approximately 150x faster than SCWRL4 at predicting side chain conformations with 98% of the accuracy. This enables us to compute the distribution of side chain positions at every step of molecular dynamics at modest cost. Second, Upside provides a Boltzmann probability distribution over rotamer states, enable molecular dynamics using exact forces from the approximate side chain ensemble. SCWRL4 is not well suited for continuous dynamics because it provides only the lowest energy conformation, causing inevitable discontinuities in any attempt to compute forces using SCWRL's predicting side chain conformations.

2.7 Molecular dynamics

Strictly speaking, the parameters obtained from the maximum likelihood training are only optimal for side chain packing for a fixed, native-like backbone geometry. However, we believe that they also encode energetics that can be applied to molecular dynamics simulations. Specifically, in the limit that the model is flexible enough to model the true side chain interactions and there is unlimited training data, the maximum likelihood method would recover the true side chain interaction. Even without having the true form of the side chain interaction, the maximum likelihood parameters assign high probability to the observed rotamer states, thereby including at least some of the underlying physics. The degree to which packing suffices provides a useful energy function to simulate correct backbone structures will be investigated below.

There are caveats to using side chain packing parameters for protein dynamics, even though the parameters are in principle governed by the same physics. The first issue is that a free backbone may move to an unlikely conformation that differs qualitatively from the configurations in the crystallographic data set (e.g., poorly packed or less dense) and the parameters may assign an inappropriately low energy to this conformation. This happens most commonly with the strong parameters of side chain-backbone hydrogen bonding. The backbone may adjust to enable more side chain-backbone hydrogen bonds than are physically realistic because this interaction is geometrically-constrained and hence, inordinately strong. This problematic effect has been observed and may have multiple causes. The first cause is that insufficient strength in other interactions (such as hydrophobic burial) may provide insufficient penalty to the unusual configurations that maximize side chain-backbone hydrogen bonding. The second cause is that the data are trained on low temperature crystal structures. Hydrogen bonds and other interactions may stabilize at low temperature, additional structure may form, and/or the minor populations with side chain-backbone hydrogen bonds broken may not be identified or present in the electron density. We anticipate that for the case of side chain-backbone hydrogen bonding, hydrogen exchange protection factors or NMR observables may be able to resolve the true populations of these interactions. Crystal packing artifacts, as well as limiting training to crystallizeable sections of proteins, may also affect the generated parameters. These artifacts are expected to have a weak effect, though they may bias the model to unphysically bias against unstructured loop regions.

To test the suitability of adapting the side chain packing model to molecular dynamics, simulations were run from the native state of a set of small, fast-folding proteins (protein set adapted from [15]. To create a reasonable protein dynamics model, backbone springs, backbone sterics, hydrogen bond energy, and a basic Ramachandran potential were added to the side chain model. The Ramachandran potential is derived from a coil library[22] as a statistical potential. The hydrogen bond strength is chosen using trial simulations. We choose the hydrogen bond strength $-1.8\,kT$ to maximize the median fraction of simulation frames with RMSD less than 5Å from the native state over the course of a short simulation. For simulation details, see the appendix 2.10. Note that because alanine and glycine have no side chain rotamer states, and hence no training to match the native χ-angles can be conducted, the ALA-ALA, ALA-GLY, and GLY-GLY potentials are completely determined by the regularization. Interactions of ALA and GLY with other residue types are optimized, however, as rotamer states of the other residues provide information on the ALA-X and GLY-X interactions.

The simulations results show that for the majority of proteins, the Upside model does not assign the lowest free energy to the native structure (Fig 2.9). For most proteins, the native structure in the Upside model is however temporarily stable, indicative a local minimum of the free energy surface. The model relaxes quickly to its preferred structure.

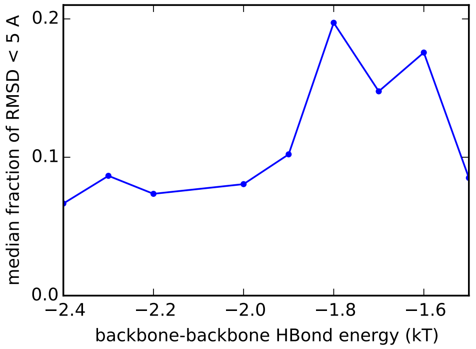


Figure 2.8: Accuracy over short duration simulations. The backbone hydrogen bonding strength is not determined by the packing optimization, so we search for the strength that gives the best simulation accuracy. This is the only parameter in the model directly optimized for simulation accuracy. To assess accuracy, we look at the fraction of each simulation with RMSD to native of less than 5 Å, and we compute the median accuracy over all proteins in the test set. All other results shown are for a backbone hydrogen bond energy of $-1.8\,kT$.

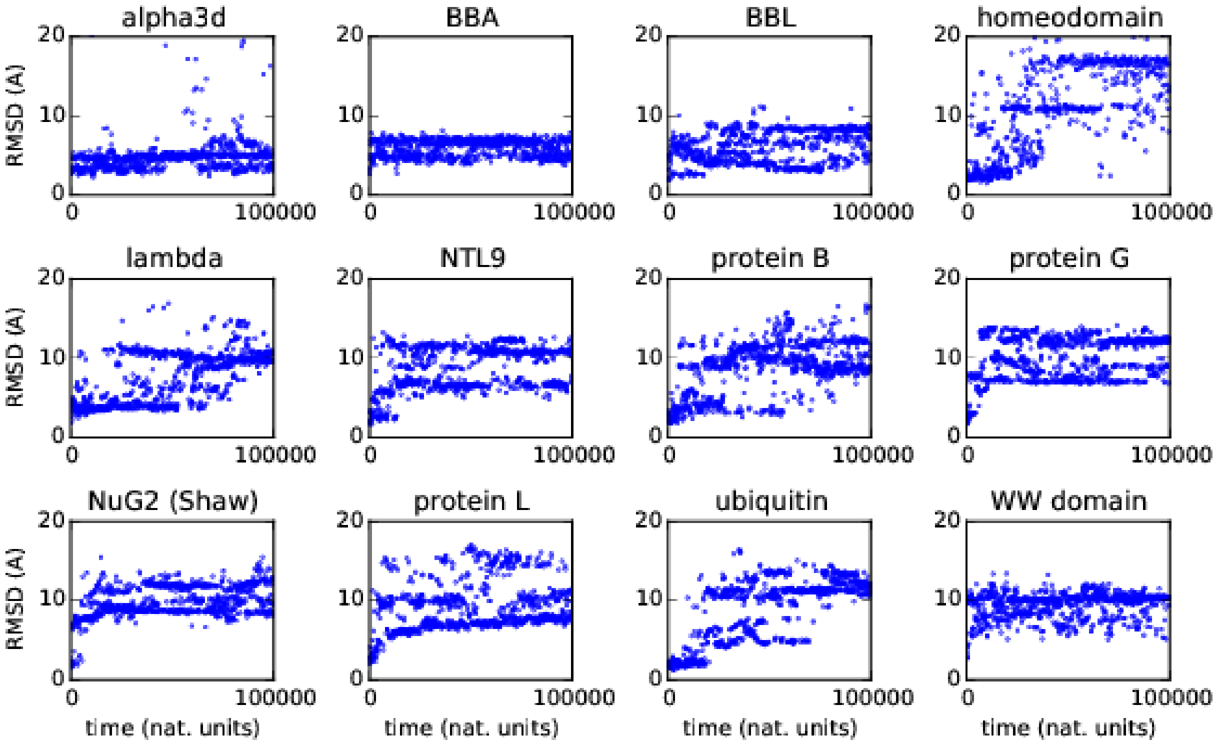


Figure 2.9: RMSD to native over replica exchange simulation trajectories with hydrogen bond energy $-1.8\,kT$.

2.8 Related Work

In the vast literature of coarse-grained modeling, we highlight several strands of work that relate to our modeling. The major features of our model include the following: molecular dynamics on three atoms but with a dynamic ensemble of side chains, optimized discretization of the side chain states to best represent the protein interactions in the coarse-grained model, statistical potential with optimized and state-dependent bead locations and orientations, training a protein interaction model for folding using side chain packing accuracy, and a side chain model with an explicit side chain entropy.

A large body of work, exemplified by SCWRL[14], have studied the prediction of side chain configurations by discrete rotamer states. SCWRL achieves greater than 90% χ_1 accuracy for predicting the most likely rotamer states by minimizing the energy that combines observed rotamer state frequencies and an atomic interaction model[14]. A variety of algorithms have been developed for solving for the highest probability side chain states given the pair interaction values[4, 26]. Kamisetty et al.[11] have worked on scoring protein interaction complexes using a self-consistent approximation to the side chain interactions. Earlier simulation work by Koehl and Delarue[13] use 1-residue mean field techniques to approximate ensembles of side chain conformations but fail to account for the pairwise correlations of the side chain rotamer states. All of these works use atomically-detailed descriptions of the side chains paired with simple or molecular dynamics interaction forms. Their highly detailed side chain with many χ-angles for each residue makes it difficult to perform dynamics sufficiently quickly for folding, and the use of existing interactions (instead of a newly-trained interaction model) makes it difficult to use reduced detail to speed computation. There has also been extensive work in reconstructing backbone positions from side chain beads[5] in lattice models, but these models do not perform a proper summation over possible rotamer states.

There have also been a large variety of coarse-grained techniques that use a variety of non-isotropic potentials for reduced side chain interactions. One of the most successful is the coarse-grained united residue model (UNRES)[16]. The model also uses statistical frequencies to determine the positions of the side chains but it emphasizes the parameterization of the coarse-grained model from physics-based calculations instead of statistical information. Though the potential form (Gay-Berne) used in UNRES is quite different from our work, UNRES also uses non-isotropic side chain potentials[19].

Similar to our work, Dama, Sinitskiy, et al.[3] investigate mixed continuous-discrete dynamics, where the states of molecules jump according to a discrete Hamilitonian. Their method differs from our work in a number of important ways: the authors use discrete jumps in state instead of a free energy summation over all states we employ; they do not optimize the rotamer states as we do; and they train parameters from force matching of molecular dynamics trajectories rather than from the statistical analysis of experimental data in our method.

2.9 Conclusion

We have demonstrated a fast, principled method to coarse-grain discrete side chain states to create a smooth backbone potential. This procedure results in a considerable decrease in computational time as it removes the side chain rattling and friction normally associated with a polypeptide chain moving in a collapsed state. This tracking and instantaneous equilibration of the side chains is analogous to the instantaneously-equilibrated electronic degrees of freedom with respect to the nuclear motions employed in the adiabatic Born-Oppenheimer approximation[2]. Motions are calculated only for three heavy backbone atoms, yet the model contains considerable structural detail including hydrogen bonds involving both the backbone and side chains. Further, we have shown how to parameterize both a tuneable discretization of the rotamer states, and a maximum liklihood procedure to obtain physically-reasonable parameters for our coarse-grain model from X-ray structures. The resulting method is capable of rapid molecular dynamics sampling of protein structures.

The importance of optimizing the bead locations and directions in our model illustrates the principle that chemical intuition can only be a partial guide to accurate coarse-graining of protein interactions. The location of the interaction sites has a strong effect on our model's ability to achieve high-packing accuracy, and we expect similarly strong effects to be observed had we directly optimized for backbone conformational accuracy.

While the side chain packing optimization shows promise as a route to accurate and inexpensive molecular simulation, pairing the resulting potentials with simple Ramachandran and hydrogen bond potentials can maintain the structure of a minority of small proteins tested. Future work will develop co-training of both side chain and backbone parameters to improve simulation accuracy.

2.10 Simulation and optimization details

All simulations are run with Upside, a custom simulation engine that implements the belief propagation of side chain interactions as well as the parameter derivatives needed for gradient descent. Upside is freely available and open source[9].

The temperature is 0.7 natural units. The Ramachandran potential uses the NDRD TCB coil library[22]. The backbone hydrogen bond interaction uses both distance and angle criteria to determine hydrogen bonds. The H-O bond distance interaction starts at approximately 1.4Å and ends at 2.5Å. Both the N-H-O and H-O-C criteria half-heights are at approximately 47 degrees off of collinear.

We use Verlet integration with a time step of 0.009 units. We use the random number generator Random123 [17] to implement the Langevin dynamics with a thermalization time scale of 0.135 time units. The thermalization time scale (related to Langevin friction) is chosen to maximize the effective diffusion rate of chains while effectively thermostatting the simulation. As Langevin dynamics with any friction coefficient produces the same Boltzmann ensemble, we chose to maximize equilibration of our system rather than attempt to match a solvent viscosity.

The derivative calculations need for regularization and coordinate transforms necessary to ensure positive coefficients are handled with the Theano framework[21].

The cutoff radius for side chain-side chain interactions is 7Å, and the cutoff radius for side chain-backbone interactions is 5Å. The distance splines are zero-derivative-clamped cubic splines with a knot spacing of 0.5Å. The angular splines have a knot spacing of 0.167 in $\cos\theta$, which ranges over $[-1, 1]$.

We use the following settings for the Adam optimizer: minibatch size 256 proteins, $\alpha = 0.03$, $\beta_1 = 0.90$, $\beta_2 = 0.96$, $\epsilon = 10^{-6}$. Positivity constraints on the angular coefficients are enforced by a exponential transform. The regularization integrals over all space are approximated by sums at the knot locations of the radial and angular splines.

2.11 Belief propagation

For convenience, this appendix contains a brief description of the equations used to implement belief propagation for the side chain free energies. Given 1-residue energies $v_i(\tilde{\chi}_i)$ and 2-residue energies $v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)$, we seek probabilities $p_i(\tilde{\chi}_i)$ and $p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)$ to minimize the free energy (2.18).

It is helpful to first understand the intuition behind the belief propagation process. We seek a consistent set of one- and two-side chain probabilities for the residues compatible with the interaction potential (2.7). The probability of each residue state $\tilde{\chi}_i$ for residue i is determined by two factors. The first factor is the 1-residue energy $v_i(\tilde{\chi}_i)$ that would determine the probabilities exactly in the absence of interactions. The second factor is consistency with the side chain states of the residues in contact with residue i, where consistency is determined by the potentials $v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)$. Using these factors, the probabilities for residue i are estimated as

$$p_i(\tilde{\chi}_i) \propto e^{-v_i(\tilde{\chi}_i) - \sum_j w_{ij}(\tilde{\chi}_i)} \tag{2.31}$$

$$w_{ij}(\tilde{\chi}_i) = -\log \sum_{\tilde{\chi}_j} e^{-v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)} p_j(\tilde{\chi}_j) \tag{2.32}$$

where $w_{ij}(\tilde{\chi}_i)$ is the effective 1-body potential that residue i feels due to the interaction with residue j. The w_{ij} depends implicitly on the probability distribution $p_j(\tilde{\chi}_j)$ of residue j, so the equations (2.31) and (2.32) must be solved by self-consistent iteration until convergence of the $\{p_i\}$. This algorithm is distinguished from a standard mean-field iteration, which would be identical except the mean-field algorithm would set $w_{ij}(\tilde{\chi}_i) = \sum_{\tilde{\chi}_j} p_i(\tilde{\chi}_j) v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)$. It should be emphasized that, despite the appeal of the intuitive explanation above, the real justification of belief propagation is that the process minimizes the approximate free energy (2.18) as derived in [27]. The iteration is described more formally below, including a damping term λ to suppress oscillations during the self-consistent iteration.

For 1-residue beliefs, define $b_i^r(\tilde{\chi}_i)$ to be the round r "belief" that the i-th residue is in state $\tilde{\chi}_i$. For the 2-residue beliefs, we have two beliefs for each pair of interacting residues (i.e. any pair of residues that have non-zero interaction in any rotamer states). Define $b_{ij}^r(\tilde{\chi}_j)$ to be the round r belief for the residue pair (i,j) that residue j is in state $\tilde{\chi}_j$. The belief $b_{ji}(\tilde{\chi}_i)$ is defined similarly.

To initialize the algorithm at round 0, we take

$$b_i^0(\tilde{\chi}_i) = e^{-v_i(\tilde{\chi}_i)} \tag{2.33}$$

$$b_{ji}^0(\tilde{\chi}_i) = \sum_{\tilde{\chi}_j} e^{-v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)} b_j^0(\tilde{\chi}_j). \tag{2.34}$$

We compute the round $r+1$ beliefs from the round r beliefs according to the following equations.

$$b_i^{r+1}(\tilde{\chi}_i) = \lambda b_i^r(\tilde{\chi}_i) + (1-\lambda) \frac{e^{-v_i(\tilde{\chi}_i)} \prod_j b_{ji}^r(\tilde{\chi}_i)}{\sum_{\tilde{\chi}_i} e^{-v_i(\tilde{\chi}_i)} \prod_j b_{ji}^r(\tilde{\chi}_i)} \tag{2.35}$$

$$b_{ji}^{r+1}(\tilde{\chi}_i) = \lambda b_{ji}^r(\tilde{\chi}_i) + (1-\lambda) \sum_{\tilde{\chi}_j} e^{-v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)} b_j^r(\tilde{\chi}_j) \tag{2.36}$$

The products in equation (2.35) should be understood as taken only over residues j that interact with residue i. The damping constant λ suppresses oscillatory behavior that hinder convergence ($\lambda = 0.4$ is used in the present work). The equations are iterated until $|b_i^{r+1}(\tilde{\chi}_i) - b_i^r(\tilde{\chi}_i)| < 0.001$ for all residues i and states $\tilde{\chi}_i$.

From the converged beliefs $b_i(\tilde{\chi}_i)$ and $b_{ij}(\tilde{\chi}_j)$, we can compute the marginal probabilities

$$p_i(\tilde{\chi}_i) = b_i(\tilde{\chi}_i) \tag{2.37}$$

$$p_{ij}(\tilde{\chi}_i, \tilde{\chi}_j) = \frac{\frac{b_i(\tilde{\chi}_i)}{b_{ji}(\tilde{\chi}_i)} e^{-v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)} \frac{b_j(\tilde{\chi}_j)}{b_{ij}(\tilde{\chi}_j)}}{\sum_{\tilde{\chi}_i, \tilde{\chi}_j} \frac{b_i(\tilde{\chi}_i)}{b_{ji}(\tilde{\chi}_i)} e^{-v_{ij}(\tilde{\chi}_i, \tilde{\chi}_j)} \frac{b_j(\tilde{\chi}_j)}{b_{ij}(\tilde{\chi}_j)}}. \tag{2.38}$$

The free energy of the model is obtained by using the marginal probabilities above in equation (2.18).

2.12 Details of test proteins

Mutations from the indicated PDB structures are indicated in **bold**. The NuG2 sequence is from reference [15].

Name	PDB ID	Length	Sequence
alpha3d	2a3d	73	MGSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAA
			FESELQAYKGKGNPEVEALRKEAAAIRDELQAYRHN
BBA	1fme	28	EQYTAKYKGRTFRNEKELRDFIEKFKGR
BBL	2wxc	47	GSQNNDALSPAIRRLLAEWNLDASAIKGTGVGGRLTREDVEKHLAKA
homeodomain	2p6j	52	MKQWSENVEEKLKEFVKRHQRITQEELHQYAQRLGLNEEAIRQFFEEFEQRK
lambda	1lmb	80	PLTQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQS
			GVGALFNGINALNAYNAALLAKILKVSVEEFSPSIAREIY
NTL9	2hba	39	MKVIFLKDVKGMGKKGEIKNVADGYANNFLFKQGLAIEA
protein B	1prb	53	TIDQWLLKNAKEDAIAELKKAGITSDFYFNAINKAKTVEEVNALKNEILKAHA
protein G	1pga	56	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
NuG2 (Shaw)	1mi0	57	MDTYKLVIVLNGTTFTYTTEAVDAATAEKVFKQYAND**A**GVDGEWTY**DA**ATKTFTVTE
protein L	2ptl	61	VTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG
ubiquitin	1ubq	76	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPP
			DQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
WW domain	2f21	33	KLPPGWEKRMSADGRVYYFNHITNASQWERPSG

2.13 References

[1] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[2] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[3] James F Dama, Anton V Sinitskiy, Martin McCullagh, Jonathan Weare, Benoît Roux, Aaron R Dinner, and Gregory A Voth. The theory of ultra-coarse-graining. 1. general principles. *Journal of Chemical Theory and Computation*, 9(5):2466–2480, 2013.

[4] Johan Desmet, Marc De Maeyer, Bart Hazes, and Ignace Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.

[5] Michael Feig, Piotr Rotkiewicz, Andrzej Kolinski, Jeffrey Skolnick, and Charles L. Brooks. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins: Structure, Function, and Bioinformatics*, 41(1):86–97, 2000.

[6] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[7] Thomas Hamelryck, Mikael Borg, Martin Paluszewski, Jonas Paulsen, Jes Frellsen, Christian Andreetta, Wouter Boomsma, Sandro Bottaro, and Jesper Ferkinghoff-Borg. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PloS one*, 5(11):e13714, 2010.

[8] Edwin T Harper and George D Rose. Helix stop signals in proteins and peptides: the capping box. *Biochemistry*, 32(30):7605–7609, 1993.

[9] John Jumper. Upside. Available at `https://psd-repo.uchicago.edu/freed-and-sosnick-lab/upside-md`, 2016.

[10] John M Jumper, Karl F Freed, and Tobin R Sosnick. Maximum-likelihood, self-consistent side chain free energies with applications to protein molecular dynamics. *arXiv preprint arXiv:1610.07277*, 2016.

[11] Hetunandan Kamisetty, Eric P Xing, and Christopher J Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. *Journal of Computational Biology*, 15(7):755–766, 2008.

[12] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Patrice Koehl and Marc Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *Journal of molecular biology*, 239(2):249–275, 1994.

[14] Georgii G. Krivov, Maxim V. Shapovalov, and Roland L. Dunbrack. Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795, 2009.

[15] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[16] Adam Liwo, Jaroslaw Pillardy, Cezary Czaplewski, Jooyoung Lee, Daniel R Ripoll, Malgorzata Groth, Sylwia Rodziewicz-Motowidlo, Rajmund Kamierkiewicz, Ryszard J Wawak, Stanislaw Oldziej, et al. Unres: a united-residue force field for energy-based prediction of protein structure?orgin and significance of multibody terms. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 193–200. ACM, 2000.

[17] John K Salmon, Mark A Moraes, Ron O Dror, and David E Shaw. Parallel random numbers: as easy as 1, 2, 3. In *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1–12. IEEE, 2011.

[18] Maxim V Shapovalov and Roland L Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

[19] Adam K Sieradzan, Paweł Krupa, Harold A Scheraga, Adam Liwo, and Cezary Czaplewski. Physics-based potentials for the coupling between backbone-and side-chain-local conformational states in the united residue (unres) force field for protein simulations. *Journal of chemical theory and computation*, 11(2):817–831, 2015.

[20] John J Skinner, Wookyung Yu, Elizabeth K Gichana, Michael C Baxa, James R Hinshaw, Karl F Freed, and Tobin R Sosnick. Benchmarking all-atom simulations using hydrogen exchange. *Proceedings of the National Academy of Sciences*, 111(45):15975–15980, 2014.

[21] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[22] Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I Jordan, and Roland L Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput Biol*, 6(4):e1000763, 2010.

[23] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching. In *AISTATS*, 2003.

[24] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[25] Guoli Wang and Roland L Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[26] Jinbo Xu and Bonnie Berger. Fast and accurate algorithms for protein side-chain packing. *Journal of the ACM (JACM)*, 53(4):533–557, 2006.

[27] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

CHAPTER 3

CONTRASTIVE DIVERGENCE

3.1 Introduction

A major challenge is to extract from the collection of crystallized proteins a suitable potential that captures the physics that lead to the protein structures. We attack this problem directly by showing that a strong connection exists between the shape and location of the native basin and the rest of the protein's conformational landscape, and this connection is strong enough to train a potential energy function accurate enough for *de novo* folding simulations. Furthermore, the resulting potential is inexpensive enough to converge simulations of small proteins while running for hours to days on a single computer.

Traditional reference-state methods for determining statistical potentials are difficult to use correctly or to systematically improve. These methods compute a statistical potential using the ratio of the observed counts of residues at a specific distance as compared to the "reference" probability distribution that would be observed when the potential is absent. As Hamelryck[9] points out, these methods can be formally valid, but the true reference frequency depends on all of the other potential energy terms. This creates a self-consistency problem as the reference state for each potential depends on the values of all other potentials. The traditional solution to this difficulty is to avoid self-consistency by simply postulating a reference frequency as is done, for example, in the DOPE model[21]. The great difficulty in providing a good reference state for even simple coordinates is reflected in the great variety of reference states that have been proposed in the literature. To use the self-consistent reference-state method correctly requires computations significantly more involved than those proposed in this work, and this is probably too expensive for current computational resources.

In contrast to traditional reference-state methods, we develop a trajectory-based method to parameterize protein force-fields. Separately fitting individual parameters in the force-field inevitably leads to balance issues. Large-scale issues of energy balance, such as the hydrogen bond terms overwhelming the side chain interactions to make very long helices, may fought by weighting each term in the force-field, but more subtle issues, such as the balance between hydrophobic and charged interactions within the side chains terms, will be missed. The essence of these ideas is that the only way to correctly parameterize a statistical potential is to jointly choose the parameters to optimize the resulting protein ensemble. A contribution of this paper will be to show that we do not need to be able to converge the entire ensemble from proteins in our training set, but we may focus our attention on fluctuations near the native well. By choosing parameters to optimize the near-native basin of the protein, we will obtain a balanced and accurate set of parameters that is sufficient to fold proteins.

Further, trajectory-based training fundamentally alters the relationship between the computational speed of a model and its accuracy. Traditionally, the level of detail in both atoms represented and interaction terms are increased in order to increase accuracy, leading to slower simulation for more accurate models. When optimizing parameters based on finite trajectories, this speed-accuracy tradeoff is replaced to some extent by a speed-accuracy synergy. Models that can be equilibrated more quickly, in CPU-time, are more thoroughly explored by finite trajectories and therefore optimized closer to the ideal parameters for the set functional form. Furthermore, training with inexpensive models allows us to optimize using trajectories from approximately 500 crystal structures in the Protein Data Bank (PDB). We use the Upside model that has been shown to equilibrate extremely quickly while retaining molecular detail, as shown in chapter 2.

We will use a two part strategy for building our model. First, we will describe a physically-plausible coarse-grained model that incorporates many of the key features of protein structure and folding. To find the parameters to populate our model, we will rely on holistic assessment of accuracy, rather than parameterizing based on transfer studies or quantum calculation. We incorporate proven techniques from the machine learning community, who train very complex models on few examples with success in *de novo* prediction. Our modified contrastive divergence method is capable of training the model from exploration around the native state. Even though we only trained for a very limited metric of quality (local exploration), our parameters are plausible as evidenced by the reasonable folding thermodynamics.

We make the following contributions in this paper. We give a clear and proper statistical framework that is practical for training highly detailed protein models. We extend previous contrastive divergence work with molecular dynamics, handling of crystallographic artifacts, and much larger number of parameters. We characterize the ability of contrastive divegence to handle extreme numbers of force-field parameters in a simulation setting. We demonstrate that careful statistical handling of the parameters results in a model with transferable accuracy in three ways: (1) it extends to *de novo* folding despite only being trained on native stability, (2) it produces cooperative folding without being trained for it, and (3) it produces realistic unfolded states with high R_g, which has proven difficult for molecular dynamics models.

3.2 Coarse-grained model

This section recapitulates the Upside model and describes the differences between the Upside model used in this chapter, and the model used for folding simulations in the previous chapter. For exhaustive details of the Upside model, see chapter 4.

The key insight of chapter 2 is that a suitable self-consistent free energy can be defined from the side chain interactions, and that free energy provides an smooth potential for the simulation of backbone dynamics. Our previous work parameterized the model to maximize the accuracy of side chain conformations. While these parameters are used as the initial parameters for optimization, all interaction parameters are retrained in the current work except for the locations and directions of the side chain beads. Our current work extends the Upside model with an additional pseudo-solvation term and better-adapted backbone Ramachandran potential.

The majority of interaction parameters belong to the pair interactions between side chains. All of the pairwise interactions have the same functional form

$$\begin{aligned} V = \kappa(\,&\mathrm{unif}(r_{12}) + \\ &\mathrm{ang}_1(-n_1 \cdot n_{12})\,\mathrm{ang}_2(n_2 \cdot n_{12})\,\mathrm{dir}(r_{12})), \end{aligned} \tag{3.1}$$

where unif, ang_1, ang_2, and dir are arbitrary curves represented by cubic splines represented by cubic splines. For details of the distance coordinate r_{12} and the angular coordinates n_1, n_2, and n_{12}, see the Fig. 2.5. There are $\binom{20}{2} = 210$ types of amino acid pairs and 62 spline coefficients per pair, giving 13020 side chain-side chain interaction parameters. There are also five interaction sites on the backbone, roughly representing the H, O, N, C^{α}, and C atoms, with 54 parameters per interaction due to a smaller cutoff radius. The total number of side chain-backbone interaction parameters is 5400. The cutoff radius of the side chain-side chain interactions is 7Å and the cutoff radius for the side chain-backbone interactions is 5Å. The smaller cutoff radius for side chain-backbone interactions encourages the model to use this term to describe side-chain backbone sterics and hydrogen bonding, as well as reducing the computation expense from these terms.

We add an additional term, similar in spirit to reference [1], to capture solvation effects in a simple-minded way. For each residue, we compute the number of side chain beads within an approximate hemisphere above the C^{β}. While we could include other heavy atoms in the hemisphere calculation, it would needlessly increase the cost of the calculation. To handle the uncertainty of rotameric state that can affect the count of side chain beads, the count for different rotameric states are weighted by the prior probabilities of the rotamer states.

Mathematically, the count is given by

$$N_i = \sum_{\substack{j \\ |i-j|>2}} \sum_{\chi_i} p(\chi_i) S(|y_i(\chi_i) - y_i^{\mathrm{CB}}| - (8.\,\text{Å}), (1.\,\text{Å})) \\ S(\mathrm{angle}(y_i(\chi_i) - y_i^{\mathrm{CB}}, v_i^{\mathrm{CB}}) + 0.1, 1.), \tag{3.2}$$

where S is sigmoid-like cutoff function. Low values of N_i are typical of residues exposed to solvent and high values of N_i are characteristic of buried residues. A arbitrary energy curve is coupled to the value of N_i so that

$$V_{\mathrm{env}} = \sum_i v_{a_i}^{\mathrm{env}}(N_i). \tag{3.3}$$

While many more sophisticated solvation potentials have been derived, this term has the advantages being very fast and easily optimized by the contrastive divergence procedure, while remaining flexible enough to represent many of the solvation effects omitted by the side chain potential.

The core backbone Ramachandran potential is a simple $\sum_i V_i^{\mathrm{rama}}(\phi_i, \psi_i)$, where V_i^{rama} depends on the chemical identity of the $i-1$, i, and $i+1$ residues. The Ramachandran potentials are based on the turn, coil, or bridge (TCB) Ramachandran probability models in the NDRD backbone library[26] and an additional frequency model of Ramachandran angles for sheet residues. The single trained parameter in the model related to the backbone configuration controls the mixing ratio of TCB angles from the NDRD library with the sheet libraries that we parameterized. We introduced extra sheet probability to our backbone model to counteract an observed tendency for our model to overstabilize helices during folding. Both the NDRD probabilities and our sheet probabilities are dimer models that depend only on a single left or right flanking residue, and the following formula is used to

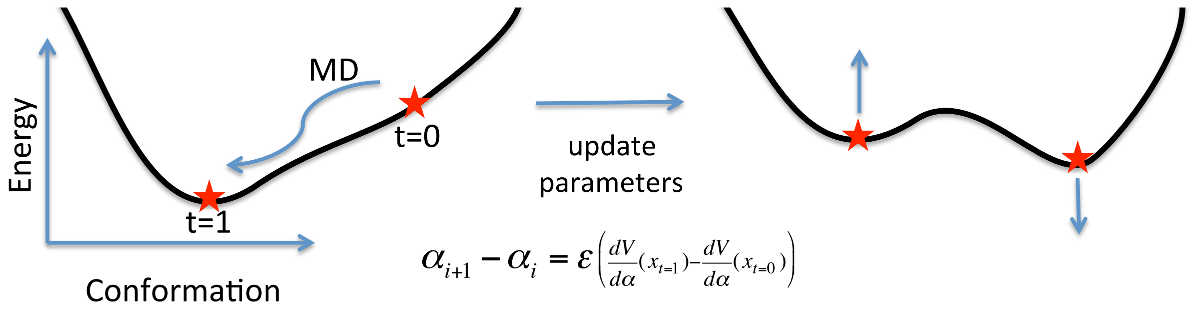

MD
t=0
t=1
Energy
Conformation
update
parameters
$\alpha_{i+1} - \alpha_i = \varepsilon\left(\frac{dV}{d\alpha}(x_{t=1}) - \frac{dV}{d\alpha}(x_{t=0})\right)$

Figure 3.1: Cartoon of the contrastive divergence algorithm

convert the base probabilities into the potential

$$V_i^{\text{rama}}(\phi_i,\psi_i) = -\log(p^{\text{TCB}}_{a_{i-1}a_i a_{i+1}}(\phi_i,\psi_i) + \frac{e^{-w}}{2}(f^{\text{right}}_{a_i a_{i-1}} + f^{\text{left}}_{a_i a_{i+1}})p^{\text{sheet}}_{a_{i-1}a_i a_{i+1}}(\phi_i,\psi_i)) \quad (3.4)$$

$$p_{a_{i-1}a_i a_{i+1}}(\phi_i,\psi_i) = \frac{1}{2}(p^{\text{right}}_{a_i a_{i-1}}(\phi_i,\psi_i) + p^{\text{left}}_{a_i a_{i+1}}(\phi_i,\psi_i)) \quad (3.5)$$

where $f^{\text{left}}_{a_i a_{i+1}}$ is the fraction of (a_i,a_{i+1}) dimers in the PDB where the i-th amino acid is in a β-sheet secondary structure. Note that this model uses the arithmetic average to combine dimer probabilities rather than the geometric average recommended in the NDRD model. Using the arithmetic average results in smoother, less sharply peaked probabilities to encourage the side chain potential to exert more influence over the backbone geometry.

The backbone non-bonded interactions are governed by a hydrogen-bonding potential and a steric repulsion. The hydrogen bonding potential depends on both distance and angle for the participating atoms, and the energy of forming a hydrogen bond is a single parameter that is chosen by contrastive divergence. The backbone atoms N, C$^\alpha$, C, and C$^\beta$ feel an repulsive interaction at approximately 1.5Å.

3.3 Contrastive divergence method

This section describes contrastive divergence heuristically. For a mathematical derivation, see section 3.16.

The motivation of contrastive divergence is to consider two ensembles, one closely linked to the crystal structures and one free to diffuse away under Langevin dynamics. Under an ideal physical model, the crystal-based ensemble and the free ensemble would be quite similar, up to artifacts of crystallographic structure determination. For an inexact coarse-grained model, systematic differences will arise between the crystal ensemble and simulation ensembles. For example, the simulation ensembles may have a higher number of backbone-backbone hydrogen bonding than are present in crystal structures. In such a case, making the energy of forming a hydrogen bond less favorable will shift the simulation ensemble to better resemble the crystal ensemble. Such a modification of the potential will have additional effects; weakening the hydrogen bonding interaction may increase the amount of hydrophobic burial. Thus, we iteratively modify all the parameters to shift the simulation ensemble to better match the crystal ensemble. The algorithm converges when there is no parameter that can distinguish the simulation ensemble from the crystal ensemble.

The unrestrained simulation ensemble is obtained through finite-time Langevin dynamics. For each protein in the training set, the simulation ensemble given by 5000 time units of dynamics, of which the first half is discarded as equilibration. This corresponds to about ten minutes of wall-clock time for each simulation. Unless the native state is particularly unstable, this time is insufficient to explore the conformational landscape much beyond the native basin. Instead this typically creates a locally-equilibrated ensemble that relaxes the crystal conformation and explores fluctuations in the near-crystal conformational basin.

The crystal ensemble is traditionally defined in contrastive divergence as consisting solely of the true data points. This δ-function distribution is problematic for protein structures because of the sharpness of the protein energy function. A slight inaccuracy in the crystallographic reconstruction, or a slightly incorrect geometry in the Upside protein interactions, will cause a very unfavorable energy. Furthermore, due to crystal packing and reconstruction artifacts, we would always expect some relaxation of the crystal structure under aqueous conditions. To reduce the impact of these issues, we replace the exact ensemble of crystal structures with the ensemble of running Langevin dynamics restrained to be near the crystal structure, approximately 1Å RMSD.

To shift the free simulation ensemble toward the crystal ensemble, we change parameters α_i in proportion to the amount that they differentiate the simulation and crystal ensembles,

$$\alpha_{i+1} = \alpha_i + \frac{\epsilon}{M} \sum_{a=1}^{M} \left(\left\langle \frac{dV}{d\alpha_i}(X) \right\rangle_{\text{restrained}} - \left\langle \frac{dV}{d\alpha_i}(X) \right\rangle_{\text{free}} \right). \tag{3.6}$$

The quantity $\left\langle \frac{dV}{d\alpha_i}(X) \right\rangle_{\text{restrained}} - \left\langle \frac{dV}{d\alpha_i}(X) \right\rangle_{\text{free}}$ represents a pseudo-derivative of the free energy of restraining the simulation to be near the crystal structure. In the limit that the simulation duration is infinite, this difference is the exact derivative of the free energy. In practice, this difference chooses a suitable direction to improve the parameters.

3.4 Handling crystallographic artifacts

The derivation of contrastive divergence presented above makes the assumption that the conformations X_a are equilibrium samples from the Boltzmann distribution of each protein, but in reality, we must work with crystal structures of proteins. While it has been shown that that the static diversity of crystal structures for different proteins conveys significant information about the dynamic ensembles of individual proteins [12]. Crystal structures deviate in a number of systematic ways from equilibrium samples, but we are most concerned about crystal packing artifacts, crystallizability bias, and errors in published structures.

We expect that our bias in working only with crystallizable sequences, thus missing intrinsically disordered regions from training, likely biases the resulting potential to disfavor coil states. The loop-stabilizing effects of crystal packing somewhat counteract this effect, as it allows longer loop regions to exist in crystal structures.

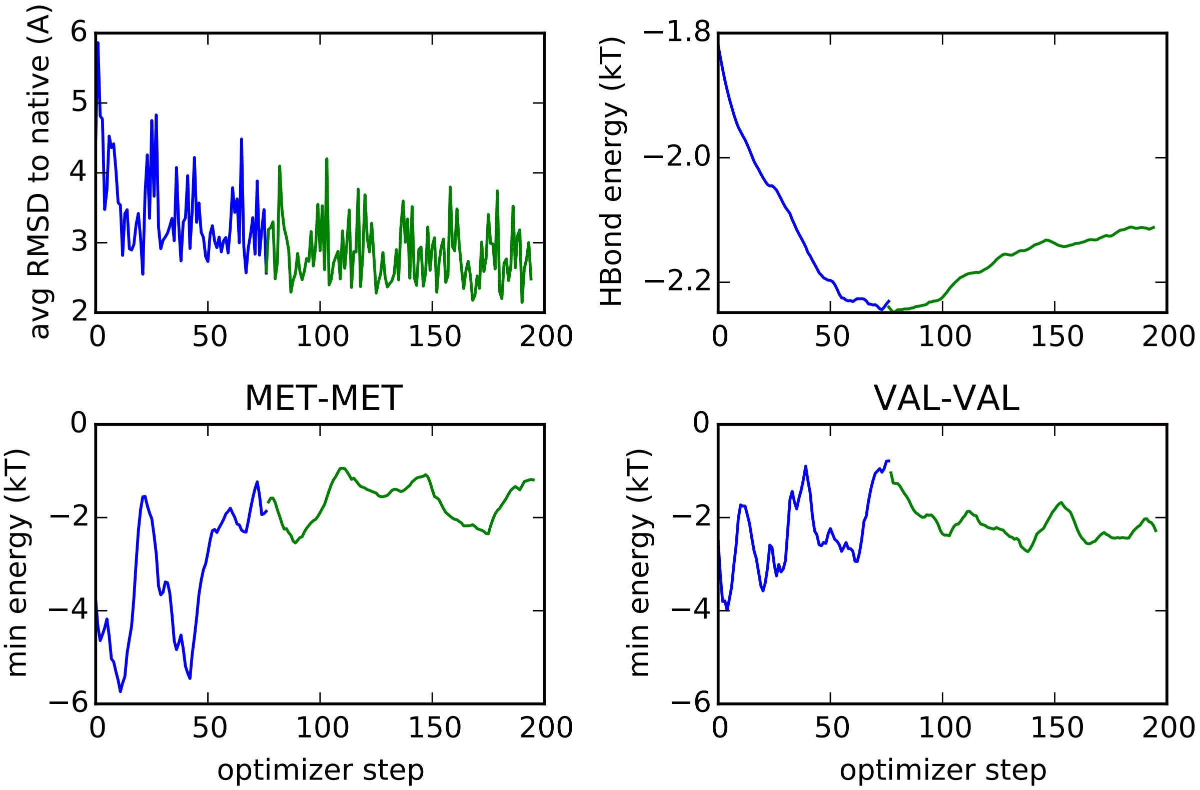


Figure 3.2: Progress of contrastive divergence training. In all plots, the blue curves indicate larger step-size training and the green plots indicate smaller step-size fine-tuning. The upper left plot show the decline in minibatch-averaged RMSD over the course of the optimization. The remaining plots show the convergence of the hydrogen bonding and side chain-side chain parameters over the optimization. The larger step-size optimization of the side chain parameters uniformly shows large oscillations that inhibit convergence.

3.5 Optimization

For practical reasons, the contrastive divergence simulations are run for a short time that does not permit complete exploration of configuration space. Each contrastive divergence simulation is run for 4000 time units, corresponding to approximately ten minutes of wallclock time per iteration. The simulations use temperature replica exchange with eight replicas to enhance barrier crossing of the contrastive divergence[20], while the temperature intervals of the replicas scale with $1/\sqrt{N_{\text{res}}}$ to encourage good replica exchange efficiency for proteins of various sizes. The progress of the replica exchange is monitored by the average best-fit RMSD-to-crystal over the simulation for each minibatch.

The initial parameters from the potential come from optimizing side chain accuracy of the model using a procedure similar to our work in the previous chapter. The contrastive divergence training rapidly improves this model as there is a quick decline in average RMSD over a minibatch from 5Å to 3Å. This decline is accompanied by rapid movement of the parameters, especially the hydrogen bond strength. At the same time, the side chain parameters show much greater fluctuations. This is likely because there are far fewer interactions for any particular pair of residue types (say ALA-GLY) than there are hydrogen bond interactions. For this reason, we expect the gradients of the side chain-side chain parameters to be far noisier than those of the hydrogen bond interactions. To reduce the fluctuations and fine-tune the results, we reduce the optimizer step size by a factor of four after two epochs. While the change in the observed RMSD has become small, there are indications that we have not converged the value of the parameters and thus are stopping the contrastive divergence early. Earlier tests showed that continuing the contrastive divergence to complete convergence does not necessarily produce better results, which has been observed in [5]. Particulary, if large barriers have been built around the native states by contrastive divergence, there may be little relaxation of the conformation during the short simulations of contrastive divergence and hence little useful information to optimize the parameters. Instead, the further fine-tuning of the contrastive divergence results may *reduce* the accuracy of the model by further optimizing against only very near-native results. Secondly, early stopping of optimization has been observed in a number of contexts to function as a regularizer that favors simpler models[6].

There is a curious behavior of the hydrogen bond strength, where it appears to converge to a significantly smaller value during the fine-tuning than during the larger optimizer steps. We speculate that the extra noise in the side chain interactions during the larger optimizer steps may in aggregate cause stronger side chain interactions for the protein. This would necessitate a large hydrogen bond energy to balance against the side chain interactions.

3.6 Accuracy of *de novo* Folding

The previous training only optimized the parameters' ability to hold a protein conformation for a short period of time, less than ten minutes of wallclock time. While contrastive divergence training has been shown to train models well for many machine learning problems [4], the accuracy must be demonstrated for our model. As a stringent test of the protein model and our training procedure, we attempt *de novo* folding of a benchmark set of small, fast-folding proteins similar to those used in [15, 1]. Before training, we removed all proteins from the training set that were homologous to any protein in the benchmark set to ensure that this would be a *de novo* prediction. The proteins are simulated using replica exchange using 16 replicas for approximately three days of wall-clock time with one processor per replica.

Two replica exchange simulations are launched for each protein with the lowest temperature of the replicas the same as the contrastive divergence temperature. The first is initialized from the native configuration of the protein to assess the stability of the experimental structure. The second simulation is initialized from a random unfolded state with ϕ and ψ angles chosen uniformly at random. The range of temperatures were chosen to be large enough to span to signficant populations of unfolded states for all proteins.

We judge the accuracy and equilibration of the model from the histogram of best-fit

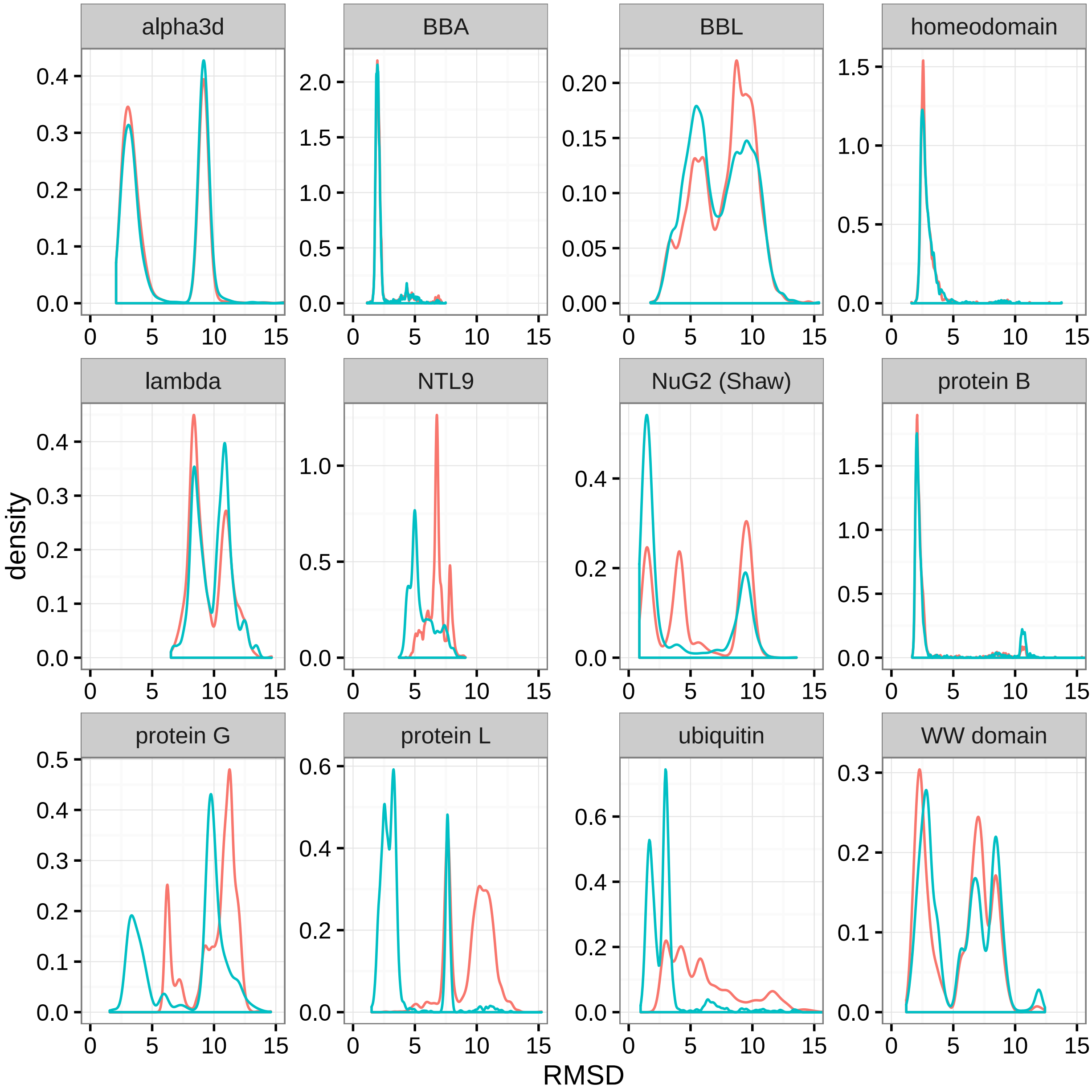


Figure 3.3: RMSD distributions after equilibration phase. Green indicates simulations started from the crystallographic native structure and red indicates simulations started from a random unfolded state.

Name	Length	Lowest RMSD (Å)
alpha3d	73	2.0
BBA	28	0.7
BBL	47	1.7
homeodomain	52	1.5
lambda	80	4.3
NTL9	39	2.6
protein B	53	1.7
protein G	56	3.7
NuG2 (Shaw)	57	0.9
protein L	61	2.6
ubiquitin	76	2.0
WW domain	33	0.9

Table 3.1: Lowest RMSD for *de novo* folding simulations.

RMSD deviations from the native structure after discarding the initial third of the simulation as equilibration. When the native-initialized and unfolded-initialized structures have similar RMSD distributions, the simulation has likely converged. In protein L and ubiquitin, we note that the ensembles are relatively far from convergence as the native- and unfolded-initialized simulations disagree strongly in their RMSD distributions.

It should be noted that, while no parameters are set based on observing the benchmark folding results, decisions on the functional form of energies are made based on the effect on folding results. While we expect that any statistical bias toward higher accuracies that arises from making force-field decisions after seeing folding results is small, we do note these "researcher degrees of freedom"[22] for full disclosure.

Multiple conformations are observed for many of the simulated proteins. This is consistent with the known properties of maximum likelihood training, where inability of the functional form of the model to represent the true Boltzmann distribution of the model will result in smaller, broader energy functions and thus unsurprising to see multiple conformations in the ensemble. Since contrastive divergence is an approximation to maximum likelihood, it is unsurprising that it inherits this conservative property.

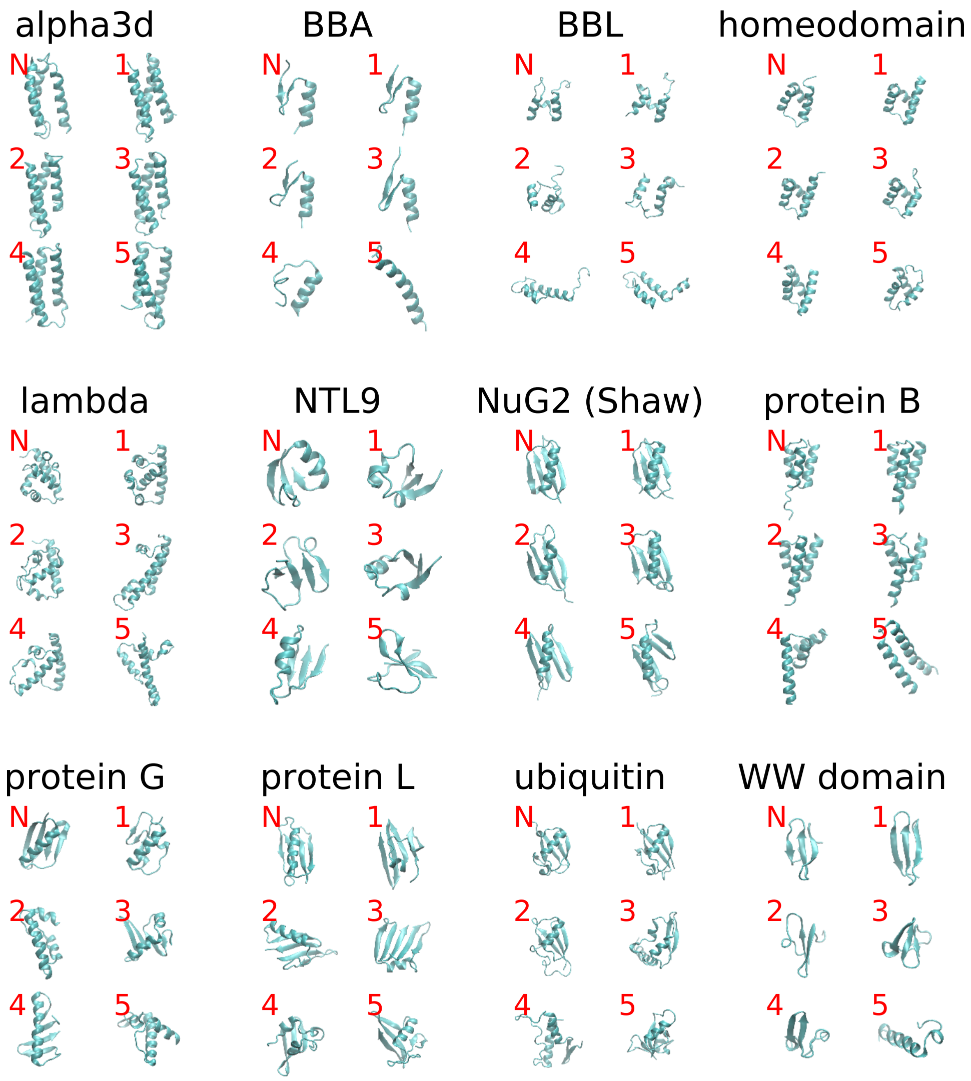
3.7 Structural characterization of low temperature conformations

To characterize the structural ensemble at low temperature, we cluster each trajectory into five clusters and choose a conformation in a high density region of each cluster to represent the cluster, as detailed in section 3.15.

The majority of the proteins, excluding BBL, show a small number of well-defined and stable basins that represent the dominant conformations of Upside for each protein. These well-defined clusters are clearly indicated in the principal component plots in Fig. 3.12. While the simulation often produces many of the these conformations quickly, the equilibration of their populations takes time, likely dominating the relaxation time of the RMSD distributions. Still, this relaxation time is on the order of hours to days of wallclock time, making it extremely quick in comparison to typical molecular dynamics simulations.

As indicated by the clustered structures, the Upside simulations tend to correctly represent the secondary structure of the proteins even as it provides a small number of distinct tertiary arrangements. This tertiary diversity is illustrated in the mirror three helix bundles for α3d and protein B, as well as the subtle re-arrangements in NuG2. As these structures coexist with similar probabilities at low temperature, we hypothesize that the short-time contrastive divergence we are using does not provide a sufficient library of large changes in the tertiary structure to enable the potential to properly distinguish the various conformations. It is difficult for short-time simulation to produce sweeping changes that preserve secondary structure but cause permutations of the tertiary structure.

As these results are obtained with an almost untuned backbone energy fucntion (only a single scalar parameter to control the amount of sheet in the Ramachandran potential), it is likely that loop conformations would be significantly improved with a more carefully tuned backbone potential. Presumably the backbone potential can be optimized using the same contrastive divergence procedure as we have used in this work. It remains only to



alpha3d

BBA

BBL

homeodomain

lambda

NTL9

NuG2 (Shaw)

protein B

protein G

protein L

ubiquitin

WW domain

Figure 3.4: Structures of the native state (N) and a representative structure within each cluster of the *de novo* trajectory in the lowest temperature replica. The clusters are ordered by average RMSD-to-native within the cluster. Representative are chosen from regions of high density in the principal component analysis plots of the trajectory.

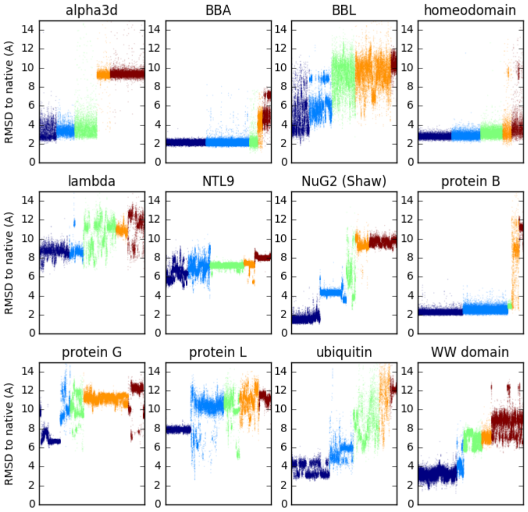


Figure 3.5: RMSD trajectories within each cluster, indicated by color. Clusters are ordered by the average RMSD-to-native. Within each cluster, the data points are in trajectory order.

choose a suitable backbone model, such as TorusDBN[3], that allows sufficient freedom for the optimization. In future work, we intend to co-adapt an existing backbone statistical potential to the Upside energy as part of the contrastive divergence procedure.

3.8 Melting behavior and unfolded states

The point of the contrastive divergence training is to capture the energy of fluctuations about the native state of the protein. As we have shown above, the energies learned by contrastive divergence are sufficient to assign significant probability to the native state at low temperatures. We are also interested in the melting and folding behavior of the model, specifically whether the excited states of the model are consistent with experimental data.

The model typically exhibits concerted melting behavior over a small range of temperatures. While the temperature of the model in Upside is not exactly comparable to a physical temperature, it is reasonable to assume $T = 1$ corresponds roughly to a temperature of 300 K. The ubiquitin transition occurs over a temperature range of approximately 0.07 temperature

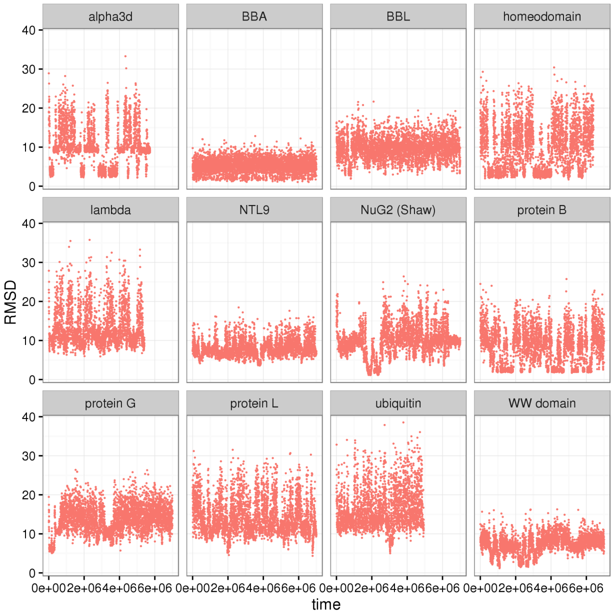


Figure 3.6: Time vs RMSD at the temperature of peak heat capacity. Heat capacity for each temperature is estimated by the fluctuation formula $C_\mathrm{p} = (\mathrm{var}\, E)/T^2$.

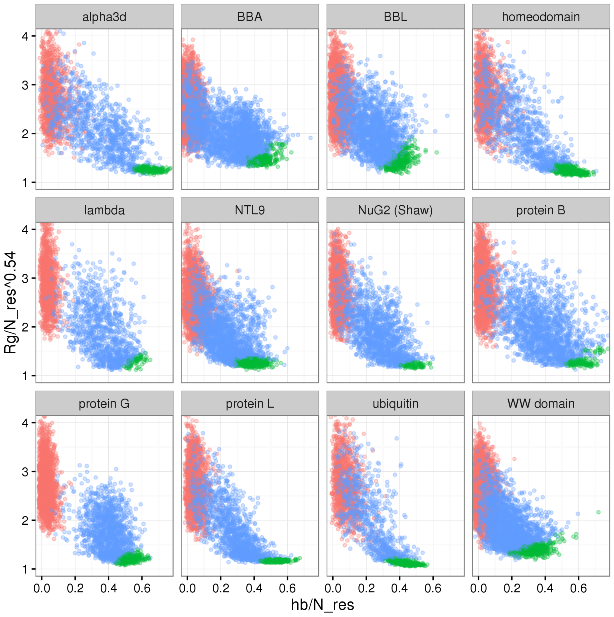


Figure 3.7: Fraction of formed HBonds vs R_g at the temperature of peak heat capacity

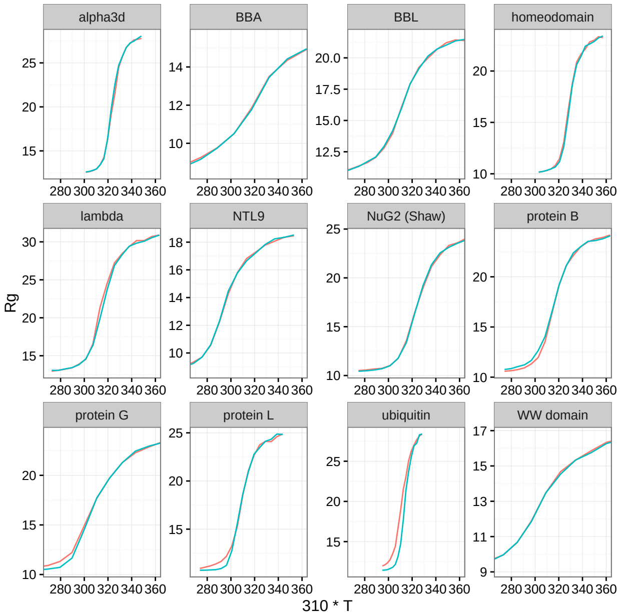


Figure 3.8: Root-mean-square R_g as a function of temperature

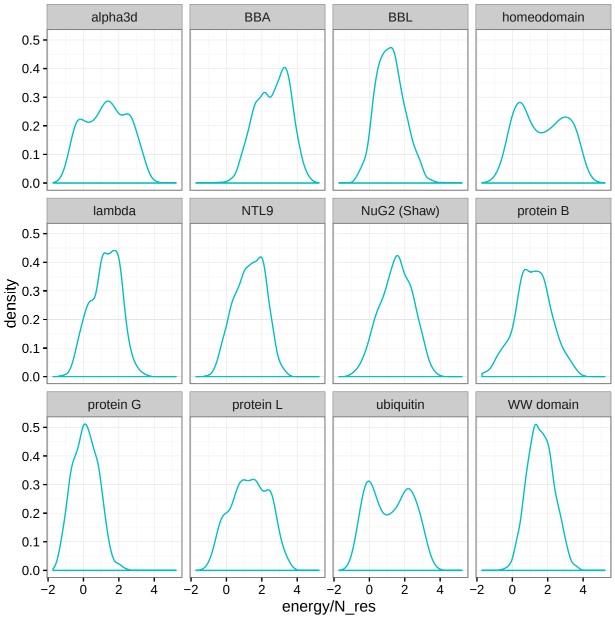


Figure 3.9: Energy frequencies at the temperature of peak heat capacity

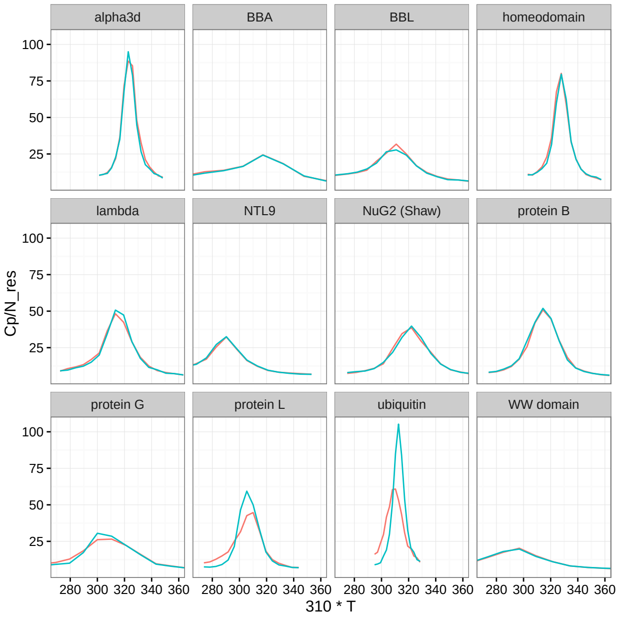


Figure 3.10: Heat capacity as a function of temperature. Heat capacity for each temperature is estimated by the fluctuation formula $C_\mathrm{p} = (\mathrm{var}\, E)/T^2$.

units, which equates to approximately 20 K in temperature. All the properties described below emerge spontaneously from training the model for low-temperature accuracy with contrastive divergence.

Furthermore, our temperature-denatured states have high R_g near the midpoint of the transition, consistent with experimental results and inconsistent with many all-atom molecular dynamics folding simulations [23, 11]. We indicate two different unfolded R_g, one at the peak of the heat capacity and the other at a high temperature, where the heat capacity is near its unfolded baseline (temperature in each case depends on the protein studied). Our R_g at the peak of the heat capacity is about 15% under the experimental values and the R_g at high temperature is approximately 10% above the experimental expectation. Both values are significantly larger than previous atomistic molecular dynamics.

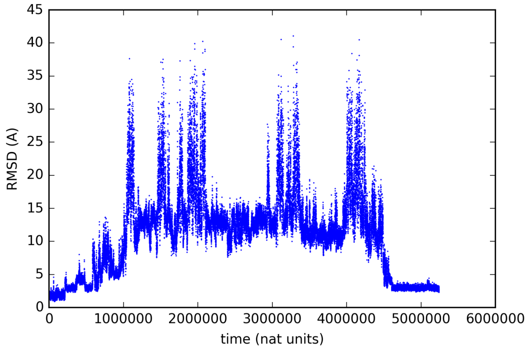


Figure 3.11: Constant temperature reversible folding trajectory of ubiquitin at a temperature of 1.003. Note that the structure becomes completely extended at high RMSD before returning to a structure with an RMSD-to-native of 2.3Å.

3.9 Accuracy of *de novo* Folding

We are able to show constant temperature, reversible folding to the experimental native state for a number of proteins in our test set. This is illustrated in Figures 3.6 and 3.11. Each of the simulations are run on a single core for a few days. The time scales of folding implied by these reversible folding trajectories implies that the time scales of the contrastive divergence simulations are far less (often a factor of 100 or more) than required to equilibrate these proteins. It is clear that the contrastive divergence is not optimizing over all conformations of the system, but only local fluctuations. The reversible folding at constant temperature also indicates that we have an model to explore folding pathways in reversible conditions. The power of our optimization procedure also means that we can take a number of qualitative models and make them quantitative by optimizing the parameters for *de novo* folding, so that we can explore folding pathways for a number of reasonable folding models without having to resort to strong native bias.

We have a model that produces a wide range of protein-like behavior for folding, but it is not yet clear which parts of the model are associated with which behavior, so that we lack a truly reduced description of the protein physics that identifies certain features as key (I could argue that this is not possible since statistical correlations tend to confound mechanistic description in ways that practitioners do not appreciate). Given that these features are produced without any specific attempt to generate them suggests that we are approaching a good model of protein dynamics. Furthermore, contrastive divergence, though its relationship to maximum likelihood, is tuned to pick up fluctuation scales in addition to the best structure. Emprically, we have noticed that a large increase in cooperativity is observed by going to optimized, precise side chain locations for the side chain model, as show in the previous chapter. This suggests that we are able to predict the improvement in the behavior of the protein physics by observing an increase in side chain prediction accuracy, which is a much easier criterion to assess. Future work will address the quality of predicted pathways and observables for individual proteins.

Note that conditional on low hydrogen bonding, the R_g at high temperature and at the peak heat capacity are quite similar. This suggests the increase in R_g for the unfolded state as temperature increases is driven by a reduction in backbone-backbone hydrogen bonds rather than side chain effects.

Based on the above results, there are two facts that must be reconciled. The first fact is the sharp phase transition with a single peak for the heat capacity. The shape of the phase transition, but not its amplitude, is consistent with a cooperative folding transition. Additionally, the relationship of average R_g to average number of hydrogen bonds is approximately linear over the transition region. The second fact is the large residual hydrogen bonding in the denatured state at the heat capacity peak, which suggests that the transition is quite non-cooperative. We propose that this may be explained by the essential feature of the contrastive divergence process, that it must balance the competing energy terms of the model so that no one energy dominates. We suspect that the very close temperatures of disrupting tertiary contacts by leaving the native state and then melting hydrogen bonds results from balancing the interaction energies of side chains pair interactions and backbone-backbone hydrogen bonds. A small tweak to the contrastive divergence training may be able to push the temperature of melting secondary structure lower so that the folding is significantly more cooperative.

3.10 Related Work

The most similar work to ours is a contrastive divergence optimization of a Gō-like protein potentials sampled with crankshaft Monte Carlo moves[19, 27]. These works optimized only tens of parameters using contrastive divergence, and the resulting model is only used to fold protein G and 16-residue peptides.

Early work focused on training protein energy terms against a library of decoys. Such efforts are problematic for a number of reasons. The first is realistic molecular dynamics energy functions have extremely rugged energy landscapes, so that the energy of a decoy may be much higher than another structure less than an angstrom away in RMSD. This ruggedness means that it is improper to score decoy prediction by energy without first relaxing the decoys to the center of the nearby energy well. The significance of using unrelaxed decoys depends on how close the decoy generation model is to the force field being trained. Another difficulty is that the conformational space of proteins is vast; where it not so, finding the lowest energy conformation for a given energy model would be easy. As a result, decoy sets are unable to exhaust conformational space and are likely to miss some conformations that incorrect protein energy functions find extremely favorable. These effects suggest that the best decoy set may be obtained simply by conformational sampling of the protein energy function. Decoy generation may provide convenient seed structures for exploration, but ultimately protein conformational sampling is needed to evaluate potential energy functions and suggest directions for improvement.

A more technical distinction against traditional protein training methods, such as Z-score optimization [16], is that it is relatively unimportant to know how well the model handles average decoys. The important task is to produce a low energy ensemble (those structures no more than a few $k_\mathrm{B}T$ above the relaxed experimental structure) having a high population of low RMSD structures. While improving Z-score should correlate with a better conformational ensemble early in training, in later training it will likely become important to focus on the remaining conformations that are lower in energy than the native, irrespective of the effect on Z-score. Finally, we note that Z-score is inherently defined relative to a decoy ensemble (it is near meaningless with respect to the Boltzmann ensemble of the model). The Z-score is not defined comparing the native to all other conformations, as denatured conformations are exponentially more numerous than folded conformations. Instead, one must define a class of decoy structures that are somehow folded-protein-like in order to define the Z-score. Methods based on simulations ensembles and the associated probability density (such as maximum likelihood and contrastive divergence) are well-defined and do not need to create a class of decoys.

There has been previous work on estimating protein statistical potentials using simulation trajectories. Following the long history of work for using contrastive divergence on machine learning models, [18] have applied contrastive divergence to few-parameter protein models. The present work extends these methods to handle large-scale molecular dynamics training, where the scale is large in both the number of parameters and the amount of protein simulation training used. Furthermore, we go beyond existing work by demonstrating that our method is able to fit a model transferable to realistic unfolded states and thermal melting behavior, consistent with a cooperative, pseudo-first-order model of protein folding.

There have been several attempts to train a force-field usingmaximum likelihood. These methods are inherently limited by the need to compute the derivative of the free energy, which involves a summation over an equilibrium sample of the configurations. Such a requirement necessitates a very long simulation to update parameters. Still, the maximum likelihood approach can be viable when used with very small proteins on which the simulations converge quickly. A variant of maximum likelihood is given in [31], where decoys are generated and a maximum likelihood model is fit to adjust the parameters to distinguish between near-native and far-from-native conformations (smoothed with a Gaussian cutoff of nearness). The potential is trained on a single protein, tryphtophan cage, and then the resulting potential is applied to a number of α-helical proteins.

3.11 Conclusions

These data suggest that the Upside model may be representing the physics of protein folding well enough to be useful far from its training temperature. In particular, future work will investigate if the folding pathways of the *Upside* model are consistent with experimental data.

We have shown that extremely short simulations, paired with optimization of contrastive divergence, are capable of parameterizing an accurate coarse-grained potential. Replica exchange molecular dynamics with this energy is capable of folding many small, fast-folding proteins to sub-3Å accuracy in a matter of days on a single machine. Training the model by contrastive divergence is successful precisely because the Upside model has extremely quick equilibration for many proteins. The usual trade-off in molecular simulation is to choose between accurate, expensive simulation models (such as explicit water molecular dynamics) and inexpensive, less accurate models. With contrastive divergence training of the model, the situation is reversed to some extent. Less expensive models allow more extensive exploration during the training phase, which allows the contrastive divergence better maximize the accuracy of the chosen parameters. Coarse-grained models must still retain sufficient flexibility to represent the underlying physics, but faster models will come closer to their ideal parameters due to the enhanced exploration. Simultaneously, we have show that very large numbers of parameters (approximately 25000 in our case) are no obstacle to producing accurate proteins models using trajectory-based training. While overfitting is still a real concern, the severity is greatly reduced because contrastive divergence is training *against* the vast possibilities of alternative protein conformations explored by conformational sampling. Overstrengthing any particular interaction will lead to that interaction being present in a large number of the structures found by molecular simulations, which will cause negative feedback in the training that reduces the magnitude of that parameter. By this mechanism, contrastive divergence automatically obtains balanced parameters such that no particular interaction overwhelms the others. Even if we have clear experimental evidence about the strength of a few interactions (such as backbone-backbone hydrogen bonding), it is inadvisable to force those known interactions to their experimental values. It is more important for the different interaction terms to have balanced strength than to have precise accuracy for a subset of the model.

We have also clearly shown that straightforward contrastive divergence training produces a protein model with both cooperative folding behavior and highly-expanded unfolded states. These correlations to experiment are notable precisely because the contrastive divergence optimizations is only concerned with native state stability, not the properties of the unfolded state or the folding transition. Instead, these properties emerge from the training spontaneously, and do not depend on an accurate solvation model. Based on these successes, it is likely that this methodology can be easily extended to treat disordered regions of proteins, possibly after augmenting the training set with examples of well-characterized disordered systems.

Despite the success of this training procedures, there are also a number of weaknesses that will be addressed in future work. In the machine learning literature, contrastive divergence is noted to obtain peak accuracy before fully converging then declining accuracy afterwards [20]. This is attributed to growing interaction strengths as training proceeds that result in larger barriers that reduce exploration of the model. With reduced exploration, the contrastive divergence trains against less interesting alternative structures. There have been many suggestions in the literature and many point to continuous simulations that do not reset to the native structure on every iteration, termed Persistent Contrastive Divergence[25]. We intend to explore such alternatives with larger computing resources to better maximize the accuracy that we can extract from the Upside model.

Future work will also attempt to address equilibration difficulties for α/β-proteins. These difficulties are typified by the underconverged simulations of protein L and ubiquitin in the benchmark simulations. More expensive training like persistent contrastive divergence may still be unable to train optimal parameters when simulations like those of protein L are uncoverged after three days of continuous simulation. The origin of the sampling difficulty may be traced to the sharp folding transition in these proteins. Temperature-based techniques for accelerating simulation convergence, such as replica exchange, are noted to have difficulties accelerating first-order phase transitions [8]. The essential reason is that few to no replicas have any significant population in the transition state energies between folded and unfolded conformations. As first-order, cooperative folding transitions are experimentally well-supported [13], we would like to preserve this emergent property of our model rather than modifying the training to somehow avoid it. Instead, we intend to pursue alternative sampling techniques, such as Wang-Landau sampling [29, 28], that are known to gracefully handle first-order phase transitions. With accelerated sampling, we may expect improve accuracy from contrastive divergence.

3.12 Simulation details

The force is integrated using Verlet integration with a time step of 0.009 time units. Temperature is maintained using a Langevin thermostat with a thermalization timescale of 0.135 time units.

The Adam optimizer is used to perform gradient descent on the objective function, using the contrastive divergence pseudo-gradient in place of the true maximum likelihood gradient. The Adam parameters used are $\beta_1 = 0.8$, $\beta_2 = 0.96$ and $\epsilon = 10^{-6}$. The α parameter is varied based on the type of term to ensure stability, $\alpha_{\mathrm{SC}} = 0.5$, $\alpha_{\mathrm{env}} = 0.1$, $\alpha_{\mathrm{HBond}} = 0.02$, and $\alpha_{\mathrm{sheet}} = 0.03$. The α parameters are multiplied by 0.25 for the fine-tuning optimization.

Regularization and derivative propagation for contrastive divergence optimization are handled using the Theano library[24].

3.13 Training data and optimization

The contrastive divergence training is conducted with 456 crystal structures from the Protein Data Bank. The initial selection of structures uses the PISCES server[30] to select proteins with X-ray resolution less than 2.2Å and pairwise sequence similarity less than 30%. In structures with multiple chains, a single chain is chosen by the PISCES server. To avoid non-globular proteins or proteins with strong interactions with other subunits in the structure, random sample consensus linear regression [7] is used to identify outliers based on the relationship between $\log N_{\text{res}}$ and $\log R_g$. Only chains with between 50 and 100 residues are used to encourage fast relaxation during the contrastive divergence simulations. All proteins homologous to proteins in the benchmark folding set are eliminated from the training set. Additionally, all proteins with backbone gaps, either missing residues due to diffuse electron density or non-standard amino acids that Upside does not handle, are also excluded from the training set.

The final training set of 456 proteins is divided into 38 groups of 12 proteins each. These groups, called minibatches, called minibatches are an essential feature of algorithms derived from stochastic gradient descent. The idea is that it is very wasteful to compute the gradient direction in equation (3.6) using all 456 proteins. Using all of the proteins would give an extremely precise estimate of the gradient that nonetheless could only be used to make a small change in the objective because of the non-linearity of the Boltzmann ensemble to changing parameters. Instead, it is better to use a few proteins at a time to obtain a inexpensive, noisy estimate of the gradient then take a small step in that direction. As long as the parameter step sizes decrease appropriately, the noise of the minibatch gradients will average out and the iteration will converge. To avoid overfitting, the minibatches are cycled through sequentially so that each minibatch is only re-used every 38 steps (a full pass through the minibatches is termed an epoch). We use the Adam optimizer[14] to accelerate convergence of the parameters.

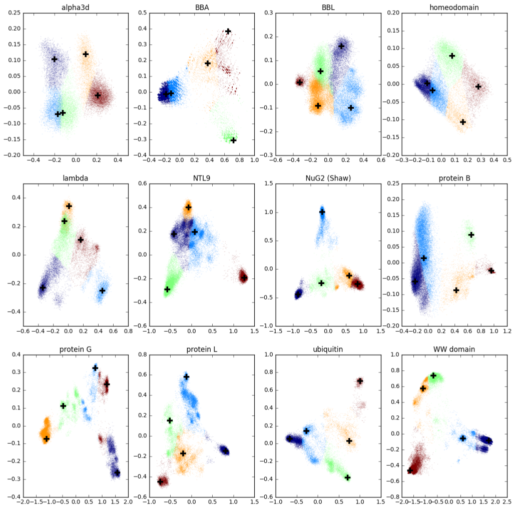


Figure 3.12: Scatter plot of the first two principal components for the lowest temperature replica of each trajectory, after discarding the first 1/3 of the simulation. The cluster of each conformation is indicated by colors, using the same color scheme as Fig. 3.5, and the cluster representative is indicated by a plus sign.

3.14 Details of test proteins

The test proteins are the same as those described in Section 2.12.

3.15 Clustering of protein structures

Figures 3.4 and 3.5 require a clustering of the low temperature structures as well the choice of a representative conformation for each cluster.

To cluster each trajectory, the conformations are mapped to their C^α contact matrices, defined so that the ij entry of the contact matrix is 1 if the position of the residue i C^α is within 8Åof the residue j C^α. The contact matrix is unrolled into a long vector, so that each conformation is represented by a 0/1-vector of length $N_{\text{res}}(N_{\text{res}+1})/2$. Principal component analysis is used to make these long vectors to five dimensions, scaled by the relative size of their variance components so that. This five-dimensional representation typically shows well-defined clusters for the trajectories we study. The resulting principal components representation of the conformations is partitioned into five clusters using the k-means++[2] implemenation in scikit-learn[17]. The choice of five clusters is arbitrary and may not be ideal to represent a natural clustering for any particular protein. The clustering is merely intended to roughly map the low temperature structures.

Representative conformations are chosen for each cluster as regions of high density in the first two principle components. For each cluster, the smoothed probability density in the first two principle components is estimated using a kernel density estimate with bandwidth given by 20% of the standard deviation in the cluster. The representative conformation for the cluster is chosen to be the conformation with the high probability density of its first two principle components according to this measure.

3.16 Derivation of contrastive divergence

We derive the contrastive divergence method as a series of approximations to the problem of best approximating the probability distribution of observed PDB structures using a force-field of an imperfect, fixed form.

We begin by assuming that we have a large collection of protein sequences $\{s_a\}$ and their associated Boltzmann distributions $p_{s_a}^{\text{true}}(X_a)$ under physiological conditions, where a represents an arbitrary label to enumerate the proteins and X_a represents the configuration of the protein (in our case, we are only interested in the backbone trace for X_a). Note that the "true" Boltzmann distribution is an unobservable idealization of the conformational ensemble of a protein under physiological conditions, and we further idealize that the true Boltzmann distribution is derived from from an extremely-complicated true potential $V_{s_a}^{\text{true}}$ by statistical mechanics,

$$p_{s_a}^{\text{true}}(X_a) = \frac{\exp(-V_{s_a}^{\text{true}}(X_a))}{\exp(-G_{s_a}^{\text{true}})} \tag{3.7}$$

$$G_{s_a}^{\text{true}} = -\log \int e^{-V_{s_a}^{\text{true}}(X)}\, dX. \tag{3.8}$$

The subscript s_a indicates that both the potential $V_{s_a}^{\mathrm{true}}$ and free energy $G_{s_a}^{\mathrm{true}}$ depend on the sequence of the protein. We may think of this as an artifact of working in the coarse-grained coordinates of the backbone trace, where the energy $V_{s_a}^{\mathrm{true}}$ really represent the free energy of the backbone coordinates after integrating away the solvent and side chain degrees of freedom. An analogous situation occurs in parameterizing all-atom molecular dynamics, where the "energy" of the system really represents the free energy of the system after integrating over the electronic degrees of freedom. Our goal is to define a parametric $V_s^{\mathrm{approx}}(X)$ that approximates the V_s^{true} for any sequence s. We drop the subscript s below where there is no chance of confusion.

For an approximate potential V^{approx}, such as the Upside model defined above, it is almost certain that V^{approx} does not have enough flexibility in its functional form to match all of the Boltzmann distributions p_a for any sequence s_a. We must instead find a V^{approx} that is "close" to V^{approx}. Defining the Boltzmann distribution of V^{approx} in the same manner as that of V^{true},

$$p_{s_a}^{\mathrm{approx}}(X_a) = \frac{\exp(-V_{s_a}^{\mathrm{approx}}(X_a))}{\exp(-G_{s_a}^{\mathrm{approx}})} \tag{3.9}$$

$$G_{s_a}^{\mathrm{approx}} = -\log \int e^{-V_{s_a}^{\mathrm{approx}}(X)}\, dX, \tag{3.10}$$

we may use the Kullback-Leibler(KL) divergence to measure the information theoretic-closeness of the associated Boltzmann distributions. This measure is defined by

$$\begin{aligned}
&\mathrm{KL}(p^{\mathrm{true}}, p^{\mathrm{approx}}) \\
&\quad = \int p^{\mathrm{true}}(X) \log \frac{p^{\mathrm{true}}(X)}{p^{\mathrm{approx}}(X)}\, dX \\
&\quad = \langle -\log p^{\mathrm{approx}}(X) + \log p^{\mathrm{true}}(X) \rangle_{\mathrm{true}} \\
&\quad = \langle (V^{\mathrm{approx}}(X) - G^{\mathrm{approx}}) - \\
&\qquad (V^{\mathrm{true}}(X) - G^{\mathrm{true}}) \rangle_{\mathrm{true}}.
\end{aligned} \tag{3.11}$$

In the last equation, we note that the KL divergence is simply the average energy difference between the true and approximate potentials (after subtracting the free energies to normalize the probabilities), where the average is taken over the true Boltzmann distribution. This makes the KL divergence highly non-symmetric between the true and approximate potentials. If the approximate potential is high (unfavorable) where the true potential is low, this will make the KL divergence much larger. Regions of configuration space where the true potential is high contribute little to the average since these regions will have low probability in the true Boltzmann ensemble. The key fact when minimizing KL divergence is that if the approximate distribution lacks the freedom to exactly match the true distribution, then the minimizing distribution will be weaker than the true distribution (i.e. less sharp) to avoid assigning highly unfavorable energy to configurations that are likely in the true distribution.

We unfortunately lack knowledge of the true energy p^{true}, so that we are unable to compute the expectation needed for the KL divergence for a concrete V^{approx}. Instead of minimizing the KL divergence, we can instead minimize

$$\langle V^{\text{approx}}(X) - G^{\text{approx}} \rangle_{\text{true}}, \tag{3.12}$$

since the remaining term $\langle V^{\text{true}}(X_a) - G^{\text{true}} \rangle_{\text{true}}$ is independent of the approximating potential. This expectation value is still intractable since we do not know p^{true}. We instead approximate

$$p^{\text{true}}(X) \approx p^{\text{empirical}}(X) = \frac{1}{M} \sum_{a=1}^{M} \delta(X - X_a), \tag{3.13}$$

where δ is the Dirac delta function and M is the number of proteins. We can approximate minimizing the KL divergence by instead minimizing

$$\begin{aligned}
&\langle V^{\text{approx}}(X) - G^{\text{approx}} \rangle_{\text{empirical}} \\
&\quad = \frac{1}{M} \sum_{a=1}^{M} \int \delta(X - X_a)(V^{\text{approx}}(X) - G^{\text{approx}})\, dX \\
&\quad = \frac{1}{M} \sum_{a=1}^{M} (V^{\text{approx}}(X_a) - G^{\text{approx}}). \qquad (3.14)
\end{aligned}$$

Minimize the expression (3.14) is exactly the method of maximum likelihood. The derivation given above illustrates two points via the connection to KL divergences. The first is what happens when the approximating energy V^{true} cannot capture the nuances of V^{approx}. In this case, the model will be overly broad so as not to assign high energy to any configurations that have low energy under V^{true}. The second salient point is that with only a finite number of samples, $p^{\text{empirical}}$ may be a poor approximation to p^{true}. Whether that is the case depends on the ability of V^{approx} to wrap itself tightly near the δ-functions associated with each sample. This is the origin of overfitting, which is especially problematic with large numbers of parameters.

We can now take the derivative with respect to an arbitrary forcefield parameter α_i in preparation to perform gradient descent to minimize (3.14). This gradient is given by

$$\begin{aligned}
&\frac{d}{d\alpha_i} \frac{1}{M} \sum_{a=1}^{M} (V^{\text{approx}}(X_a) - G^{\text{approx}}) \\
&\quad = \frac{1}{M} \sum_{a=1}^{M} \left(\frac{dV^{\text{approx}}}{d\alpha_i}(X_a) - \frac{dG^{\text{approx}}}{d\alpha_i} \right) \\
&\quad = \frac{1}{M} \sum_{a=1}^{M} \left(\frac{dV^{\text{approx}}}{d\alpha_i}(X_a) - \left\langle \frac{dV^{\text{approx}}}{d\alpha_i}(X) \right\rangle_{\text{approx}} \right), \qquad (3.15)
\end{aligned}$$

where we have used the standard statistical mechanics identity $dG/d\alpha_i = \langle dV/d\alpha_i \rangle$. While we have obtained a concrete expression for gradient descent in (3.15), we still have a major stumbling block. Computing the expectation of the derivative of the potential at X_a is straightforword given a functional form for V^{approx}, but obtaining even a reliable approximation for $\left\langle \frac{dV^{\text{approx}}}{d\alpha_i}(X) \right\rangle_{\text{approx}}$ is extraordinarily difficult. To approximate the expectation value with a finite sample, we would need to obtain Boltzmann samples from our current approximating potential. Even obtaining the single most likely configuration for our approximating potential is equivalent to finding the native state of the model, and this is very difficult for realistic pairwise potentials. Making matters more difficult, we would need to find the Boltzmann ensemble for all the proteins in our training set, and keep those Boltzmann ensembles up to date as we use gradient descent to optimize our potentials. This represents an extreme expense and it is unrealistic to obtain good converged estimates for $\left\langle \frac{dV^{\text{approx}}}{d\alpha_i}(X) \right\rangle_{\text{approx}}$ over the whole training set for anything but the simplest models of proteins. Note also that we cannot simply construct a large list of structures at some time and reweight those structures according to the potential, since the potential is constantly changing. Reweighting ensembles is only valid over very small neighborhoods of parameter space, and this procedure would depend on being able to generate an exhaustive survey of candidate structures FOOTNOTE there are contrastive divergence variants that attempt such a thing but they really work based on the driving effect of the changing potential as seen in [25].

The contrastive divergence method[10] works based on a key insight. We do not need an accurate approximation to (3.15), so long as the derivative points in direction of parameter space that improves the potential accuracy (i.e. any direction is acceptable as long as it is not uphill). The authors propose replacing Boltzmann average $\left\langle \frac{dV^{\text{approx}}}{d\alpha_i}(X) \right\rangle_{\text{approx}}$ with a finite-time Fokker-Planck average over a very short period of time *for a simulation that originates at the data point X_a*. In the Monte Carlo (MC) dynamics that the authors use, even one MC step is sufficient to produce decent optimization of the model. In our case, we replace their small number of MC steps with a short time simulation using replica exchange Langevin dynamics. As the duration of the simulation is increased, our derivative estimate will converge to the true derivative (3.15).

3.17 References

[1] Aashish N Adhikari, Karl F Freed, and Tobin R Sosnick. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. *Proceedings of the National Academy of Sciences*, 109(43):17442–17447, 2012.

[2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[3] Wouter Boomsma, Kanti V Mardia, Charles C Taylor, Jesper Ferkinghoff-Borg, Anders Krogh, and Thomas Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.

[4] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *AISTATS*, volume 10, pages 33–40. Citeseer, 2005.

[5] Guillaume Desjardins, Aaron Courville, Yoshua Bengio, Pascal Vincent, and Olivier Delalleau. Parallel tempering for training of restricted boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 145–152. MIT Press Cambridge, MA, 2010.

[6] David Duvenaud, Dougal Maclaurin, and Ryan P Adams. Early stopping as nonparametric variational inference. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1070–1077, 2016.

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[8] Jes Frellsen, Ole Winther, Zoubin Ghahramani, and Jesper Ferkinghoff-Borg. Bayesian

generalised ensemble markov chain monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 408–416, 2016.

[9] Thomas Hamelryck, Mikael Borg, Martin Paluszewski, Jonas Paulsen, Jes Frellsen, Christian Andreetta, Wouter Boomsma, Sandro Bottaro, and Jesper Ferkinghoff-Borg. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PloS one*, 5(11):e13714, 2010.

[10] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[11] Malene Ringkjøbing Jensen and Martin Blackledge. Testing the validity of ensemble descriptions of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 111(16):E1557–E1558, 2014.

[12] Abhishek K Jha, Andrés Colubri, Karl F Freed, and Tobin R Sosnick. Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13099–13104, 2005.

[13] Hüseyin Kaya and Hue Sun Chan. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Structure, Function, and Bioinformatics*, 40(4):637–661, 2000.

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[16] Adam Liwo, Matthew R Pincus, Ryszard J Wawak, Shelly Rackovsky, S Ołdziej, and Harold A Scheraga. A united-residue force field for off-lattice protein-structure simulations. ii. parameterization of short-range interactions and determination of weights of

energy terms by z-score optimization. *Journal of computational chemistry*, 18(7):874–887, 1997.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Alexei A Podtelezhnikov, Zoubin Ghahramani, and David L Wild. Learning about protein hydrogen bonding by minimizing contrastive divergence. *Proteins: Structure, Function, and Bioinformatics*, 66(3):588–599, 2007.

[19] Alexei A Podtelezhnikov and David L Wild. Inferring knowledge based potentials using contrastive divergence. In *Bayesian Methods in Structural Bioinformatics*, pages 135–155. Springer, 2012.

[20] Ruslan R Salakhutdinov. Learning in markov random fields using tempered transitions. In *Advances in neural information processing systems*, pages 1598–1606, 2009.

[21] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein science*, 15(11):2507–2524, 2006.

[22] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, page 0956797611417632, 2011.

[23] John J Skinner, Wookyung Yu, Elizabeth K Gichana, Michael C Baxa, James R Hinshaw, Karl F Freed, and Tobin R Sosnick. Benchmarking all-atom simulations using hydrogen exchange. *Proceedings of the National Academy of Sciences*, 111(45):15975–15980, 2014.

[24] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[25] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

[26] Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I Jordan, and Roland L Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput Biol*, 6(4):e1000763, 2010.

[27] Csilla Várnai, Nikolas S Burkoff, and David L Wild. Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: a maximum likelihood approach. *Journal of chemical theory and computation*, 9(12):5718–5733, 2013.

[28] Fugao Wang and DP Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):056101, 2001.

[29] Fugao Wang and DP Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050, 2001.

[30] Guoli Wang and Roland L Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[31] Bartłomiej Zaborowski, Dawid Jagieła, Cezary Czaplewski, Anna Hałabis, Agnieszka Lewandowska, Wioletta Z?mudzin?ska, Stanisław Ołdziej, Agnieszka Karczyn?ska, Christian Omieczynski, Tomasz Wirecki, et al. A maximum-likelihood approach to force-field calibration. *Journal of chemical information and modeling*, 55(9):2050–2070, 2015.

CHAPTER 4

THE UPSIDE MODEL FOR COARSE-GRAINED PROTEIN PHYSICS

4.1 Introduction

The purpose of this chapter is to give a detailed account of the Upside model for protein physics and to explain the myriad aspects of the potential energy function. Modelling choices are made both for physical realism and for the ability to stably integrate Hamilton's equations using the Verlet algorithm. Detail-averse readers may skip this chapter without compromising their understanding of the rest of the text, although they are encouraged to read subsection 4.3 for a description of interesting effect when converting between Ramachandran frequencies and Ramachandran potentials.

While this chapter does not describe the software implementation of the Upside model in detail, descriptions of the components of the Upside potential will include a reference to the class in the Upside program that implements them.

4.2 Basic backbone terms

4.2.1 Bond springs

Harmonic potentials are used to restrain the N–C$_\alpha$, C$_\alpha$–C, and C–N distances to physical values. Upside uses the standard form of a harmonic potential,

$$V_{\text{bond}}(x_1, x_2) = \frac{1}{2}k(|x_1 - x_2| - r_0)^2, \tag{4.1}$$

where the equilibrium distance r_0 is chosen to be 1.453Å, 1.526Å, and 1.300Åfor the three types of bond springs, respectively. The spring constant k is chosen to be 48. energy units / Å^2. While the bond distances are based on crystallographic data, the choice of spring constant is made in an attempt to maximize the allowable Verlet time step. Since bond vibrations are often the highest frequencies in a molecular system, the step size for molecular simulation is often chosen to avoid instability of the Verlet integrator due to instabilities in integrating the motion of bonded atoms. Hence, simulation efficiency may be increased by increasing the integrator time step after reducing the spring constant for bonded atoms. A limit on the extent to which the bond spring constant may be reduced is discussed in the next section.

Bond springs are implemented in the `DistSpring` class.

4.2.2 Angle springs

Angle springs are used to restrain the bond angles for N–C$^\alpha$–C, C$^\alpha$–C–N, and C–N–C$_\alpha$ groups. The form of the potential is

$$V_{\text{angle}}(x_1, x_2, x_3) = \frac{1}{2}k((\widehat{x_1 - x_2}) \cdot (\widehat{x_3 - x_1}) - a_0)^2. \tag{4.2}$$

Note that the dot product gives $\cos\theta$ where θ is the bond angle. The integration stability of using the more standard θ instead $\cos\theta$ to be investigated, as angular springs can cause integrator instability when the spring constant k is high (in Upside, k is chosen to be 175.). Note that this angular term limits the allowable spring constant for the distance springs at a fixed integration time step. When one of the flanking atoms in an angular spring approaches the central atom, the force due to the angular potential increases proportional to $1/r$. Since these high forces are likely to push the system outside the stability conditions of the Verlet integrator, a sufficiently high spring constant must be used for the distance restraints to avoid near approach of atoms participating in an angular interaction.

The equilibrium value a_0 is given by

$$a_0^{\text{N–C}^\alpha\text{–C}} = \cos 109.5^\circ \tag{4.3}$$

$$a_0^{\text{C}^\alpha\text{–C–N}} = \cos 120^\circ \tag{4.4}$$

$$a_0^{\text{C–N–C}_\alpha} = \cos 120^\circ. \tag{4.5}$$

Angle springs are implemented in the `AngleSpring` class.

4.2.3 Dihedral springs

Dihedral springs are used to restrain the ω dihedral angles. They use the functional form

$$V_{\text{dihedral}}(x_1, x_2, x_3, x_4) = \frac{1}{2}k(\text{minimage}(\text{dihedral}(x_1, x_2, x_3, x_4) - \theta_0))^2, \tag{4.6}$$

where minimage(ϕ) adds a multiple of 2π to ϕ to give an angle in the range $[-\pi, \pi]$. Simulations employ only fixed *cis* or *trans* angles without interconversion since the derivative of this potential is discontinuous at the antipodal point to the equilibrium value. While the fixed-ω constraint is limiting for describing the Boltzmann ensemble of unfolded proteins, it is natural for describing kinetics as *cis*/*trans* interconversion takes place over much longer time scales than most small-protein folding. The dihedral springs could be modified to have a continuous derivative using the potential

$$V_{\text{dihedral}}^{\text{modified}}(x_1, x_2, x_3, x_4) = k^{\text{modified}}(1 - \cos(\text{dihedral}(x_1, x_2, x_3, x_4) - \theta_0)), \tag{4.7}$$

in analogy with the von Mises distribution on the circle. A pair of these potentials could be used to implement *cis*/*trans* interconversion, but this would require changes to the Ramachandran potentials which distinguish *cis*- and *trans*-proline.

The equilibrium θ_0 is chosen to 0 or π as appropriate. The spring constant k is chosen to be 30. Note that the dihedral springs place constraints on the spring constants of the angular springs. If either of the internal angles of the dihedral group approach π (collinear atoms), the derivative of the dihedral potential varies rapidly as the dihedral angle is undefined for collinear atoms. The angle potential spring constants must be sufficiently large to avoid these integration instabilities.

Dihedral springs are implement in the `DihedralSpring` class.

4.3 Ramachandran angle potential

There are a variety of physical effects and statistical correlations in protein structures that can be represented using a Ramachandran potential, which requires careful care to capture only those effects that at not subsequently captured by the side chain potential.

The conceptually simplest method to construct to construct a Ramachandran potential is to first create a statistical model of the backbone conformations of protein in the Protein Databank, such as the TorusDBN[2] model. By taking the negative logarithm of such a model, one may obtain a potential energy surface for sampling from the probability model. While such a model is conceptually straightforward to construct, it is fraught with problems for interpreting the physicality of such a potential. Separately parameterizing the Ramachandran potential forces all other potential terms to model only the terms statistically independent to the potential described by the Ramachandran angles. Such a division is simple for terms like the bond springs between backbone atoms. Physically, there is little important correlation to be expected between bond fluctuations and Ramachandran propensities. The situation for side chain and hydrogen bonding terms is quite a bit more complicated. Consider determining the helical fraction for a short peptide sequence in solution. A sufficiently powerful, statistical Ramachandran probability model may capture the helicity of this peptide correctly (i.e. the helical probability of the peptide matches that of experiment). While this would be a success of probability modeling for Ramachandran potential, it is a challenge for the hydrogen bond and side chain potentials of the model. Since the helicity is correct using only the Ramachandran potential, the hydrogen bond potential and side chain potential must contribute nothing in aggregate to the helical probability of the peptide. This is obviously unphysical, as surely the hydrophobicity of the residues or the strength of hydrogen bond interactions are largely responsible for determining the helicity of the peptide. Indeed, the effects have been moved out of the physical side chain and hydrogen bonding interactions to be put in the Ramachandran potential. While this may work for an isolated peptide with no tertiary structure, it is extremely hard to make such a model transferable to larger systems, where the hydrophobic effects that must be omitted for the peptide may be driving the formation of tertiary structure. This problem is quite general to statistical potentials. Given the dihedral angles of protein, it is possible to recover the atomic positions of the backbone to high accuracy, especially if care is taken to avoid steric clash. For this reason, a sufficiently sophisticated statistical Ramachandran potential is able to describe arbitrary protein physics, yet the stated purpose of Ramachandran potentials is only to capture sequence local interactions.

The above argument shows that it is imperative to limit the capacity of the Ramachandran potential to avoid capturing protein physics that should be described in the hydrogen bonding and side chain interactions. Yet, it is not clear that the appropriate course of action is to use similar dihedral potentials to those used in all-atom molecular dynamics. These potentials are parameterized from quantum calculations of the energies of backbone torsions and depend heavily on the side chain interactions to set local conformational preferences. Such direct energetic models put a great strain on the side chain interactions to describe the secondary structure, such as the secondary structure preferences of β-branched residues. While this is appropriate for an all-atom model, it may be difficult to describe these conformational preferences in a coarse-grained model. Unfortunately, we must anticipate which effects are likely to be described by the remaining terms of the Upside potential when we choose the form of the Ramachandran potential.

We use a simple form for the Ramachandran potential

$$V_{\mathrm{Rama}}(\phi_1, \psi_1, \ldots, \phi_N, \psi_N) = \sum_i V_{\mathrm{aa}_{i-1},\mathrm{aa}_i,\mathrm{aa}_{i+1}}(\phi_i, \psi_i). \tag{4.8}$$

This dynamically-independent residue model for the Ramachandran angles is unable to model long-range correlations that are needed to describe secondary structure formation, so that the hydrogen bond and side chain interactions are responsible for much of the driving force of secondary structure formation. It still influences secondary structure preference through angular probabilities. The construction of the $V_{\mathrm{aa}_{i-1},\mathrm{aa}_i,\mathrm{aa}_{i+1}}$ potentials is a bit involved, however. The starting point for Upside's Ramachandran model is the NDRD Ramachandran libraries [7]. These libraries model the Ramachandran probabilites of a residue given the identity of the residue and the identity of its left *or* right neighbor (i.e. dimer maps), which are modeled using a hierarchical Dirichlet mixture. The Dirichlet mixture implies that the probability density is a (possibly-infinite) sum of Gaussian-like functions, which allows the probability model to be continuous and differentiable for all inputs. The model is hierarchical in the sense that it uses a hierarchical Bayesian prior so that the prior distribution for a dimer is influenced self-consistently by the one-body distributions of each residue type.

One of the most important modelling decisions for the Ramachandran interaction is the choice of what residues in the PDB to use for our statistical model. As we explained above, it makes little sense to use all of the PDB residues, since we hope only to model the interactions that are missed by our side chain model. A popular choice is a so-called "coil library", that is to model only those residues which have coil structure according to the DSSP classification [4]. This choice is only one among many, and it is not clear that it strikes the optimal balance needed to capture only the protein physics not treated elsewhere in the model. We use a slightly more parameteric approach. Our starting point is a library of residues in the turn, coil, and bridge DSSP states (TCB states), which is modeled in [7]. This library separately provides $f^{\mathrm{L}}_{\mathrm{aa}_1,\mathrm{aa}_2}(\phi_2,\psi_2)$, which is the smoothed Ramachandran frequency for residues of type aa_2 given that there is a residue of type aa_1 *preceding* it in sequence, and $f^{\mathrm{R}}_{\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2)$, which is the smoothed Ramachandran frequency for residues of type aa_2 given that there is a residue of type aa_1 *following* it in sequence. To combine these to two frequencies into an approximation of the triplet frequency, we use a simple mixture

$$f^{\mathrm{mixture}}_{\mathrm{aa}_1,\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2) = \frac{f^{\mathrm{L}}_{\mathrm{aa}_1,\mathrm{aa}_2}(\phi_2,\psi_2) + f^{\mathrm{R}}_{\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2)}{2}. \tag{4.9}$$

This mixture is more diffuse than that we would obtain using the stricter ratio rule recommended in [7],

$$f^{\mathrm{mixture}}_{\mathrm{aa}_1,\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2) \propto \frac{f^{\mathrm{L}}_{\mathrm{aa}_1,\mathrm{aa}_2}(\phi_2,\psi_2) * f^{\mathrm{R}}_{\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2)}{f^{\text{1-residue}}_{\mathrm{aa}_2}(\phi_2,\psi_2)}, \tag{4.10}$$

where the 1-residue frequency is used to avoid double-counting effects that depend only one residue. While either rule is reasonable, we have noted better results from the less peaky distributions obtained with the mixture rule.

Since these Ramachandran models empirically tends to create very helical Upside models, we add an additional parameter λ to increase the proportion of β-sheet angles in the model. To do so, we have fit a simple gaussian kernel density estimate for angles classifed as sheet in DSSP, as well as the fraction $f_{\mathrm{aa}_1,\mathrm{aa}_2}$ of residues of each dimer type that are in β-sheets in a representative sample of the PDB. We combine the TCB Ramachandran frequencies and the sheet frequencies according to weighted mixture,

$$f^{\mathrm{final}}_{\mathrm{aa}_1,\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2;\lambda) = \frac{(1-g)f^{\mathrm{TCB}}_{\mathrm{aa}_1,\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2) + g * e^{-\lambda} f^{\mathrm{sheet}}_{\mathrm{aa}_1,\mathrm{aa}_2,\mathrm{aa}_3}(\phi_2,\psi_2)}{(1-g) + g * e^{-\lambda}}, \tag{4.11}$$

where g is the average of the DSSP sheet frequencies of the dimers $(\mathrm{aa}_1,\mathrm{aa}_2)$ and $(\mathrm{aa}_2,\mathrm{aa}_3)$. The purpose of this combining rule is to only add sheet frequencies for residues observed to be in the sheet conformation in the PDB. The parameter λ gives the contrastive divergence

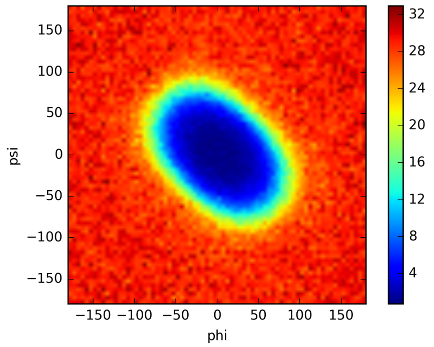


Figure 4.1: Probability density of Ramachandran angles for the central residue of a free alanine chain with no Ramachandran potential for the central residue (colors scale gives the probability density values in units of 1/degrees^2). The chain experience only short-range excluded volume interactions, and interactions between adjacent residues are excluded from the potential. The depression near $\phi = \psi = 0$ occurs due to a high probability of $i-1, i+1$ steric overlap when the i-th residue is near the origin of ϕ/ψ space.

optimization a way to tune the amount of added sheet bias to compensate partially for the effect of other terms in the model. To obtain the final Ramachandran energy term, we perform a naive Boltzmann inversion (with a small modification below) to give

$$V_{\mathrm{rama}}(\phi_1, \psi_1, \ldots, \phi_N, \psi_N) = -\sum_i \log f^{\mathrm{final}}_{\mathrm{aa}_{i-1}, \mathrm{aa}_i, \mathrm{aa}_{i+1}}(\phi_i, \psi_i; \lambda), \tag{4.12}$$

where λ is a single tuneable parameter.

The Ramachandran coordinates ϕ and ψ are implemented in the `RamaCoord`, and the Ramachandran potential is implemented in the `RamaMapPot`.

4.3.1 Reference state correction for Ramachandran modeling

The potential V in (4.12) is sufficient to ensure that the Ramachandran frequencies of the Boltzmann distribution are equal to the f^{final} functions in the case that the only other terms are the bond, angle, and dihedral springs given above. The situation is more complicated when even short-range steric interactions are present to give proper atomic radii to the N, C_α, and C atoms of the backbone, interactions which are described in more detail below.

The bias of Ramachandran angles for a small peptide with zero Ramachandran potential is depicted in Fig. 4.1, which shows a preference against angles near $\phi = \psi = 0$. The edge of the bias region is sufficient to significantly alter the helical probability of proteins. The bias depends only weakly on the Ramachandran potential of the flanking residues. For this reason, we correct the bias by added a reference state potential

$$V_{\text{rama}}^{\text{ref}} = \sum_i \log f^{\text{ref}}(\phi_i, \psi_i). \tag{4.13}$$

Because the dependence on the Ramachandran distribution of the flanking residues is weak, there is no need to iteratively correct the bias.

4.4 Hydrogen bonding potential

Since the configuration of the protein chain is represented only by N, C_α, and C, we lack the amide hydrogen (H) and carbonyl oxygen (O) atoms necessary to represent the physics of hydrogen bonding. We remedy this difficiency by noting that the position of these atoms can be reconstructed to high accuracy from the position of the backbone atoms. To find the position of H and O, we compute

$$\begin{aligned}
x_{\text{H}} &= x_{\text{N}} + (0.88\text{Å})d_{\text{H}} && (4.14)\\
x_{\text{O}} &= x_{\text{C}} + (1.24\text{Å})d_{\text{O}} && (4.15)\\
d_{\text{H}} &= \text{sp3}(x_{\text{C}}, x_{\text{N}}, x_{\text{C}_\alpha}) && (4.16)\\
d_{\text{O}} &= \text{sp3}(x_{\text{C}_\alpha}, x_{\text{C}}, x_{\text{N}}) && (4.17)\\
\text{sp3}(x_1, x_2, x_3) &= \text{unitvec}\left(\text{unitvec}(x_2 - x_1) + \text{unitvec}(x_2 - x_3)\right) && (4.18)\\
\text{unitvec}(x) &= \frac{x}{|x|}. && (4.19)
\end{aligned}$$

The coordinate computation is implemented in the `Infer_H_O` class, and the resulting coordinates are used heavily in subsequent Upside potential calculations. The directions d_{H} and

d_{O} are the N–H and C–O bond vectors respectively, which are used to assess angular components of both hydrogen bonding and side chain interactions. In Upside, the N-terminal and proline residues have no amide hydrogen, while the C-terminal residue has no carbonyl oxygen. No capping is currently implemented in Upside.

Given the position and bond vectors for the hydrogen and oxygen atoms, a hydrogen bond score for each possible hydrogen bond is computed according to

$$\begin{aligned}
h_{ij} &= \mathrm{radial}(|x_i^{\mathrm{H}} - x_j^{\mathrm{O}}|)\,\mathrm{angular}(d_{ij}^{\mathrm{HO}} \cdot d_i^{\mathrm{O}})\,\mathrm{angular}(-d_i^{\mathrm{H}} \cdot d_{ij}^{\mathrm{HO}}) && (4.20)\\
d_{ij}^{\mathrm{HO}} &= \mathrm{unitvec}(x_i^{\mathrm{H}} - x_j^{\mathrm{O}}) && (4.21)\\
\mathrm{radial}(r) &= \mathrm{sigmoid}\left(\frac{r - (1.4\text{Å})}{0.10\text{Å}}\right)\mathrm{sigmoid}\left(\frac{(2.5\text{Å}) - r}{0.125\text{Å}}\right) && (4.22)\\
\mathrm{angular}(a) &= \mathrm{sigmoid}\left(\frac{a - 0.682}{0.05}\right). && (4.23)
\end{aligned}$$

The hydrogen bonding score h_{ij} is non-symmetric in i and j and quantifies the extent to which residue i and residue j form a hydrogen bond as the donor and acceptor, respectively. The hydrogen bond score is in the range $[0, 1]$, where the hydrogen bond score is one approximately when $1.4\text{Å} < r < 2.5\text{Å}$, $\theta_{\mathrm{HOC}} < 47°$, and $\theta_{\mathrm{OHN}} < 47°$.

Given the hydrogen bond score of each possible hydrogen bond, a hydrogen bond score in the range $[0, 1]$ is assigned to each hydrogen or oxygen atom. This score can be interpreted as the confidence that the atom is participating in some hydrogen bond. We compute the atomic hydrogen bond score as

$$\begin{aligned}
s_i^{\mathrm{H}} &= 1 - \prod_{j \neq i}(1 - h_{ij}) && (4.24)\\
s_j^{\mathrm{O}} &= 1 - \prod_{i \neq j}(1 - h_{ij}). && (4.25)
\end{aligned}$$

The asymmetry between the definitions of s_i^{H} and s_j^{O} reflects the asymmetry in the definition of h_{ij}. Interpreting h_{ij} as a probability of the event that residues i donates a hydrogen bond to residue j and assuming that all such events are independent, (4.24) represents the probability that residue i is a hydrogen bond donor for *any* other residue. This score is constrained to the range $[0, 1]$ and the sum of these scores provides a differentiable count of the number of hydrogen bonds.

The Upside energy for backbone-backbone hydrogen bonding is given by

$$V_{\text{hbond}} = E_{\text{hbond}} \sum_i (s_i^{\text{H}} + s_j^{\text{O}}), \tag{4.26}$$

where E_{hbond} is a scalar quantity that approximately represents half of the energy of forming a single hydrogen bond. The hydrogen bond scores are further used in the side chain potential energy functions.

The hydrogen bond score is implemented in the `ProteinHBond` class and the backbone-backbone hydrogen bond energy is implemented in the `HBondEnergy` class.

4.5 Placement of candidate side chain sites

4.5.1 Residue position and orientation for side chain placement

For specified χ-angles, the side chain atomic positions are fixed relative to the position and orientation of N, C_α, and C atoms. We want to define the position and orientation of our interaction sites in a side chain reference frame with N, C_α, and C in defined positions, then apply rigid body rotation and translation to move the side chain interaction sites to their appropriate positions for a given backbone structure.

In Upside, we define the reference backbone positions (in Å) to be

$$x_{\text{N}}^{\text{ref}} = (-1.21, -0.26, \quad 0.) \tag{4.27}$$

$$x_{\text{C}_\alpha}^{\text{ref}} = (-0.02, \quad 0.56, \quad 0.) \tag{4.28}$$

$$x_{\text{C}}^{\text{ref}} = (\quad 1.23, -0.30, \quad 0.). \tag{4.29}$$

The transformation between the reference frame backbone and the current configuration of residue i is given by a rotation matrix U_i and a translation i, which are chosen to minimize the RMSD,

$$U_i, t_i = \arg\min_{U,t} \frac{1}{3} \sum_{a \in (\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C})} |x_a - (U x_a^{\mathrm{ref}} + t)|^2. \tag{4.30}$$

The tranformation U_i and t_i are computed independently for each residue i using the quaternion algorithm of [5]. The computation of the alignment is handled by the `AffineAlignment` class.

Given data (scalars, vectors, or points) in the reference frame of residue i, we compute equivalent data in the simulation frame of the residue using the usual rules for covariant transformation of spatial data. Scalar values, such as the probabilities of individual χ rotamer states, are invariant under transformation. Vectors, such as bond vectors, are rotated by U_i. Points, quanties which define a position in space, are transformed by both U_i and i. Summarizing,

$$s_i = s_i^{\mathrm{ref}} \qquad \text{scalar transformation} \tag{4.31}$$

$$d_i = U_i d_i^{\mathrm{ref}} \qquad \text{vector transformation} \tag{4.32}$$

$$x_i = U_i x_i^{\mathrm{ref}} + t_i \qquad \text{point transformation.} \tag{4.33}$$

The placement of (optionally Ramachandran-dependent) data in simulation frame is handled by the `PlacementNode` class.

The current definition of the Upside potential uses N_i, C_i^α, and C_i as its configuration space with U_i and t_i as computed values, but this may not be the most efficient choice for simulation. Very little would change in the model if U_i (likely in quaternion form) and t_i were considered the dynamical variables and N_i, C_i^α, and C_i were computed values. Indeed, using the rigid body coordinates as the dynamical variables may allow larger and more efficient integrator steps. While using rigid body coordinates as dynamical variables has not yet been pursued due to the added technical challenge of rigid body integration and thermostatting, it is a promising area to speed simulation. If handled properly, it is likely that the Boltzmann ensemble of the protein would be hardly affected for quantities of interest.

4.5.2 Backbone sites

Backbone steric interactions are described using an ordinary pair interaction,

$$V_{\text{steric}} = \sum_{i+1<j} \sum_{a\in(\mathrm{N},\mathrm{C}_\alpha,\mathrm{C},\mathrm{C}_\beta)} \sum_{b\in(\mathrm{N},\mathrm{C}_\alpha,\mathrm{C},\mathrm{C}_\beta)} 4\,\mathrm{sigmoid}\left(\frac{|x_{ia}-x_{jb}|^2-(3\text{Å})^2}{(3\text{Å})(0.1\text{Å})}\right), \tag{4.34}$$

where x_{ia} is the position of the atom a of residue i. The position of atoms are actually placed using the rigid body placement described above (necessary for C_β but also true for the other atoms).

Backbone sterics are implemented in the `BackbonePairs` class.

4.6 Interactions that contribute to side chain placement

As described in the previous chapters, the rotamer states of the side chains are determined by self-consistent belief propagation on single and pair interactions. Below, we describe the individual interactions that combine to determine the interaction energies of the side chain states. The subscript χ will be used to indicate the rotamer state for each iteraction residues, so that x_{χ_i} represents the position of residue i corresponding to rotamer χ_i.

This section will avoid discussing in detail energy functions describe in previous chapters.

4.6.1 Rotamer prior probabilities

The Ramachandran-dependent rotamer prior probabilities are adapted from [6] by summing the probabilities of fine-grained states within each coarse-grained state used in Upside. The implementation as a scalar (non-rotated) residue potential is straightforward, except to note that the probabilities are converted to 1-residue energies using a negative logarithm transform

$$V_{\text{prior}}(\{\chi_i\}) = -\sum_i \log p_{\chi_i}(\phi_i, \psi_i). \tag{4.35}$$

The purpose of this energy is to account for within-residue side chain-backbone interactions that may be more difficult for Upside to model using only pair interactions. For example, β-branched amino acids have strong χ preferences due to the need accommodate the backbone.

This potential is implemented using a scalar component of a `PlacementNode` class. The same placement class handles placing the location and direction for each of the side chain beads.

4.7 Side chain-backbone interactions

There are two terms for side chain-backbone interactions, both of which uses the pair interaction form defined in section 2.5.

The first term controls the interaction with the hydrogen and oxygen atoms with the side chains. The position and direction of these atoms are given in section 2.5. The interaction is modulated by $\kappa_i = (1 - s_i)^2$, where the s_i is given by (4.24) or (4.25) as appropriate, so that the side chain interaction only occurs for atoms not already participating in a backbone-backbone interaction. This interaction is implemented in the `HBondCoverage` class.

The second term allows the interaction of side chain atoms with three sites (position and direction) on the protein backbone. The form of the interaction is identical to the hydrogen and oxygen interaction, except $\kappa_i = 1$ so that there is no modulation of this interaction. The interaction sites are initialized at the N, C_α, and C locations but move somewhat as they are optimized in the side chain maximum likelihood training. This interaction is also implemented by the `HBondCoverage` class but with different parameters than above.

4.8 Side chain-side chain interactions

The side chain-side chain interactions also the follow the form of (2.24) and are not modulated. The side chain-side chain interactions are implemented in the `RotamerSidechain` class. Though the capability is currently unused, the model supports having multiple beads for each side chain. This may improve the modeling of residues such as lysine or threonine that have a mixture of polar and hydrophobic interactions.

4.9 Computing the side chain energy

The belief propagation algorithm for computing the side chain free energy is detailed in 2.11.

An important detail is that the belief propagation is not a convex optimization, and so the stationary point to which the algorithm converges may be dependent on the initialization of the algorithm. We initialize the algorithm to the probabilities implied by the 1-rotamer energies, including the prior frequency energy and all the side chain-backbone interactions. Note that the latter interactions are *1-rotamer* interactions despite representing the interactions of two beads, since each involves only a single side chain. The strong accuracy of the prior probability of each rotamer at predicting crystallographic rotamer states [6] suggests that initializing the rotamer probabilities to the 1-body probabilities has a high likelihood of ensuring convergence to the global minimum of the approximate side chain free energy.

The theoretical soundness of the belief propagation free energy would be enhanced by monitoring the belief propagation to identify if the optimization converges to different basins on successive steps of the Verlet integration of the backbone. Such basin hopping will be noted in an anomalously large change in the free energy resulting from a discrete change in basin to which the algorithm converged. There is no guarantee that the old and new basins have equal free energies; rather they may be artifact of sliding down one hill or another by chance. If a large upward jump in energy were noted, presumably it could be treated with Monte Carlo accept or reject step to some accuracy (possibly by embedding the algorithm in a hybrid Monte Carlo framework[3]). As such jumps are expected to be rare, though a careful study has not been performed, this potential source of error is currently uncontrolled.

The `solve_for_marginals` method in the `RotamerSidechain` class computes the belief propagation solution for the side chain ensemble.

4.10 Environment Interaction

The Upside model as described above has no explicit solvation energy, only those effects implied by the attraction of hydrophobic side chains to the backbone and other hydrophobic side chains. Most approximate solvation interactions are designed for atomic simulation and are inappropriately expensive for Upside. Furthermore, many models need to be modified to deal with the reduced level of detail inherent in Upside's representation of the protein. Instead, the solvation model is quite similar to the simple model presented in [1], except that the solvation potential is modified to be differentiable and to be optimized by contrastive divergence. We call the term an environment interaction because it reflects the energy of the typical environment of a given residue type, with the intention that this reflects a reasonable energy for solvation.

4.10.1 Counting surrounding residues

We define a count of surrounding residues in a similar manner to [1] but modify the construction so that it is differentiable.

The main component of the environment interaction is to count the number of side chains beads within a fixed radius of the C_β atom in a hemisphere above the atom. We define

$$b_i = \sum_j p_{\chi_j}(\phi_j, \psi_j)\,\mathrm{sigmoid}\left(\frac{|x_{ji}| - (8\text{Å})}{1\text{Å}}\right)\mathrm{sigmoid}\left(\frac{d_i^{\mathrm{C}_\beta} \cdot \hat{x}_{ji} + 0.1}{1.0}\right) \tag{4.36}$$

$$x_{ji} = x_j^{\mathrm{SC}} - x_i^{\mathrm{C}_\beta}, \tag{4.37}$$

where $x_i^{\mathrm{C}_\beta}$ is the position of the C_β on residue i and $d_i^{\mathrm{C}_\beta}$ is the corresponding C_α–C_β bond vector. The sigmoid parameters are chosen to approximately maximize distinctiveness of the b_i burial distributions for different residue types. In a well-formed definition of burial, the burial distribution for a hydrophobic residue like valine should be very distinct from the burial distribution of a charged residue like aspartic acid.

The side chain probability $p_{\chi_j}(\phi_j, \psi_j)$ appearing in the definition (4.36) is the prior probability of the side chain bead, not the marginal probability from belief propagation. The derivative of the marginal probabilities with respect to the side chain positions are complex, much moreso than the derivative of the free energy with respect to the side chain positions. Furthermore, for intellectual self-consistency, the side chain rotamers ensemble should account for burial interactions. Unfortunately, belief propagation is not defined for many-body interactions, such as would arise from a nonlinear function of the burial b_i. It would be possible to extend belief propagation to handle many-body terms at a lower level of approximation, although it is not clear how accurate such an approximation would be and whether it would be guaranteed to converge. Due to the derivative difficulties of using the marginal probabilities without perturbing for burial and the intellectual difficulties of extending belief propagation to handle many-body terms, we simply avoid the issues by using only the prior (in the sense of equation (4.35), not the values of the previous step) probabilities to define the environment interactions. The precise probabilities used to define burial are unlikely to compromise the ability of the b_i to distinguish hydrophobic and hydrophillic residues.

The burial calculation is implemented in the `EnvironmentCoverage` class.

4.10.2 Computing the non-linear energy

Given the burial values b_i, the environment energy is an arbitray smooth function of the burials, parametrized as a natural cubic spline. The total burial energy is given by

$$V_{\text{burial}} = \sum_i v_{\text{aa}_i}(b_i), \tag{4.38}$$

where the v_{aa_i} function depends only on the residue type aa_i of residue i. This function should be considered as a method to correct the deficiencies in the model of the hydrophobic effect implied by the pairwise side chain energies. As a solvation model, it is very crude but the limits of the model are partially compensated by the co-training of the environment energy with the side chain energies, so that at least the energy terms are co-adapted.

The environment energy is implemented in the `NonlinearCoupling` class.

4.11 References

[1] Aashish N Adhikari, Karl F Freed, and Tobin R Sosnick. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. *Proceedings of the National Academy of Sciences*, 109(43):17442–17447, 2012.

[2] Wouter Boomsma, Kanti V Mardia, Charles C Taylor, Jesper Ferkinghoff-Borg, Anders Krogh, and Thomas Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.

[3] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[4] Abhishek K Jha, Andres Colubri, Muhammad H Zaman, Shohei Koide, Tobin R Sosnick, and Karl F Freed. Helix, sheet, and polyproline ii frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28):9691–9702, 2005.

[5] Pu Liu, Dimitris K Agrafiotis, and Douglas L Theobald. Fast determination of the optimal rotational matrix for macromolecular superpositions. *Journal of computational chemistry*, 31(7):1561–1563, 2010.

[6] Maxim V Shapovalov and Roland L Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

[7] Daniel Ting, Guoli Wang, Maxim Shapovalov, Rajib Mitra, Michael I Jordan, and Roland L Dunbrack Jr. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. *PLoS Comput Biol*, 6(4):e1000763, 2010.

CHAPTER 5

FUTURE DIRECTIONS

The probabilistic view of protein coarse-graining is most vibrant and important when we take seriously that the potential energy encodes a Boltzmann distribution, and work directly to understand and optimize the full ensemble. Given the growing scale of computational resources as well as large experimental data sets, we expect the methods espoused in thesis to increase in importance. Furthermore, the rigorous coarse-graining typified in our side chain interaction methods allow us to cross the boundary between the sharp detail of individual interactions and smoother landscape of slow-degrees of freedom. Such methods will allow the field to simulataneously attack the sampling and parameterization challenges of protein simulation. In the future, we expect a synergy between computational speed and simulation accuracy to supplant their current adversarial relationship precisely due to the ability to rigorously coarse-grain simulations.

This thesis has demonstrated a path forward to use statistical information for rigorous coarse-grained simulation of protein physics. We expect that the Upside model and associated software will be a strong starting point both for *de novo* prediction of structure and dynamics as well as an integration point to meld experimental observables (NMR chemical shifts, evolutionary features, mutational data, etc.) with a strong and sensible prior on protein structure. We further expect that the side chain free energy approximations that we have developed will expand to a rich vein of methods which blur the distinction between included and omitted degrees of freedom in coarse-grained methods. Finally, as the total throughput of molecular simulation increases, we expect that the parameterization of even all-atom molecular dynamics will come to represent a more equal hybrid of quantum and statistical information, and that experiences in pure statistical models like Upside will greatly inform practitioners about the strengths and limitations of statistical training to inform protein dynamics.

5.1 Future improvements to the Upside model and training

5.1.1 Enhanced equilibration for simulations and training

The ability to both train and run the Upside model is limited by the time required to equilibrate a simulation trajectory. While the Upside model is designed for fast equilibration, it may still take days or longer to equilibrate the Boltzmann ensemble for some proteins, even those with fewer than 100 amino acids. Currently, the only enhanced sampling techniques used in Upside are replica exchange and pivot moves. Replica exchange, however, is ineffective at accelerating the sampling of first-order phase transitions[2], and most small protein folding is approximately a first-order phase transition[3]. Instead of temperature-enhanced sampling, energy-enhanced sampling using the Wang-Landau[6] and other methods should provide faster decorrelation for the Upside model.

Currently, Upside makes no use of multi-core parallelism outside of replica exchange. While this is reasonable for very small proteins where the overhead of parallelism can be significant, there are many potential applications of Upside to larger systems for both protein association and conformational change that are greatly limited by the single-threaded nature of Upside. The Upside model is amenable to parallelization. Most of the computational cost occurs in evaluating the spline interactions necessary for the side chain interactions, a trivially-parallelizable task. Future work will address parallelism in the Upside model.

Finally, persistent contrastive divergence[5] may be used to enhance the decorrelation of structures during the contrastive divergence training. In this scheme, the individual contrastive divergence trajectories are not reset on each pass through the training set, and new proteins simulations are initialized from the last structure of their previous simulations. This allows much greater exploration of the energy landscape. As long as the parameters do not change too quickly, as controlled by the decorrelation time of the protein simulations, this method can converge to the true maximum likelihood potential energy rather than the approximately fluctuation-optimal potential energy that terminates contrastive divergence optimization.

5.1.2 Improved treatment of hydrogen bonding

Upside assigns a fixed enthalpy to the formation of a backbone-backbone hydrogen bond, a term which assigns zero enthalpy to all other states. Hydrogen bond partners in a protein are conceptually in one of three states – protein-protein hydrogen bonded, protein-solvent hydrogen bonding, or desolvated in a hydrophobic environment. Desolvation of the hydrogen bond donor or acceptor is expected to be energetically costly, while the relative energy of protein-protein and protein-solvent hydrogen bonding is far less clear. Since Upside has only a single energy associated to backbone-backbone hydrogen bonding, this energy must both prevent donor or acceptor desolvation and describe the energetic balance between solvent hydrogen bonding to backbone hydrogen bonding. Adding an additional desolvation penalty should allow a much smaller energy difference between protein-protein and protein-solvent hydrogen bonding without risking an unphysical collapse that produces large numbers of desolvated hydrogen bonds.

5.1.3 Understanding the important physics of protein folding

Coarse-grained modeling is often seen as a way to extract and understand the essential elements of protein structure and dynamics, but the success or failure of a single coarse-grained model often provides only cryptic evidence for the importance of the interactions that it describes. Upside can successfully fold many proteins without the solvation-like environment term. It is unreasonable to conclude that the partial success of a no-solvent model proves that solvation is unimportant to protein folding. Instead, it just shows the ability of side chain pair interactions to mask the absence of the solvation terms. Only by constraining the model to reproduce a large number of experimental observables, such as side chain packing and unfolded-state physics in addition to structure prediction, may we start to understand the truly essential elements of protein folding. The ability of the Upside model to make a wide range of experimental predictions, both dynamic and ensemble-averaged, will allow us to explore the important physics of protein folding using a variety of terms that include or exclude different types of protein physics *while re-training to achieve the optimal parameters for each model.* By systematic experimentation, we should be able to better characterize the essential elements of protein folding.

5.2 Applications of the Upside model

Since *Upside* is a general method for protein simulations, there are a wide variety of potential future applications. We highlight a few applications that are either novel or uniquely suited to Upside's strengths.

5.2.1 Prediction of native state fluctuations and hydrogen exchange

Since the contrastive divergence training of Upside inherently samples the near-native states of proteins, we expect the model to be especially accurate at describing the subglobal fluctuations of the native state commonly found through hydrogen exchange. Using replica exchange or even simple constant temperature simulation, we can map the local unfoldings and other conformational changes of proteins to compare our ensemble to the predictions of hydrogen exchange. Such a comparison should have the dual benefits of enhancing our understanding of the Upside model as well as helping to interpret the results of experiments with local unfoldings. This would will likely require understanding the effects of denaturant in the Upside model, a subject of future research.

5.2.2 Prediction of protein-protein binding

Upside's high accuracy and computational speed make it a natural candidate to explore protein-protein association, especially in cases where flexibility is needed to form the binding interface. While the interaction parameters derived from contrastive divergence on single-domain proteins should be fairly transferable to protein-protein association, it is not clear that these parameters would be sufficiently accurate to make a state of the art prediction method for protein association. Instead, we may use crystal structures of protein complexes to fine-tune the interaction parameters through contrastive divergence. This work is ongoing with Nabil Faruk and early results are promising.

5.2.3 Co-prediction of structure for homologous sequences

As a result of extensive genetic sequencing, it is common to have a collection of homologous protein sequences without knowing the structure of any sequence in the collection. Furthermore, if the sequences are sufficiently similar, the sequences likely have similar structures. We would like to have a method to leverage the ensemble of homologous sequences to enhance the accuracy of structure prediction for each member of the collection.

Effective methods have been developed by other researchers[4] in the case that hundreds to thousands of homologous sequences are available. In that case, residues typically in contact in the protein structures will have correlated residue types, reflecting the need to maintain relationships such as having opposite charge or occupying a given volume. When there are sufficiently many structures to resolve these statistical correlations, residue contacts can be predicted with good accuracy[7]. Typically, these statistical methods make little use of the properties of protein structures; in many techniques, the contact predictions are even invariant to random shuffling of the residue order.

In the case where only a few homologous sequences are available, say two to ten proteins, there is little hope to statistical correlations of the sequences without strong assumptions about the nature of protein structures. We propose to use Upside as a strong Bayesian prior to aid the structure prediction. The simplest scheme directly incorporates the assumption that homologous sequences have closely related contact matrices into the potential energy function. We may simulate all of the homologous mutants as nearly-independent simulations. The potential energy would consist of independent Upside potential energies for each protein augmented by an energy term that increased the probability of a contact in proportion to the number of other proteins that also formed that contact for corresponding residues in a multiple sequence alignment. In the case of a single protein, this represent ordinary structure prediction through simulation. In the case of multiple proteins, each protein feels a force encouraging it to adopt (a superset of) the average contact matrix of all of the homologous sequences. The key idea is that homologous proteins will each have a low free energy for the true native structure, but each protein may also have low free energy for incorrect structures. So long as the incorrect structures are different for different homologues, they will add incoherently to the potential, and thus every structure except the true native will be weakened by the additional potential.

The advantages of the proposed technique are significant. This technique is far more data-efficient than statistical analysis. Even using two sequences is likely to greatly enhance the accuracy of structure prediction. Additionally, this cooperative folding incorporates strong, well-calibrated information about protein structures from the Upside model, an independent source of information from the statistical correlations. Finally, this method does not assume that the contact matrices of homologous proteins are extremely similar; each protein is free to have its own structure and contact matrix, which may be helpful in case a minority of proteins are strongly incompatible with the structures of the majority of the collection.

5.2.4 Computational simulation of mutational scans

Given the high efficiency of Upside simulation, we can directly simulate the large collection of mutants studied in a typical mutational scan. A typical method to study these proteins is an alanine-scan to compute ϕ-values at each residue. Instead of predicting the results through careful study of a wild-type protein simulations[1], we may perform an em in silico alanine-scan by running each of the mutant in Upside at reasonable computational cost. Such studies are likely to provide much greater insight both into the nature of the observed fractional ϕ-values as well as the trustworthiness of Upside simulation for the protein in question.

5.2.5 Prediction of conformational change

Pathway prediction can be performed in molecular dynamics using a variety of techniques, but the efficiency and accuracy of these methods are typically strongly tied to the ability of the scientist to propose a reasonable starting trajectory for atomic molecular dynamics. We expect that Upside will be useful for studying conformational change, both in studying conformational changes difficult to sample with atomic molecular dynamics and to provide suitable initial trajectories that can be further refined with atomic methods.

5.3 References

[1] Robert B Best and Gerhard Hummer. Microscopic interpretation of folding ϕ-values using the transition path ensemble. *Proceedings of the National Academy of Sciences*, 113(12):3263–3268, 2016.

[2] Jes Frellsen, Ole Winther, Zoubin Ghahramani, and Jesper Ferkinghoff-Borg. Bayesian generalised ensemble markov chain monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 408–416, 2016.

[3] Hüseyin Kaya and Hue Sun Chan. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Structure, Function, and Bioinformatics*, 40(4):637–661, 2000.

[4] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.

[5] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008.

[6] Fugao Wang and DP Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):056101, 2001.

[7] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *arXiv preprint arXiv:1609.00680*, 2016.