Optimization of BCL::Fold for Protein Folding de novo and with Cryo-EM Restraints

By

Michaela Sever Fooksa

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Chemical and Physical Biology

August 11, 2017

Nashville, Tennessee

Approved:

Jens Meiler, Ph.D.

Hassane Mchaourab, Ph.D.

ACKNOWLEDGMENTS

A number of people have helped me immensely throughout the process of my research and writing my thesis. I would like to thank my adviser, Jens, for all his guidance, as well as everyone in the Meiler lab for aways being willing to help out and have thoughtful discussions. I want to thank my friends Katie, Shonali, Leah, and everyone else for always supporting me through the ups and downs in my work, while also being there to remind me to have a healthy work-life balance. Lastly, I want to send the most love and gratitude to my parents and my brother Adam, whose unwavering support leaves me confident in all that I have done and will do in the future.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN Artificial Neural Network

ASIC Application-Specific Integrated Circuits

BCL BioChemical Library

CASP Critical Assessment of Protein Structure Prediction

CASP Critical Assessment of protein Structure Prediction

CCC Cross Correlation Coefficient

CS Chemical Shift

CV Collective Variable

DED Direct Electron Detector

EM Electron Microscopy

EMDB Electron Microscopy Data Bank

EPR Electron Paramagnetic Resonance

GA Genetic Algorithm

GDT Global Distance Test

HMM Hidden Markov Models

KBP Knowledge-Based Potential

LCS Longest Continuous Segment

MD Molecular Dynamics

MP Membrane Protein

MSA Multiple Sequence Alignment

MSM Markov state models

NMR Nuclear Magnetic Resonance

NOE Nuclear Overhauser Enhancement

PAM Partitioning Around Medoids

PDB Protein Data Bank

PSP Protein Structure Prediction

RDC Residual Dipolar Coupling

REMD Replica Exchange Molecular Dynamics

RMSD Root Mean Squared Deviation

SDSL Site-Directed Spin Labeling

SSE Secondary Structure Element

SVM Support Vector Machine

XL MS Cross Linking Mass Spectrometry

Chapter 1

Introduction

1.1 Computational Protein Structure Prediction Methods

Predicting the tertiary and quaternary structure of a protein based solely on its amino acid sequence has been referred to as the ”Holy Grail” of structural biology. Chapter 2 constitutes a comprehensive review of the current state of computational protein structure prediction. Specifically, I have written chapters 2.6, 2.7, 2.9, and 2.10.6, as well as portions of chapters 2.2 and 2.3.

The software used in the remainder of this work is named the BCL (BioChemical Library), and it was developed in the Meiler Lab at Vanderbilt University. The basic premise of *de novo* folding with the BCL is shown in Figure 1.1. The knowledge-based potentials that are used to evaluate the overall energy of each protein model come from statistics in the PDB. Specialized scoring terms can also be incorporated at this stage based on the folding protocol, including experimental restraints or specific scoring for membrane proteins. This work is concerned with improving the current methodology for the BCL's folding algorithm (BCL:Fold) either completely *de novo* or coupled with experimental data from cryo-Electron Microscopy (EM).

1.2 Cryo-Electron Microscopy

There are currently upwards of 100,00 known three-dimensional structures of proteins in the Protein Data Bank (PDB) [1], the vast majority of which were determined by X-ray crystallography. In general, crystallography does provide a reliable atomic structure of a protein crystal, but the requirement of a protein to crystallize introduces many problems. First, the process of crystallization can introduce artefacts and trap a protein in a non-native conformation. In addition, the time investment in determining suitable crystallization con-

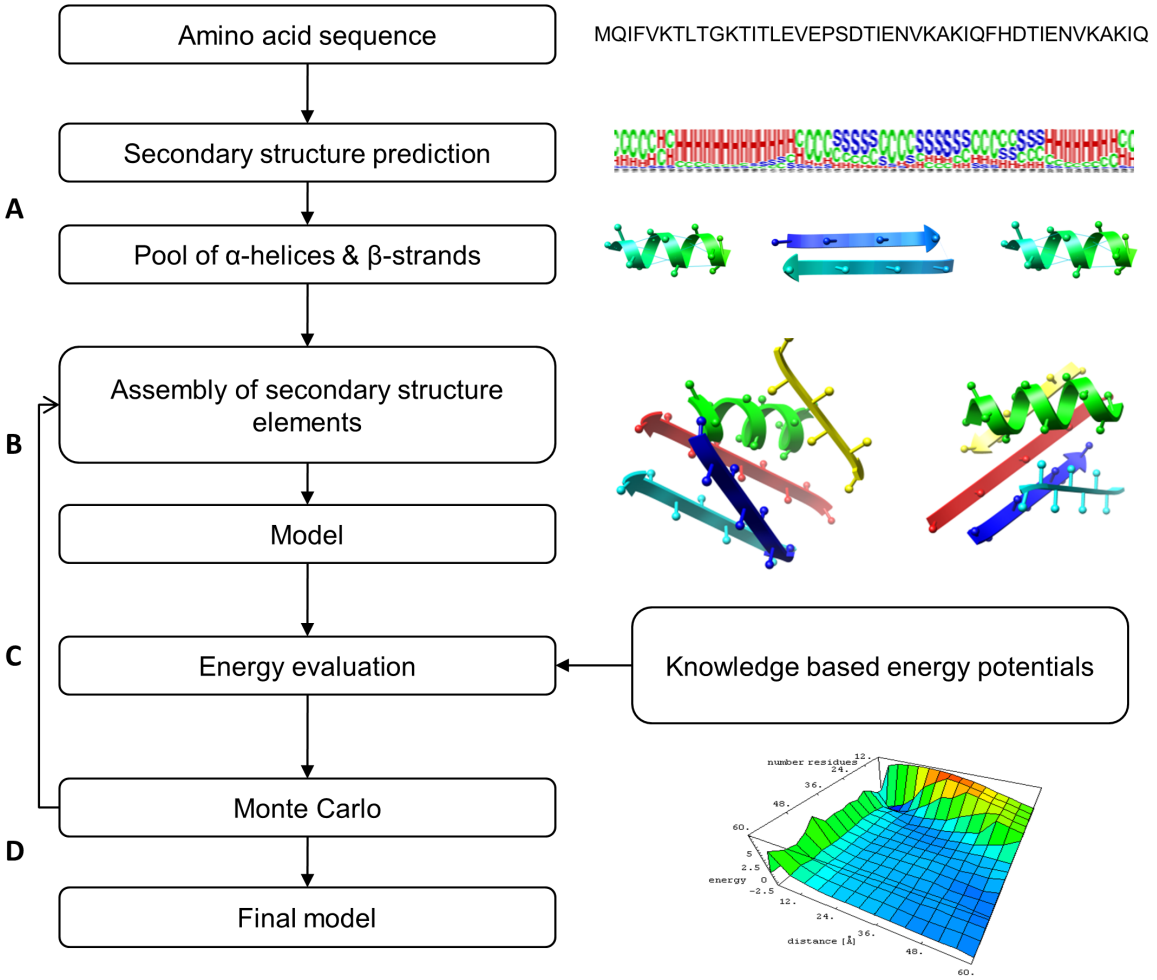

Figure 1.1: BCL De Novo Folding Pipeline. From a protein's primary structure, secocndary structure elements (SSEs) are predicted using a consensus prediction of various machine learning methods in order to create a pool of SSEs. Next, these SSEs are assembled in three-dimensional space using a stochastic Monte Carlo sampling algorithm that applies moves to entire SSEs. Each model is evaluated using knowledge-based energy potentials, and new models are either accepted or rejected based on the Metropolis criterion.

ditions can be enormous, and many proteins defy crystallization altogether.

An important class of molecules whose structures are underrepresented in the PDB are membrane proteins (MPs). MPs make up about 30% of the eukaryotic proteome, but only about 1.4% of protein structures in the PDB [1]. This dichotomy is especially problematic considering the fact that MPs constitute about 60% of drug targets [2]. Proper understanding of how a drug interacts with a protein relies on an intimate knowledge of the protein's structure. MPs are notoriously difficult to crystallize, sparking the need for alternative experimental techniques to structurally characterize these proteins.

One such technique is cryo-electron microscopy (EM). Studying protein structure by cryo-EM has the distinct advantage that proteins are not required to crystallize, which opens up cryo-EM's usefulness as a tool to study a breadth of differenty types of proteins. Rather, the proteins are plunge-frozen and imaged in their native state. This approach allows the direct inspection of a native-like protein, and the multiple conformations present also allow for scientists to investigate protein dynamics. Cryo-EM is thus particularly suitable for MPs, as they need only to be reconstituted and not coaxed into crystallization.

Cryo-EM has historically yielded lower-resolution images. At medium resolutions (30 - 12 Å), the overall shape and organization of macromolecules can be gleaned, but important functional parts of a protein are not visualized until higher resolutions. At about 10 Å resolution, α -helices can start to emerge as density rods, and β -strands can be visualized at about 5 Å resolution. Individual side chain character can start to be visualized at about 3 Å resolution.

Recent developments in imaging and image processing technology have allowed cryo-EM images to attain higher and higher resolutions. One such technology is the development of the direct electron detector (DED). DEDs have the advantage over their predecessor, the charge coupled device, in that they can directly detect electrons rather than first relying on converting electrons to photon counts. This process can introduce backscattering of electrons and loss of resolution. Another important technological advance was the advent

of movie mode. In movie mode, images are captured at intervals fractions of seconds apart and then recombined. In this way, any blurring effect caused by particle drift can be corrected for. Additional technological improvements, as well as the general increase of computational space and power available, have allowed many recent cryo-EM images to achieve atomic-level detail. In recent years, the atomic structure of the human ribosome was determined to 3.5 Å resolution by cryo-EM [3, 4]. In 2013, Yifan Cheng at the University of California, San Francisco determined the structure of TRPV1, a small MP, to 3.4 Å resolution [5], and the structure of γ-secretase, another MP implicated in Alzheimer's disease, was determined to a resolution of 3.4 Å in 2015 [6]. The number of protein structures determined by cryo-EM has been increasing at a rapid rate, and as these technologies become more and more widespread, cryo-EM is predicted to become more popular as a method of structural determination. EM can help elucidate the structure of many MPs.

Chapter 2

Towards Solving the Protein Folding Problem Computationally

2.1 Abstract

The protein folding problem has been a grand challenge in molecular biology for over half a century. Theories have been developed that provide us with an unprecedented understanding of protein folding mechanisms. However, computational simulation of protein folding is still difficult, and prediction of protein tertiary structure from amino acid sequence is an unsolved problem. Progress toward a satisfying solution has been slow due to challenges in sampling the vast conformational space and deriving sufficiently fast and accurate energy functions. Nevertheless, several techniques and algorithms have been adopted to overcome these challenges, and the last two decades have seen exciting advances in enhanced sampling algorithms, computational power, and tertiary structure prediction methodologies. This review aims at summarizing these computational techniques, specifically conformational sampling algorithms and scoring strategies that have been frequently used to study protein folding mechanisms or to *de novo* predict protein tertiary structures. We hope that this review can serve as an overview on how the protein folding problem can be studied computationally and, in cases where experimental approaches are prohibitive, help the researcher choose the right computational approach for the problem at hand. We conclude with a summary of current challenges faced and an outlook on potential future directions.

2.2 Introduction

Protein folding is a process of molecular self-assembly during which a disordered polypeptide chain collapses to form a compact and well-defined three-dimensional structure. A grand challenge in structural biology has been to understand the process by which

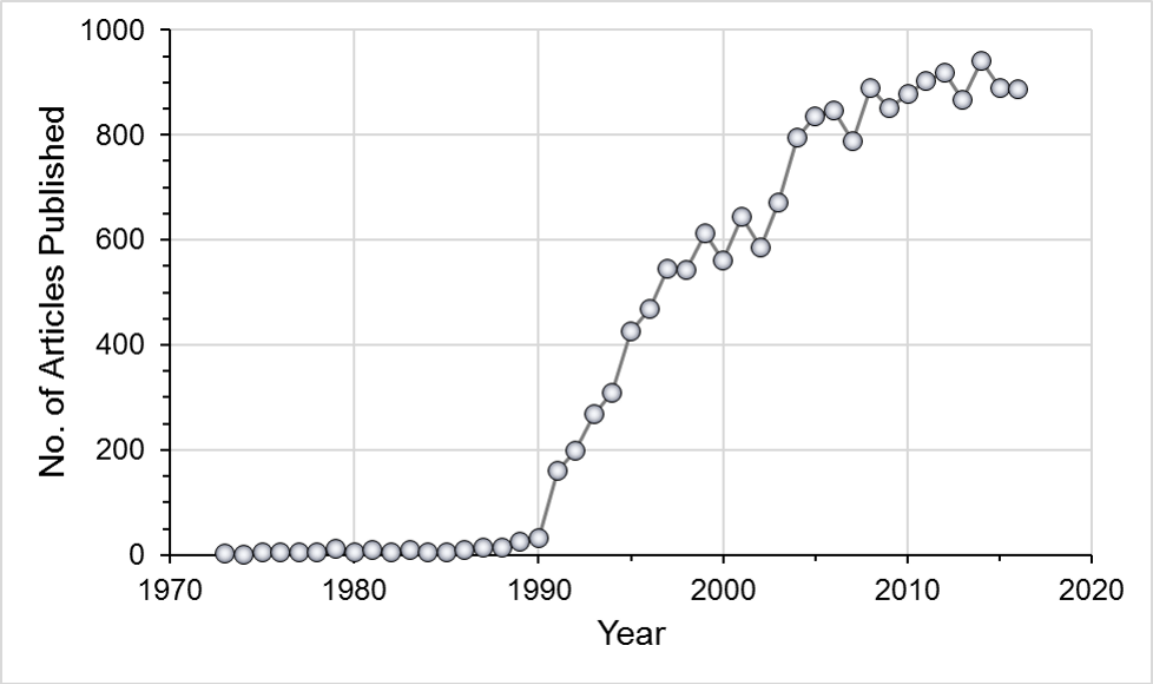

Figure 2.1: The number of articles published each year (1973 - 2016) with the phrase ”protein structure prediction” or ”protein folding” in either the title, abstract, or author keywords. The data were taken from Web of Science.

proteins fold into their functional tertiary structure (folding mechanism) and to predict this tertiary structure from amino acid sequence (structure prediction), two tasks that are collectively known as the protein folding problem.[7, 8, 9] Solving this problem is of far-reaching impact as it will not only reveal the missing link between sequence and structure but also provide molecular biologists with a theoretical framework and practical tools for applications such as drug design and protein engineering. As a result, an enormous amount of effort has been contributed to study the protein folding problem by the scientific community. This is illustrated by Figure 2.1, which shows the striking growth in the number of articles published each year on this problem since Anfinsens thermodynamic hypothesis of protein folding, that protein native state resides in the global minimum of Gibbs free energy, was formally stated in 1973. [10] A comprehensive review of the study of this problem is deemed impossible for an article of this kind. As many excellent review articles on the theory and experimental validation of protein folding have been published

over the years,[11, 12, 13, 14, 15, 16, 17, 18, 19, 20] here we focus our discussion on computational methods for studying folding mechanisms and predicting tertiary structures. Specifically, we limit our discussion to protein folding simulations and de novo protein structure prediction at atomic detail, as methods based on coarse-grained representation of protein structures were recently comprehensively reviewed.[21] In addition, due to space limitations, we are not able to cover the complete literature of this topic, and we apologize to those whose contributions have not received the deserved attention.

The two key components of any folding simulation or structure prediction methods are efficient sampling of conformational space and accurate evaluation of the energy of sampled conformations. Hence, the main body of this article is devoted to discussing different algorithms and their advances toward efficiently sampling extensiveconformational space followed by approaches and progress toward accurate energy functions. To put the discussion under the theoretical framework of protein folding, we first briefly summarize different views on the mechanisms of protein folding. The interplay between sampling algorithms and energy functions is concretely illustrated by discussing some representative methods shown to be relatively successful in the Critical Assessment of protein Structure Prediction (CASP) experiment.[22, 23] Finally, we present a summary on the progress and outline specific challenges that future development in the field will likely overcome.

2.3 Various Views on the Mechanism of Protein Folding

The conformational space accessible to a polypeptide chain is astronomically large, a systematic search for the functional structure of a polypeptide chain with 100 residues would take an amount of time even longer than the age of the universe. The fact that proteins fold on a biologically meaningful timescale, with some attaining to their functional structure in just a few microseconds, led Levinthal to conclude that there must be a specific folding pathway that each protein follows when it folds to the native state [24]. This classical view of protein folding assumes a sequential model and postulates a well-

defined sequence of intermediates which follow one another so as to carry the protein from the unfolded random coil to a uniquely folded native state.[24, 25, 26] However, the observation that kinetic folding intermediates (molten globules) form asynchronously over a range of timescales fostered the multipath funneled energy landscape paradigm of protein folding.[13, 17, 19, 27, 28, 29] In this new view, it is inferred that proteins must fold into their unique native state through multiple unpredictable pathways that involve the progressive organization of an ensemble of partially folded intermediates on a rugged free energy surface that resembles a funnel (Figure 2.2). The native state resides in the deepest basin of the funnel according to Anfinsens thermodynamic hypothesis.[10] A more recent formulation of the mechanism of protein folding is centered around of concept of foldons.[16] In whats called the foldon-based hypothesis, a protein starts folding by forming an initial seed foldon through unguided search, and it follows a foldon-determined folding pathway as the seed foldon guides subsequent foldons in a folding upon binding way. While this hypothesis states that proteins fold along a definite path after formation of the initial foldon, the foldon formation at the initial stage is assumed to be accomplished through a disordered multitrack search.[16]

2.4 Conformational Sampling is a Bottleneck

A polypeptide chain with a typical size can adopt an astronomical number of conformations. For example, if we assume that the backbone torsional angles each can adopt three stable conformations, then a protein made up of a polypeptide chain consisting of 100 amino acid residues would have a total of about 10^{49} conformations. This is a gross underestimate, given that the degrees of freedom of sidechains are ignored. It is agreed that conformational sampling remains to be a bottleneck of de novo structure prediction.[30, 31, 32, 33, 34] Nevertheless, there has been exciting improvement in sampling algorithms based on statistical mechanical principles or guided by experimental or predicted restraints, all of which are further accelerated by improvements in hardware speed and power. For the

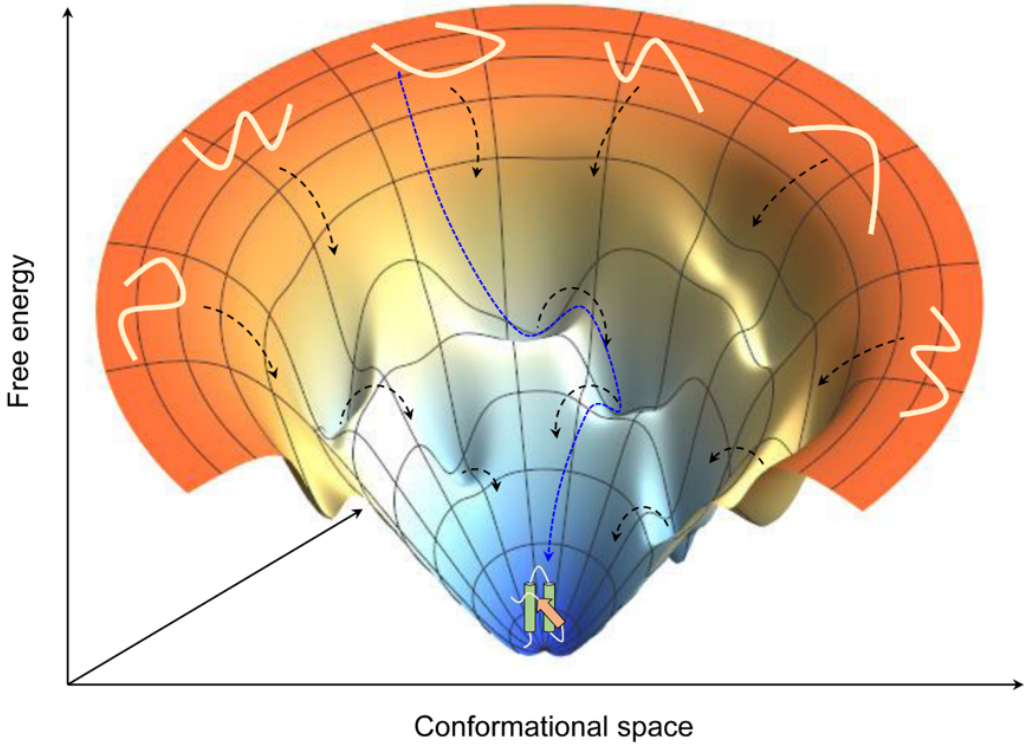

Figure 2.2: Surface Rendering of a Hypothetical, Simplified Folding Energy Landscape. Extended polypeptide chains at the top of the funnel might navigate through a myriad of different routes (black dashed arrows) or a single definite path (blue arrow) to the native state, sitting at the bottom of the funnel. Some routes involve transient intermediates while others might have significant local energy minima that are so deep that the protein is kinetically trapped in them.

convenience of discussion, we divide conformational search methods into the following three broad categories: molecular dynamics simulations, Monte Carlo simulations, and genetic algorithms. For each category of algorithms, we give a general formulation of the algorithm and a review of the latest studies in which the algorithm was applied to *de novo* protein structure prediction.

2.4.1 Unbiased Molecular Dynamics Simulations

Molecular dynamics (MD) simulation is a widely used computational technique for exploring the macroscopic properties of molecular systems through explicit computation of microscopic particle motions. MD has had enormously influential applications in biomolecular systems and has been heavily used to study motion-related phenomena such as protein folding, conformational flexibility, protein structure determination from NMR, ligand-protein interaction, and protein-membrane interaction.[35, 36, 37, 38, 39, 40, 41, 42, 43] The two essential elements of a MD simulation are the interaction potential for the particles and the equations of motion governing the dynamics of the particles.[44, 45] Here, we describe how MD simulations explore the phase space of a molecular system. A typical MD run involves generation of successive microstates of a molecular system by solving Newtons equations of motion for all atoms simultaneously with femtosecond timesteps (Equation 1).

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{F}_i = -\frac{\partial U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)}{\partial \mathbf{r}_i} \tag{2.1}$$

where $\mathbf{r}_i$ and $U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ denote position vector and potential energy of point mass i, respectively. $\mathbf{F}_i$ denotes the force acted upon point mass i. The result of the simulation is a trajectory of microstates that specifies how the system explores the phase space as time goes by [44]. The result of the simulation is a trajectory of microstates that specify how the system evolves in phase space [44]. In principle, equilibrium properties can be computed by averaging over the trajectory if it is of sufficient length to give a representative

ensemble of the microstates of the system. Unfortunately, the usefulness of MD in studying long timescale biological phenomena is often limited due to inadequate sampling of all relevant conformational states of a system. Even when the energy barriers separating two topologically different low energy regions of the conformational space are of order k_BT, traversing them by random thermal fluctuation cannot be achieved within a reasonable amount of time.

A wide range of biologically interesting phenomena occurs over timescales on the order of milliseconds, several orders of magnitude beyond the reach of traditional MD simulations. As a result, studying processes that involve major conformational changes, such as protein folding, activation, and deactivation, by MD simulations has been traditionally challenging.[46] The very first successful *de novo* protein folding simulation via MD was notably made by Duan and Kollman, who folded the villin headpiece (a 36-mer) in explicit solvent for two months on parallel supercomputers to obtain models up to 4.5 Åfrom the experimental structure [47]. This small protein was later folded by Pande and coworkers to 1.7 Åwith a total simulation time of 300 μs or approximately 1000 CPU years with the help of worldwide-distributed computers [48].

Substantial progress has been made during the past decade or so to extend the folding times accessible by conventional MD simulations through efficient parallelization of MD codes or MD-specialized hardware (Figure 2.3).[49] The MD-specialized parallel method Desmond and its associated massively parallelized machine Anton, developed relatively recently in the Shaw research group, allow for conducting millisecond timescale MD simulations of systems with tens of thousands of atoms in just a few weeks [50, 51]. Anton is built on MD-specific ASICs (application-specific integrated circuits) that interact in a tightly coupled manner using a high-speed communication network. Its ability to efficiently perform simulations on the timescales over which many physiologically relevant processes take place expands substantially the set of problems for which the MD approach is tractable. Armed with this specialized set of software and hardware, Shaw et al were

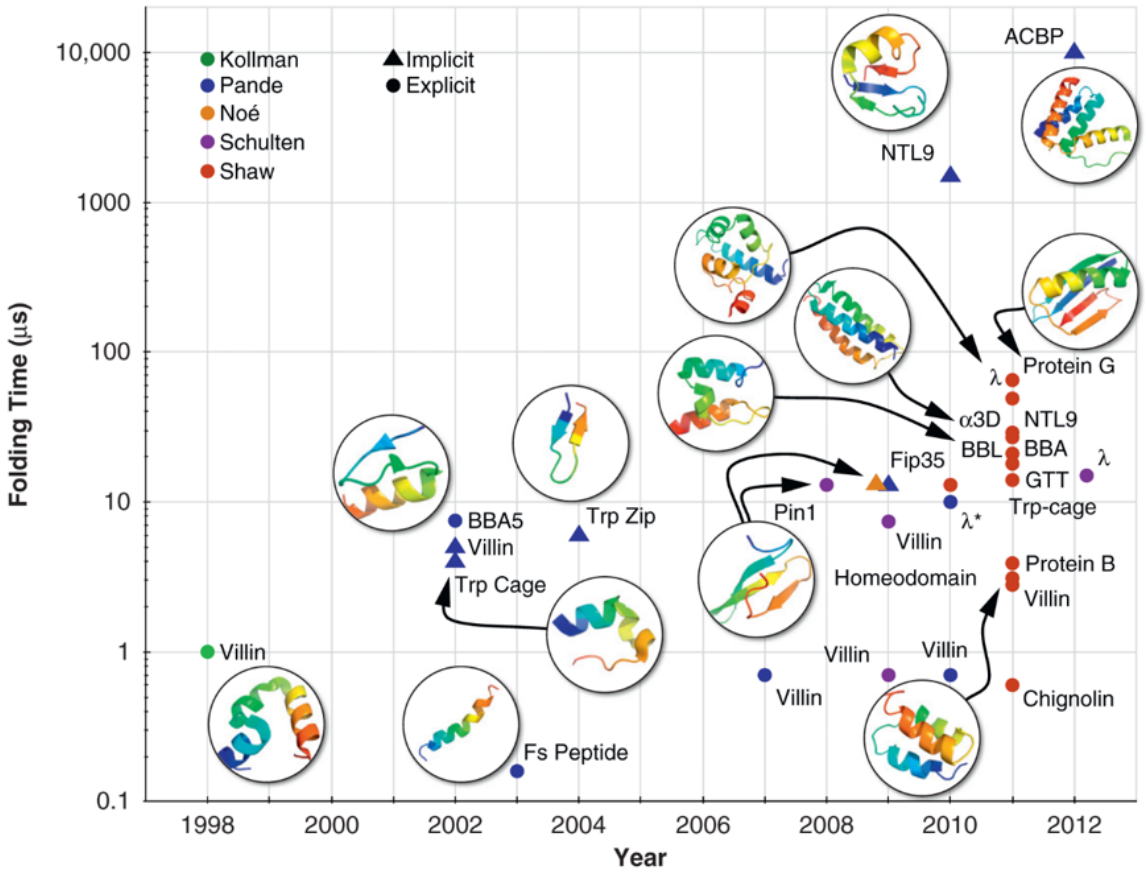

Figure 2.3: Folding timescales accessible to MD simulations have increased exponentially since Duan and Kollman folded the villin headpiece by MD simulations in explicit solvent. [47] Shown are all proteins simulated using unbiased, all-atom MD in empirical force fields reported in the literature. Adopted, with permission, from reference . See reference for reference to each folding simulation highlighted in the figure.

able to simulate protein folding from extended random coils [52, 53] and structural origin of slow diffusion in protein folding [54], protein-ligand recognition [55, 56], mechanism of nucleotide exchange in G proteins [57], and mechanisms of kinase activation and inhibition, at realistic timescales. The *de novo* folding simulations conducted by Shaw et al generated evidence in favor of the single-pathway view of protein folding (Figure 2.2). For example, equilibrium simulations of WW domain captured multiple folding and unfolding events that consistently follow a well-defined folding pathway. [53] Subsequent folding simulations of 12 fast-folding proteins tend to emulate the foldon-based distinct pathway picture.[52]

A different approach to overcome the sampling challenge of MD is through statistical analysis of multiple independent trajectories or aggregating independent short simulations using Markov state models (MSM) to make a complete model of system dynamics.[49, 58, 59] The MSM effectively pieces together this complete model from independent trajectories, allowing for prediction of kinetic phenomena on timescales much longer than the individual trajectories used to construct the model.[49] While the MSM-based "multi-trajectory" approach has some advantages over the reaction coordinate-based single trajectory analysis, such as identifying areas of phase space for adaptive sampling,[60, 61] insights gained from MSM analysis unfortunately disagreed with the foldon-based single pathway view of folding. For example, while Shaw et al concluded that folding of the WW domain follows a definite pathway where the first hairpin folds first,[53] using MSM to analyze the same simulation trajectories, Lane et al detected a parallel statistically significant pathway where the second hairpin of the WW domain folds first.[62] Similar analysis conducted by Beauchamp et al [63] on the MD trajectories of 12 small fast-folding proteins by Lindorff-Larsen et al[52] also arrived at conflicting conclusions regarding the folding mechanism of half of simulated systems.

2.4.2 Enhanced Sampling Techniques in MD

The roughness of energy landscapes with many local minima separated by high-energy barriers makes adequate conformational sampling a challenging task. MD trajectories often do not reach all biologically relevant conformations, a problem that can be addressed by employing enhanced sampling algorithms.[64, 65] Several enhanced sampling techniques in simulations of biological systems have been developed such as replica exchange molecular dynamics (REMD) and metadynamics.[65]

The replica-exchange method was developed to overcome the multitude of local minima separated by high energy barriers.[66] Many molecular simulation scenarios require ergodic sampling of energy landscapes that feature many minima, and barriers between minima can be difficult to overcome at ambient temperatures over accessible simulation timescales. Replica-exchange simulations seek to enhance the sampling in such scenarios by running n non-interacting copies of the system $C_i\,(i = 1, \cdots, n)$ in parallel each at a different temperature T_i in the canonical ensemble (Figure 2.4). The non-interacting nature of this artificial compound system ensures that each states weight factor $(C_1, C_2, \cdots, C_n)$ is given by the product of Boltzmann factors of each copy.

$$w = \exp\left\{-\sum_{i=1}^{n} \beta_i U_i\right\} \tag{2.2}$$

Compared to a standard Monte Carlo simulation, which affects the conformation of only one copy, REMD explores the energy landscape by periodically exchanging the conformations of replicas. The probability of transition of a compound system such that the conformations between a pair of copies $\left(C_i, C_j\right)$ are exchanged is

$$p = \min\left(1, e^{\Delta}\right) \tag{2.3}$$

Where

$$\Delta = \left(\beta_j - \beta_i\right)\left(U_j - U_i\right) \tag{2.4}$$

Exchange of the conformations of replicas decreases auto-correlation, thus enabling replicas to reach thermal equilibrium faster than without exchange. While it is not necessary to restrict the exchange to copies with neighboring temperature (e.g. $j = i + 1$), doing so will be optimal, since the transition probability decreases exponentially with the difference in temperature between copies [67]. It is also worth noting that while exchange of conformations between copies must be conducted in a Monte Carlo way, there is no restriction on which algorithms should be used for updating the conformation of an individual copy locally. In fact, several variants of REMD have been developed.[68] For example, a replica-exchange Monte Carlo (REMC) technique was implemented in the threading-based structure prediction pipeline QUARK and tested in CASP11.[69]

Metadynamics is a class of methods that eases sampling by introducing a time-dependent biasing potential that acts on a selected number of coarse-grained order parameters, often referred to as collective variables (CVs).[70, 71, 72, 73] CVs are generally nonlinear functions of the atomic positions of the simulated system that should ideally distinguish between all relevant metastable states. Some simple but informative CVs used in protein folding simulations are number of C contacts, number of backbone H-bonds, and helicity of the backbone, and the free energy surface is usually plotted as a function of these CVs.[70] The added biasing potential is introduced through successive addition of small repulsive Gaussian kernels deposited along the system trajectory in CV space (Figure 2.4).[72, 73] The added Gaussian kernel is a function of the current position and the previous position of the system in the CV space, and its intended purpose is to discourage the system from revisiting configurations that have already been sampled, thus accelerating sampling. The final summation of the deposited Gaussian kernels also gives an unbiased estimate of the free energy landscape of the system. In contrast to these advantages, it is, however, far from trivial to decide when to stop a simulation and find a set of CVs proper for describing

the process of interest.[72, 73]

Both REMD and metadynamics have been used to de novo fold several small peptides and proteins. The first example of using REMD to sample a folded structure starting from a completely unfolded state is probably the study of Rhee et al where a 23-residue BBA5 protein was folded by whats called multiplexed REMD.[74] Using REMD simulations in implicit solvent, Pitera et al folded a 20-residue designed Trp-cage peptide starting from an extended coil to a state ¡ 1.0 ÅC_α RMSD from conformations in the NMR ensemble.[75] Recently, Jiang et al folded a diverse set of 14 fast folding proteins from their unfolded states using REMD with a residue-specific force field.[76] A similar study by Nguyen et al included a larger set of 17 proteins; while they successfully folded most proteins, misfolded structures are thermodynamically preferred for 3 proteins.[77]

2.5 Monte Carlo Simulation

MD simulation is without a doubt a required technique if one wishes to study folding pathway or kinetics computationally. However, for tertiary structure prediction of large proteins whose energy landscapes are populated with many local minima separated by high barriers, Monte Carlo (MC) simulation can be much more efficient (Figure 2.5a). It is, in fact, the underlying search engine of some of the most successful *de novo* tertiary structure prediction methods [69, 78, 79, 80] and our method BCL::Fold.[81] Unlike MD simulations where successive conformations of the system are connected through time, in a MC simulation, each new conformation of the system depends only upon its immediate predecessor. The technique of MC simulation was introduced as the first computer simulation of a molecular system in 1952.[82] Nowadays, the term Monte Carlo is often used to describe a simulation whenever random sampling is performed.

A MC simulation explores the phase space of a system by randomly perturbing the current conformation by actions such as moving a single atom or molecule or adjusting dihedral angles. The energy of the new conformation is then evaluated using the energy

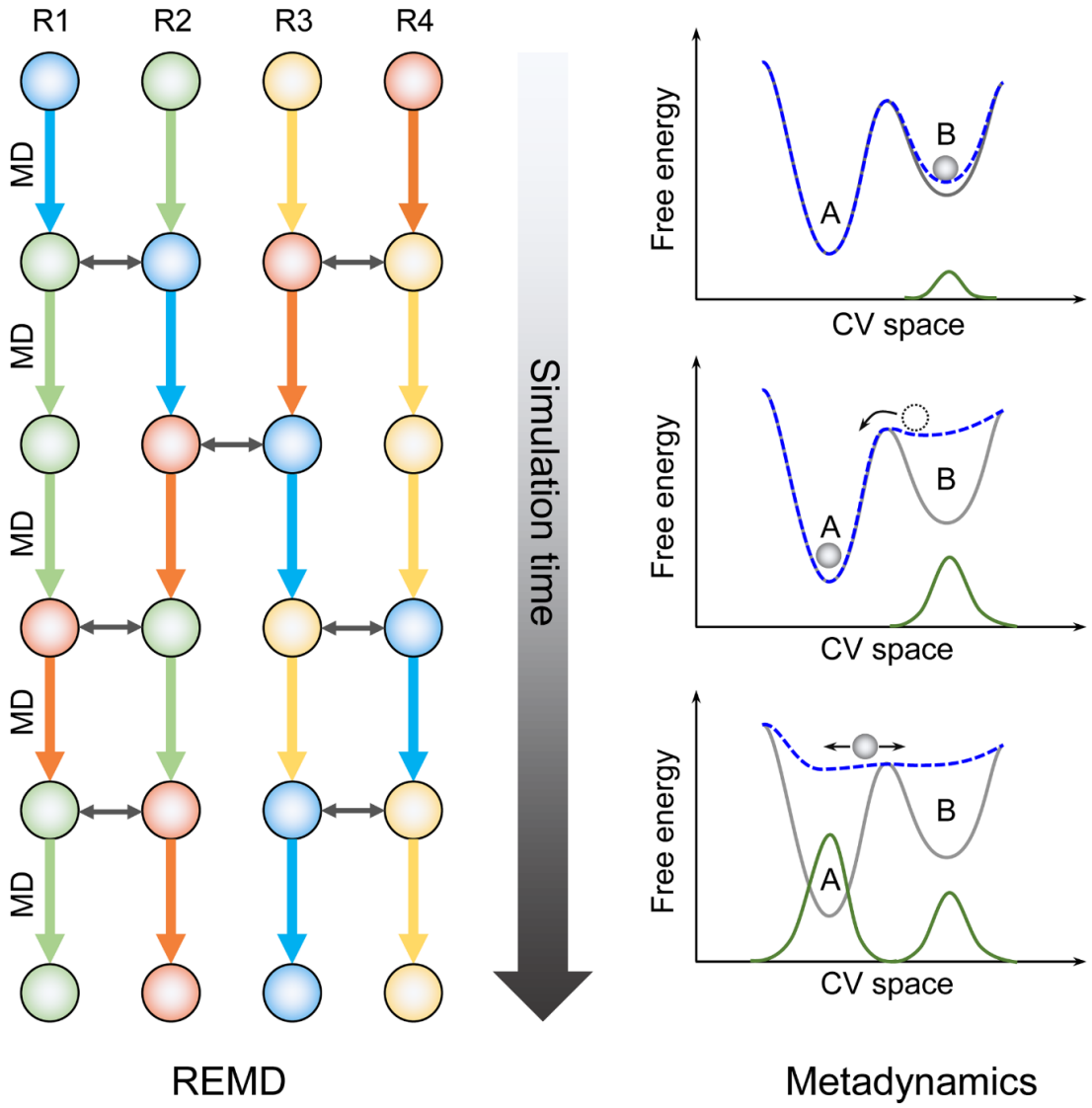

Figure 2.4: A sketch of the process of REMD and that of metadyanmics. REMD: a set of non-interacting replicas (4 in this illustration), each runs at a different temperature. In an efficient REMD, replicas at neighboring temperatures are swapped (shown as double-headed arrows) based on Metropolis criterion and all replicas will experience swapping. Metadynamics: this illustrative system has two minima A and B (grey curve). The system trapped in B is lifted by progressive deposition of repulsive Gaussian kernels (green curve) and the free energy landscape changes accordingly (blue dashed curve). After B is filled up, the system moves into A which is filled up similarly. When the simulation completes, the green curve gives a first rough negative estimate of the free energy landscape of the system.

function. If the new conformation is lower in energy than its predecessor, it is accepted as a starting conformation for the next iteration. If the energy is higher, the new conformation is accepted with a probability based on the infamous Metropolis criterion (Equation 3).[82] This is often done by comparing the Boltzmann factor of the new conformation to a random number between 0 and 1, and the new conformation is accepted if its Boltzmann factor is greater than the random number and rejected otherwise. While the essential search algorithm of MC-based structure prediction methods is the same, they differ in the starting components for assembling 3D models and in the repertoire of MC moves implemented for perturbing the model.[83]

Primitive MC sampling can be computationally expensive and thus inefficient at finding global energy minima. Typically, these methods are coupled with some optimization technique that vastly decreases computational expense by directing the progression of the MC simulation toward global energy minima. One optimization technique is gradient-based sampling, where MC iterations are directed down local property gradients, i.e. the potential next state with the lowest energy is selected. For instance, gradients can be calculated based on side chain rotameric states [84] or, in the HP-lattice model,[85] the movement of a residue in various directions.[86] However, when the conformational space is continuous rather than discrete, gradient descent becomes unfeasible because the energy cannot be calculated for every step forward. The most popular optimization approach shown to effectively accelerate the convergence of a MC simulation is probably simulated annealing.[87] The essential feature of this technique is that it combines Monte Carlo sampling of conformational space at an initially elevated temperature with an appropriate cooling scheme over the course of the simulation. The cooling scheme, if gentle enough, theoretically ensures the system will reach the global minimum. In turn, the probability of a higher energy step being accepted decreases over time, and models are directed toward the global energy minimum.[88] Many powerful *de novo* tertiary structure prediction methods integrate this MC simulated annealing approach;[21] we include a detailed discussion on some se-

lected examples of such methods (see Examples of methods for *de novo* tertiary structure prediction).

2.6 Genetic Algorithms

Genetic algorithms (GAs) are an optimization procedure based on the process of evolution that occurs in nature. GAs have been used in a variety of applications. Some prominent ones include automatic programming, machine learning, and population genetics [89]. Generally, a GA initializes the optimization process by randomly generating an initial population of trial solutions each encoded as string of bits, also called a chromosome (Figure 2.5b). Offspring are produced by applying nature-inspired operations, namely mutations and crossovers on bit strings. Mutations are introduced into strings by flipping one or more bits, whereas crossovers between two individuals consist of randomly selecting a crossover site and exchanging the left segment of one string with the right segment of the other (Figure 2.5b). The fittest offspring are selected for continual refinement via the iteration of multiple generations [90].

A large number of studies of the use of GAs for *de novo* protein structure prediction and protein folding simulation have been made [90, 91, 92, 93, 94, 95, 96, 97, 98, 99] since the pioneering work of Dendekar and Argos on *de novo* folding simulation of a model protein of a four β strand bundle [100] and that of Unger and Moult on searching for global energy minimum on the 2D HP lattice model [101]. The simplest protein representations used in GAs is the 2D HP model developed by Lau and Dill.[102] In this model, amino acids are of only two types: hydrophobic (H) or polar (P). The sequence is folded on a 2D square lattice on which bonds are orthogonal to each other. Folded structures are evaluated by a so-called ”hydrophobic potential” where each pair of non-bonded direct hydrophobic contact (occupying neighboring non-diagonal lattice vertices) receives -1. Using HP lattice models avoids the computational power needed for all-atom models while still capturing the general principles that govern protein folding, and they can be extended to account for

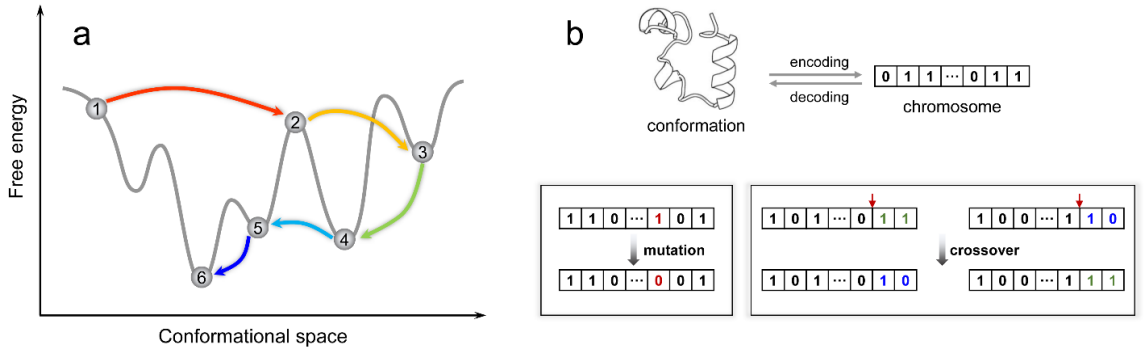

Figure 2.5: Monte Carlo simulated annealing and genetic operations in genetic algorithms. a) A Monte Carlo simulated annealing procedure allows the system to ”freely” navigate the free energy surface. For example, transition from state 4 to 5 would be prohibitive to MD simulations due to the high energy barrier separating them. b) In genetic algorithms, conformations are encoded as bit strings (or real-valued arrays) called chromosoms. A mutation operation flips the bit value at a randomly selected site, whereas a crossover operation takes a pair of chromosomes and exchanges parts of chromosomes split at a randomly selected crossover site.

physicochemical characteristics of individual residues such as size, hydrophobicity, and charge. In more detailed models, proteins can be represented as a sequence of pairs of dihedral angles that describe the backbone degrees of freedom of each residue. Mutations can simply be introduced by changing the dihedral angle of a residue and cross-overs by swapping randomly assigned sections of two sequences [90, 93]

2.7 Energy Functions are Evolving Objects

An essential part of almost all successful protein folding simulations or protein tertiary structure predictions is an energy function that is a good approximation to the energy landscape of real proteins. Protein energy functions can be roughly divided into two classes: physics-based force fields and knowledge-based potentials.[103] Historically, physics-based forces fields are coupled with MD or MC simulations to study protein dynamics or calculate free energies,[104, 105, 106, 107] whereas knowledge-based potentials are mostly used for fold recognition or tertiary structure prediction.[108, 109, 110] Before

we give a detailed account on them, we remind the reader that both of these two types of energy functions are evolving objects. To improve accuracy, further parameter optimization for physics-based force fields is required and statistics need to be rederived for knowledge-based potentials when energy function deficiencies are identified or data sets of better quality become available.

2.7.1 Physics-Based Force Fields

Physics-based force fields are classical mechanical models that approximate the potential energy of chemical systems. Force field models ignore the electronic motions in a system and only considers interactions among nuclei. Compared to *ab initio* quantum mechanical methods, force fields are much more computationally efficient while giving a quite acceptable level of accuracy. A force field has a functional form and a set (usually very large) of associated parameters taken together to model bonded and non-bonded interactions in a system. The functional form of a force field is often a compromise between accuracy and computational efficiency and depends on the level of resolution (all-atom or coarse-grained), chemical nature (inorganic, small organic, or biomolecular), and target properties of the systems to be modeled. Nevertheless, most force fields generally contain five components. The first three of them, so-called bond stretching, angle bending, and torsion, model bonded interactions. The last two components describe electrostatic and van der Waals non-bonded interactions [44].

$$\begin{aligned} U(\mathbf{r}^N) = \sum_{bonds} \frac{k_b}{2}(l-l_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta-\theta_0)^2 + \sum_{torsions} \frac{k_\varphi}{2}[1+\cos(n\varphi-\gamma)] \\ + \sum_{electrostatics} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} + \sum_{VDW} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \end{aligned} \tag{2.5}$$

This functional form looks relatively simple but we must keep in mind that the set of parameters associated with it is very large. For example, the term that models bond stretching (a harmonic potential) has for each bond type a force constant k_b and an equilibrium bond

length l_0. These parameters of course must be determined by fitting the force field to a given set of data obtained from experiments or quantum mechanical calculations. Depending of the size of the data set, parameter optimization may be conducted in a number of ways: trial and error, least-squares fitting [111] or recently machine-learning algorithms [112].

Well-known examples of force fields intended for modeling proteins include CHARMM [113, 114, 115, 116, 117, 118], AMBER [119, 120, 121, 122], OPLS [123, 124], GROMOS [125, 126], MARTINI [127, 128]. Their functional form invariably contains the five terms of Equation 5, the major difference between them lies in specifics in the treatment of non-bonded interactions and the levels of resolution covered. For modeling non-bonded interactions, the original CHARMM and AMBER force fields both incorporate 12-10 Lennard-Jones potential to explicitly model hydrogen-bonding [114, 120], whereas the OPLS and GROMOS force field uses only the electrostatics and van der Waals terms [123, 125]. The need for more efficient evaluation of non-bonded interactions arises when the number of interaction sites is large. One simple way to improve efficiency is to absorb aliphatic hydrogens into the carbon atom to which they are bonded to form 'united atoms' as was done in the united-atom version of the CHARMM and OPLS force fields, or to use a coarse-graining approach where a group of heavy atoms are combined for form a representative virtual interaction site. The MARTINI force field aiming at providing a simple model that is computationally fast and easy to use adopted a 'four-to-one' coarse-graining scheme [129]. If special care is taken during parameterization, comparable level of accuracy is possible with these reduced representations while achieving considerable computational savings. Coarse-grained protein models and their applications was recently reviewed in detail by Kolinski et al [21].

Molecular mechanics force fields are traditionally coupled with MD in simulating protein dynamics and folding [130]. There have been a plethora of such studies where the utility of force fields for protein tertiary structure prediction or the accuracy of reproducing

experimental data were reported [52, 47, 48, 131, 132, 133, 134, 135, 136]. However, no agreement has been reached regarding whether force fields are sufficiently robust for these applications [137, 138]. Early analysis concluded that force fields driven by MD simulations are not particularly successful in structure prediction.[137] However, evidence has been accumulating that demonstrate that physics-based force fields are sufficiently accurate for predicting native-state structures and folding rates [52, 53, 54, 133, 138, 139]. In particular, it was pointed out the prediction of tertiary structures, folding rates, and melting temperatures appears to be more robust than is the prediction of the enthalpy and heat capacity of folding or that of the radii of gyration of unfolded states [138].

2.7.2 Knowledge-Based Potentials

Unlike physics-based force fields, which model interactions found in the most basic molecular systems using fundamental physics laws explicitly and separately, knowledge-based potentials (KBPs) are energy functions derived from statistical analyses of known protein structures and the application of the inverse Boltzmann relation to the probability distribution of geometries [140, 109, 141]. The physical meaning of KBPs has been under vigorous debate since their introduction [142, 143, 144, 145, 146, 147], although justifications of KBPs as "potentials of mean force" have been provided by analogy to the reversible work theorem in statistical thermodynamics [148] or on the basis of probabilistic arguments [78, 147]. Nevertheless, KBPs are widely used and surprisingly effective in scenarios including but not limited to protein structure prediction [78, 149, 80, 150], refinement of NMR structures[151, 152], fold recognition [153, 154], protein-ligand or protein-protein interactions [155, 156, 157, 158], and protein design [159]. Thus, in this article, we summarize the formalism of KBPs, specific implementations of different types of potentials, and their applications instead of concerning about the physical interpretation of KBPs. A KBP energy function is a linear combination of individual potentials with each capturing a

specific type of interaction. The most common formulation of such scoring functions is:

$$E\left(C|S\right)=\sum_{ij}w_{ij}\left(-kT\ln\frac{p\left(c_j|s_i\right)}{p\left(c_j\right)}\right) \tag{2.6}$$

where $E(C|S)$ is the energy of conformation C given that the underlying amino acid sequence is S. $p(c_j|s_i)$ is the probability that a given sequence s_i adopts conformation c_j, whereas $p(c_j)$ is an unconditional probability that any sequence fragment adopts conformation c_j. $\frac{p(c_j|s_i)}{p(c_j)}$ can be thought of as an "equilibrium constant" of a hypothetical chemical reaction: *random sequence, unique conformation*$\rightarrow$ *unique sequence, unique conformation* [144]. The types of individual potentials incorporated into a KBP scoring function is essentially only limited by the type of statistical relations that can be practically extracted from known protein structures. Thus, different implementations of KBPs remain rather diverse. In addition to the above inverse Boltzmann formulation, other formations of individual KBP terms have also been widely used. For example, the KBP under the modeling package ROSETTA was formulated based on Bayesian theorem [78]. This approach was also adopted by Woetzel et al recently to derive the KBP for a secondary structure element-based protein folding algorithm [81, 160, 161, 162]. In their Discrete Optimized Protein Energy, or DOPE, Shen and Sali compute the negative logarithm of the joint probability density function of a given protein [150].

The types of individual potentials incorporated into a KBP energy function are essentially only limited by the type of statistical relations that can be practically extracted from known protein structures. Depending on its intended purpose, a KBP may include individual potentials that fall into one or several categories. We elaborate three such potentials in the following and refer the reader to references [78, 80, 160] for examples of other potentials.

1) *pairwise distance-dependent potential* that approximates residue contact energies [140, 109, 141, 163]. The concept of pairwise distance-dependent potentials was first in-

troduced in the pioneering work of Tanaka and Scheraga, who related residue contact frequencies to the free energies of formation of corresponding interactions using the simple relationship between free energy and equilibrium constant [164]. Their work was followed by that of Miyazawa and Jernigan, who formalized the theory of residue contact potentials using quasi-chemical approximation [165, 166]. However, these early implementations of contact potentials are in fact not distance-dependent except that a single cutoff distance was used to define residue contact. A real pairwise distance-dependent potential was first introduced by Sippl [163] and this was followed by an explosion of different statistical potentials [149, 150, 160, 167, 168, 154, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180]. Such pair potentials are usually formulated at residue level where inter-residue distances are measured between Cβ atoms or sidechain centroids in reduced representation of amino acid residues to promote computational efficiency. However, atomic level formulation usually gives better discriminatory power albeit at the cost of more computational resourse [148, 149, 150, 173, 175, 181].

2) *solvent accessibility-based environment potentials* that represent the interactions of individual residues with their local environment [80, 154, 182, 183]. Residue environment potentials are often included to account for solvation effects. Precise calculation of solvent accessibility requires full atomic structure and is time-consuming. In tertiary structure prediction scenarios where reduced representations of residues are used, good approximations to solvent accessibility such as residue contact numbers provide significant computational savings [160, 184, 185, 186]. It should be noted that in addition to transforming solvent accessibility statistics to energy-like potentials using the inverse Boltzmann relation, they have also been incorporated into KBP scoring functions as a penalty term to disfavor models where residue-specific solvent accessibilities disagree with expected solvent accessibilities [80].

3) *potentials of torsion angles* that evaluate backbone Φ, Ψ torsion angles and/or the preference of sidechain rotamers [154, 187, 188]. It is well known that only certain com-

binations of Φ, Ψ torsion angles are populated in proteins [189] and significant correlations exist between side-chain torsion angle probabilities and backbone Φ, Ψ angles.[190] Including such potentials has been shown to enable the energy function to exclude conformations that have unlikely combinations of torsion angles. In a study by Kocher et al where several types of potentials were tested to recognize protein native folds, potentials representing backbone torsion angle preferences recognized as many as 68 protein chains out of a total of 74.[154] This result was striking given the fact that backbone torsion potentials consider solely local interactions along the chain and are well known to be incapable of determining the full 3D fold.[154] Potentials of torsion angles have also been used to refine structures generated from NMR data.[151, 152] Kuszewski et al incorporated a database-derived torsion angle potential into the target function for NMR structure refinement, resulting in a significant improvement in various quantitative measures of quality (Ramachandran plot, side-chain torsion angles, and overall packing).[151] In a similar way, Yang et al constructed a database of 2405 refined NMR structures.[152]

2.8 Improving Sampling and Scoring with Restraints

Due to their intrinsic inaccuracies, a common issue with energy functions is that incorrect conformations may be scored comparably to (or even better than) the native state,[110] lending the energy function inability to recognize the native state (Figure 2.6a). This issue be remedied by incorporating sparse experimental data as restraints, which offers some structural information that by itself is insufficient to completely determine the proteins structure (Figure 2.6b, c).

2.8.1 Sparse Experimental Data as Restraints

Restraints from sparse experimental data drastically decrease conformational space that needs to be sampled to only those structures consistent with the data. Many software suites implement algorithms to couple their *de novo* prediction methods with limited experimental

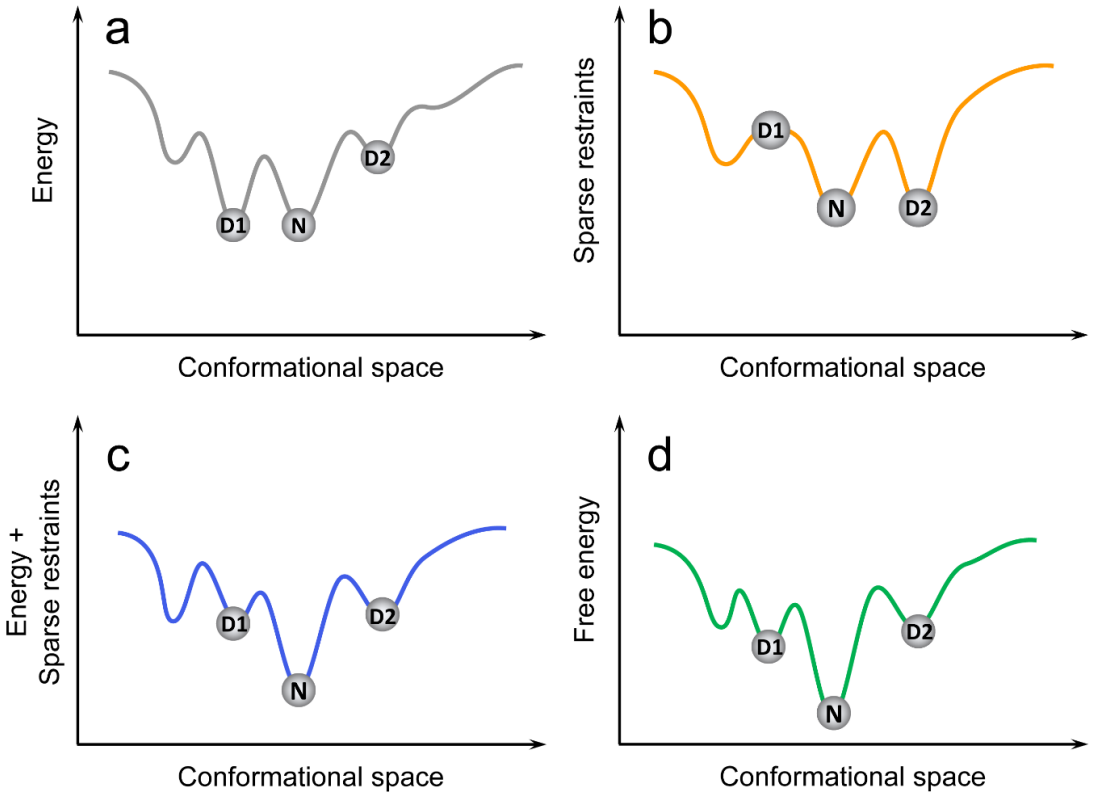

Figure 2.6: Cooperative effects of energy functions and sparse restraints on a hypothetical protein. a) The energy function has two comparable minima, lending itself the inability to tell decoy D1 from the native state N; b) A scenario where decoy D1 violates some restraints and is thus penalized by the restraint score. However, as sparse restraints by themselves are insufficient to completely determing the protein's structure, there exist decoys, such as D2, that satisfy the restraints as well as the native state N does; c) Adding sparse restraints to the energy function renders an energy surface closer to the protein's real free energy surface, such that only one global minimum corresponding to the native state exists; d) The real free energy surface of the protein.

data, including those from nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), cross linking mass spectrometry (XL MS), and electron microscopy (EM).

NMR rivals X-ray crystallography as a technique by which an entire protein structure can be unambiguously determined. Solution-state NMR can determine the structure of relatively small proteins (¡ 20 kDa), but intensive experimental techniques and analysis of NMR spectra are required to determine a high-resolution structure of a protein. Each residue typically requires upwards of 15 constraints. Oftentimes, NMR spectroscopy can provide some degree of low-resolution information about the global conformation of a protein, even for larger proteins [191, 192]. These sparse restraints, including chemical shifts (CSs), Nuclear Overhauser Enhancements (NOEs), and residual dipolar couplings (RDCs), do not provide enough information to fully determine the structure of a protein, but can be used in conjunction with computational PSP software. CSs provide information about the protein backbone conformation, while NOEs and RDCs give information about the global fold of the protein. *De novo* PSP software can take advantage of just CSs [193], CSs and NOEs [194], or all three types of restraints [195].

Site-directed spin labeling (SDSL) and EPR can be used to glean information about proteins of nearly any size in their native environments. In addition, only a small amount of sample is required for structural interrogation by EPR. The accessibility and mobility of the spin labels can be used to determine the exposure and topology of SSEs [196, 197]. Distances between spin labels can be detected up to 60 $\mathring{A}$, and can give insight into the overall fold of the protein as well as different conformational states [198, 199]. However, it is not feasible to use EPR to determine the full structure of a protein. EPR is rather experimentally intensive, as it requires the introduction of unpaired electrons into the protein. This generally requires mutating all cysteines to alanine or serine, followed by mutating the desired labeling sites to cysteine and coupling with spin labels. This technique will only give a small part of structural information about the protein, so these sparse EPR

data can be used in conjunction with computational protein structure prediction methods [200, 201, 202]. The selection of sites to spin label is integral to the efficacy of structure determination by EPR [200].

Similarly, XL-MS experiments can be used to determine inter-atomic distances that serve as experimental restraints. XL-MS can be used with proteins in their native states, and has proven to be compatible with relatively large proteins, flexible proteins, and membrane proteins [203, 204, 205]. In addition, the samples used can be heterogeneous and dynamic, as the output of XL-MS experiments is an average. The basis of XL-MS is the ability of two functional groups of a protein to form covalent bonds if they are within a certain distance of one another. These cross links can occur both inter- and intramolecularly. The proteins are then enzymatically digested, and MS is used to identify these cross links and surface labels [206, 207, 208].

EM provides data similar in format to that of X-ray crystallography, that is, a density map of a protein or complex. The data are thus less sparse than many of the aforementioned experimental techniques, but EM has historically provided lower-resolution density maps, from which an atomic structure cannot be gleaned. However, even low resolution EM density maps are integral for identifying the overall organization of large molecular complexes. In recent years, EM technologies have progressed such that density maps with resolutions in the range of 4 8 $\AA$ can regularly be attained, at which level SSEs can be visualized and even some side chain character can be visualized [209]. Many computational modeling methods have been developed that work with EM density maps, either in fitting previously solved structures into density maps, determining the topology and location of SSEs [210, 211], performing comparative modeling, and *de novo* protein structure prediction.

Most *de novo* protein structure prediction algorithms require the use of a segmented density map, which can be accomplished with the use of various segmentation algorithms [212, 213, 214]. Then, SSEs can be extracted from the density map either manually or with

the use of algorithms that automate the selection of helices and/or sheets from a segmented density map [210, 215, 216, 217]. Next, *de novo* modeling algorithms can use these data with the density map and primary sequence of the protein in order to create a full structural model either via optimization [218] or using Monte Carlo methods [219, 220, 221].

2.8.2 Predicted Contacts as Restraints

If no experimental restraints are available for the protein, secondary and tertiary restraints can be predicted from an amino acid sequence based on existing structures. Secondary structures can be predicted using machine learning methods. Artificial neural networks (ANNs) can be used to predict secondary structures from position-specific scoring matrices [222, 223], reduced amino acid representation [224], or multiple sequence alignments (MSAs) [225, 226]. Methods have also been developed specifically to predict membrane protein topology from amino acid sequence using ANNs [224, 227, 228], support vector machines (SVMs) [229], or Hidden Markov Models (HMMs) [230, 231].

It is a long-standing observation that three-dimensional protein folds can be predicted from sufficient information regarding the proteins inter-residue contacts [232, 233]; the addition of even relatively sparse information about tertiary contacts into an algorithm's scoring function can help improve protein models [234]. Recently, the incorporation of long range contact predictions has resulted in the most effective *de novo* protein structure prediction algorithms [235, 236]. Several algorithms have been devised to predict these contacts using the principle of correlated mutations. In general, amino acid contacts that stabilize the protein fold are assumed to evolve complementarily if one residue of a contact is mutated, the other will likely also mutate to a reasonable interaction partner.

In order to identify pairs of correlated mutations, amino acid pairs can be scored based on their physicochemical similarity using the McLachlan matrix [237], which is based on the frequencies of observed mutations in homologous proteins. Correlated mutations can

also be scored by mutual information between MSAs based on the equation

$$MI = \sum_{ab} f\left(A_i B_j\right) \log \frac{f(A_i B_j)}{f(A_i) f(B_j)} \tag{2.7}$$

The above equation indicates that the mutual information between two protein sites i and j is computed by summing over amino acid pairs ab for every amino acid type a and b, where $f(A_i B_j)$ is the observed relative frequency of ab at columns ij and $f(A_i)$ is the observed relative frequency of amino acid type a at position i. The identification of these correlated mutations is used in many methods of multiple sequence alignment [232, 238, 239, 240, 241] from which tertiary contact predictions can be extrapolated.

In recent years, numerous algorithms have come out that account for covariance caused by indirect inter-residue coupling effects, which has led to improvement in prediction of correlated mutations [241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254]. These methods are based on the observation that two residues aligned in multiple sequence alignments may exhibit statistical dependencies even though they are distant in physical space, which usually arises from chains of interacting pairs of residues. Also, information regarding the conservation of certain residues regardless of their tertiary contacts must be considered for correlated mutations to properly represent actual three-dimensional contacts. Many methods have been devised that decouple direct from indirect residue coevolution, primarily based on statistical methods. Covariation-based contact prediction has also proven successful as a scoring metric for *de novo* folding [244, 245].

Machine learning methods, including ANNs [255, 256, 257, 258, 259], genetic algorithms [260, 261], random forests [262], HMMs [263, 264], and SVMs [265, 266], have also arisen as successful methods to predict 3D contacts These methods use various features to predict contact maps. Some of the most successful of these machine learning methods for contact prediction are hybrid methods that predict contacts based on both physicochemical features and evolutionary features, using MSAs as part of their training data sets

[267, 268, 269, 270].

2.9 Examples of Methods for de novo Tertiary Structure Prediction

Protein structure prediction methods can be broadly grouped into template-based modeling, where construction of target models involves threading the target sequence through the structure of homologous proteins (templates), and *de novo* structure prediction, where target models are constructed from sequence alone, without relying on similarity at fold level between the target sequence and any of the known structures.[32, 137, 271, 272] Template-based modeling is based on the premise that tertiary structures of proteins in the same family are more conserved than their primary sequences.[273, 274, 275] While it can produce very accurate models for target sequences if templates with sequence identity ¿ 50

2.9.1 FRAGFOLD

FRAGFOLD was developed based on the rationale that proteins tend to have common structural motifs at the super-secondary structural level.[276, 277, 278] In FRAGFOLD, 3D models are built by assembling super-secondary structural fragments from a library of highly resolved protein structures with MC simulated annealing and evaluated with a knowledge-based energy function. FRAGFOLD was initially tested in CASP2,[276] CASP4,[277] and CASP5.[278] Its success in predicting the fold of NK-Lysin marked the first correct de novo blind prediction of a proteins fold.[276]

The super-secondary structural fragments considered include α-hairpin, α-corner, β-hairpin, β-corner, $\beta-\alpha-\beta$ unit, and split $\beta-\alpha-\beta$ unit. Favorable super-secondary structural fragments are selected based on the quality of threading. Threads that contradict the reliable regions of predicted secondary structure by PSIPRED [279] are skipped. In addition to this sequence-specific fragment list, a general fragment list that consists of all tripeptide, tetrapeptide, and pentapeptide fragments is also constructed from a library of highly resolved protein structures. The knowledge-based potential function in FRAG-

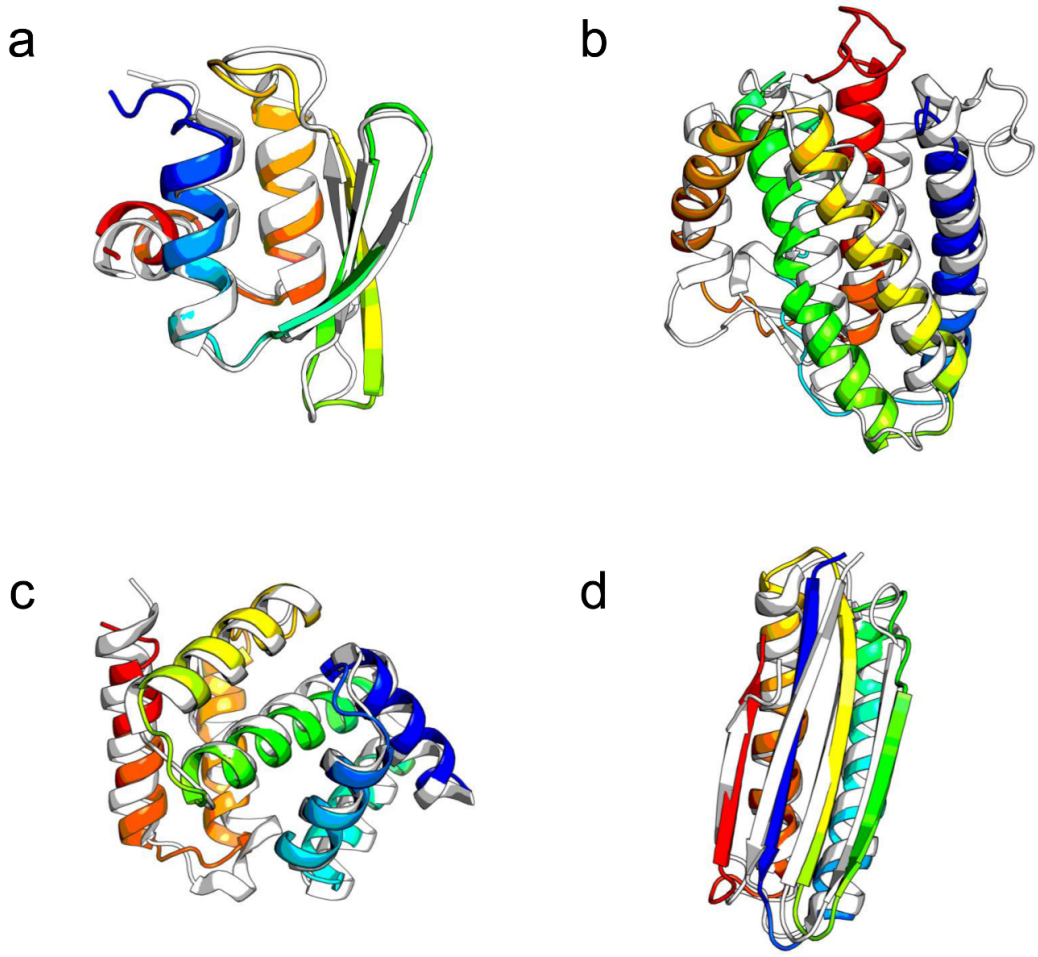

Figure 2.7: Highlights of *de novo* structure prediction in CASP experiments. Predicted structure models (rainbow) are superimposed with the crystal structures (grey). a) Rosetta-predicted structure model superimposed with a crystal structure (PDB code: 1whz) of CASP6 target T0281, hypothetical protein from *Thermus thermophilus* Hb8. This model is astonishingly close to the crystal structure with a C_α RMSD of 1.6 Å. b) I-TASSER-predicted structure model superimposed with a crystal structure (PDB code: 4dkc) for the CASP10 ROLL target R0007, interleukin-34 protein from *Homo sapiens*. c) Superimposition of a QUARK-predicted structure model with a crystal structure (PDB code: 5tf3) of the cASP11 target T0387, hypothetical protein YPO2654 from *Yersinis pestis*. This model has a C_α RMSD of 2.9 Åfrom the crystal structure. d) Superimposition of a BCL::Fold-predicted structure model with a solution NMR structure (PDB code: 2mq8) of CASP11 target T0769, a *de novo* designed protein LFR11 with ferredoxin fold. While this target is in the category template-based modeling, BCL::Fold assembled models for it without relying on any homologous templates.

FOLD initially consists of a set of pairwise potentials, a solvation potential, a term for penalizing non-compact folds, a term for penalizing steric clashes, and a term that accounts for hydrogen-bonding.[276] The total energy of a 3D model is computed according to:

$$E_{total} = W_1 E_{short-range} + W_2 E_{long-range} + W_3 E_{solvation} + W_4 E_{clash} + W_5 E_{h-bonding} + W_6 E_{compact} \tag{2.8}$$

where W_{1-6} are tunable weights, $E_{short-range}$ represents the sum of pair potentials for residue separation of 6 or less, $E_{long-range}$ represents the sum of pair potentials for residue separation of 7 or more. $E_{solvation}$ for a residue α is computed in a similar fashion:

$$E^{\alpha}_{solvation}(r) = -RT \ln \left[\frac{f^{\alpha}(r)}{f(r)} \right] \tag{2.9}$$

where r is the degree of residue burial characterized by contact number [280], $f^{\alpha}(r)$ represents the frequency of occurrence of residue α with burial r and $f(r)$ represents that frequency of occurrence of all residues with burial r. $E_{compact} = d_{max} - d_{cutoff}$ and penalizes non-compact folds defined as folds in which the maximum distance between any pair of C_{α} atoms is greater than a predetermined cutoff. This energy function was recently complemented with predicted contacts as a restraint term [281]. Jones and coworkers found that combining statistical potentials with predicted contacts by PSICOV[282] is significantly better than either statistical potentials or predicted contacts alone [281].

2.9.2 Rosetta

The Rosetta algorithm for *de novo* protein structure prediction employs MC simulated annealing to assemble protein-like 3D models from fragments of unrelated protein structures with similar local sequences using an energy function based on Bayes' theorem.[283, 78] The algorithm is based on the experimental observation that local sequence preferences bias, but do not uniquely determine, the local structure of a protein.[283] Rosetta has turned

out to be one of the most successful methods indicated by results from CASP experiments [79, 284, 285] and several other studies (see Figure 2.7a for an example).[286, 33]

Model construction in Rosetta is performed via a sequence of fundamental conformation modification operations termed fragment insertion. For each fragment insertion, a sequence segment of three or nine residues is selected, and the torsion angles of these residues are replaced with the torsion angles of a homologous fragment selected from a ranked list of fragments of known structure.[78] Fragment insertions that decrease the energy of the resulting conformation are accepted and those that increase the energy are accepted according to the Metropolis criterion.[82] Derivation of the Rosetta scoring function was based on a Bayesian separation of the total energy into components that describe the likelihood of a particular structure, independent of sequence, and those that describe the fitness of the sequence given a particular structure [78].

$$P(structure|sequence) = \frac{P(sequence|structure)P(structure)}{P(sequence)} \tag{2.10}$$

The original Rosetta scoring function is coarse-grained: terms corresponding to solvation and electrostatic effects are based on observed residue distributions derived from known protein structure databases, and hydrogen bonding is not explicitly described. However, preferences of β-strand pairing geometries and β-sheet patterns are included. Steric clashes are generally penalized, while van der Waals interactions are not explicitly modeled. A more physically realistic, atomic-level scoring function was developed later for applications requiring more detailed structurally information. In this "fine-grained" version of the energy function, van der Waals interactions are modeled with a 6-12 Lennard-Jones potential. Solvation effects are included, using the Lazaridis-Karplus model, and hydrogen-bonding is explicitly accounted for using a secondary structure- and orientation-dependent potential derived from high-resolution protein structures [287]. Energetics of local interactions are described using an amino acid- and secondary structure-dependent potential for

backbone torsion angles. The reader is referred to reference 240 for a mathematically more detailed description of the Rosetta scoring function.

2.9.3 I-TASSER

Recent CASP experiments have shown significant advantages of integrating various techniques such as threading, de novo modeling and atomic-level structure refinement approaches into a single pipeline of tertiary structure prediction.[22, 285, 288, 289, 290, 291] The I-TASSER method,[292, 293, 294] which implements TASSER296 in an iterative mode, is one example of the composite approaches. I-TASSER has been particularly successful as shown by recent CASP experiments (see Figure 2.7b for an example).[292, 293, 291, 69]

I-TASSER uses a sophisticated threading scheme, which compares the target sequence with template structures using profile-profile alignment, for selection of the most probable structure fragments. Aligned regions of the target sequence are modeled by connecting template fragments through a random walk of C_α-C_α bond vectors of variable lengths. Unaligned regions are simulated on a cubic lattice system for computational efficiency. Initial full-length coarse-grained models are refined via REMC simulation where two kinds of moves are implemented: off-lattice rigid fragment translations and rotations of the aligned regions and on-lattice 26 bond movements and multi-bond sequence shifts of unaligned regions.[295] The models of the first-round TASSER simulation are clustered and the cluster centroids are submitted to a second-round TASSER simulation to remove physically unrealistic interactions. Finally, backbone atoms and sidechain rotamers are added to the model with the lowest energy from the second round.[294] The energy function of I-TASSER includes the original TASSER knowledge-based potential and a new burial potential based on neural network-predicted accessible surface area (ASA).[294] The original TASSER potential consists of long-range pair interactions of sidechain centers of mass, local C_α correlations, hydrogen-bond, hydrophobic burial interactions, propensities for pre-

dicted secondary structures, protein specific pair potentials of sidechain centers of mass, and tertiary contact restraints extracted from the threading templates.[296]

2.9.4 QUARK

QUARK is an algorithm for *de novo* protein structure prediction using REMC simulations guided by a consensus knowledge-based energy function. In contrast with Rosetta and I-TASSER that assemble fragments of fixed sizes, QUARK assembles 3D models from small structure fragments of multiple sizes from 1 to 20 residues. To increase the structural flexibility and the efficiency of conformational search, QUARK also implements a set of movements consisting of free-chain constructions and fragment substitutions between decoy and fragment structures.[80] The QUARK algorithm has been shown to be highly successful in recent CASP experiments (see Figure 2.7c for an example).[80, 69]

QUARK generates structure fragments for target sequences by threading sequence segments through a library of non-homologous experimental structures. Multiple features such as solvent accessibility, real-value Φ and Ψ angles, and secondary structure types as predicted from back-propagation neural networks are used to improve generation of structure fragments. Optimization of 3D models is performed via REMC simulations that start with initial models assembled by chaining randomly selected fragments with varied sizes. Conformational sampling of each replica is done through residue-level, segment-level, and topology-level movements. After each running cycle, the conformations between every two adjacent replicas are exchanged according to the Metropolis criterion.[82] Protein structure models built by QUARK are evaluated by a composite knowledge-based energy function consisting of atomic-level pair potentials, hydrogen-bonding potential, SSE packing potentials, heuristic terms that account for excluded volumes, solvent accessibility, and radius of gyration.[80]

2.9.5 BCL::Fold

The BCL::Fold algorithm developed in our group seeks to overcome the limitations of protein size and fold complexity by assembling idealized secondary structure elements (SSEs) into 3D models, thereby facilitating the sampling of non-local contacts as shown by our benchmark study [161, 81]. Thus, BCL::Fold may be a promising tool for structure prediction of proteins with high contact order.[297, 298, 299] It's also worth mentioning, that in contrast to the other four methods, which heavily reply on the availability of homologous template structural fragments (short or long), BCL::Fold is "truly" *de novo* in the sense that no template structure is needed at any state of the algorithm. While BCL::Fold was not ranked among the most successful methods, we would still like to highlight the CASP11 target T0769. While this protein is in the category of template-based modeling, meaning that a suitable template can be identified that covers all or nearly all of the target, BCL::Fold predicted a model with a C_α-RMSD of 1.8 Åto the released solution NMR structure without relying on any homologous templates (Figure 2.7d).[162]

In BCL::Fold, the necessary complexity reduction of the sampling space is achieved by assembling SSEs from a predetermined pool of SSEs using a Monte Carlo simulated annealing algorithm and omitting more flexible loop regions. A high-quality pool of SSEs can be readily created using machine learning-based secondary structure prediction methods such as PSIPRED.[222] BCL::Fold implements a comprehensive list of SSE-based MC moves, which are categorized into six main categories: adding SSEs, removing SSEs, swapping SSEs, single SSE moves, SSE-pair moves, and moving domains consisting of multiple SSEs.[81] Models generated by BCL::Fold are evaluated by a knowledge-based consensus energy function called BCL::Score,[160] which consists of potentials residue pair interaction, residue environment, SSE packing, β-strand pairing, loop length, radius of gyration, contact order, secondary structure prediction agreement. Separate penalizing energy terms were also included to exclude conformations with clashes between amino acids or secondary structure elements and loops that cannot be closed.[160] BCL::Score

can also be complemented with experimental or predicted restraints to improve selection of native-like models.[186, 185, 300, 195]

2.10 Outlook

In the past decade, weve seen hardware and algorithmic advances that enabled researchers to perform millisecond timescale simulations of protein folding, and weve also seen development of methodologies that predicted tertiary structure with better accuracy for proteins with larger size. Despite these achievements, there is still a long list of challenges on the way toward a solution to the protein folding problem.

On the folding mechanism side, even though long simulations have been available, unambiguous scientific results learned from such simulations have thus far been modest and theories about how proteins fold have not yet converged.[49] First, it is still being debated whether proteins fold via a single definite pathway or multiple parallel pathways. Although both views have received support from simulations and experiments,[16, 17] additional simulations with more robust trajectory analysis and experimental validation are required to disambiguate conflicting results. Second, realistic folding simulations have thus far been limited to small proteins (¡ 100 residues), it is questionable whether folding mechanisms revealed by these simulations are generalizable to larger proteins. Thus, simulating the folding of larger proteins will likely be a major trend for the next decade. Finally, full integration of simulations to experimental workflows has been slow. Closer interaction between simulations and experiments such that simulations be tested by experiments and in turn aid in the interpretation of experimental results and guide the design of future experiments will have greater impact on the field.

On the structure prediction side, larger proteins, especially those with multi-domains, stay a significant challenge to *de novo* structure prediction methodologies. These proteins are often characterized by their high contact order and long folding time.[297, 301] Conformational sampling of these proteins is usually inefficient and is complicated not only by

protein size, but also by the considerable number of non-local contacts, which are formed by residues far apart in sequence but usually critical for structural stability. [302, 230] Consequently, tools for *de novo* structure prediction are not likely to become practically useful for structure prediction for any but very small, sometimes medium-sized proteins.[30, 32] Other challenging targets, especially for methods whose energy functions heavily rely on statistics extracted from known structures, may also include proteins with rare and unusual folds.[289] Accurate prediction of tertiary structure for these challenging targets certainly requires the joined forces of high-performance hardware, efficient algorithms for conformational sampling, accurate energy functions, and, last but not least, valuable experimental restraints.

Chapter 3

Optimization of EM-Fold

3.1 Introduction

Coupling computational protein structure prediction methods with experimental data greatly decreases the conformational space that computational techniques need to sample. The BioChemical Library (BCL), developed in the Meiler Lab at Vanderbilt University, can perform protein structure prediction either completely *de novo* or coupled with experimental restraints. The basic premise of the BCL protein folding pipeline is the assembly of SSEs. From a protein sequence, machine learning methods are used to predict a pool of SSEs. Each SSE is used as an entire entity in the subsequent assembly of these SSEs in three-dimensional space using a Monte Carlo Metropolis sampling algorithm with simulated annealing. The quality of each model is evaluated using a series of knowledge-based energy potentials [303]. These scoring functions are optimized for the type of protein being folded, and experimental data can be included at this stage.

The use of the BCL folding algorithm with EM data is called EM-Fold. When assembling models using EM-Fold, density rods that represent SSEs must first be gleaned from an experimental density map either manually or using various automatic segmentation algorithms [304, 305, 306, 307]. Next, structure prediction occurs via two main stages: assembly and refinement. In the assembly stage, these density rods are filled with SSEs from the pool. The Monte Carlo moves used during this stage include such large-scale movements as adding and removing SSEs from the model, swapping two SSEs, and flipping SSEs. The next stage, the refinement stage, consists of much smaller moves that keep each SSE within its density rod; these moves include small bends and rotations. Each potential model is scored not only in terms of the energy of the arrangement, but also its feasibility. Models are severely penalized if they do not fully occupy the density map and if it is not

possible for a loop to exist between subsequent SSEs [308, 309].

EM-Fold is currently optimized for medium-resolution (5 - 10 Å) density maps, where SSEs can be seen but no highly detailed information is present. The goal in this work is to optimize the algorithm for use with EM density maps in a higher resolution range (3 - 5 Å), where more side chain information is present, but a protein cannot be unambiguously traced through the density map. EM density maps are regularly being released in this resolution range.

3.2 Adding Side Chains

BCL models consist of just SSEs; no loop regions or side chain data is explicitly modeled. However, at higher resolutions, bulky side chains become apparent in density maps. The idea here is to create a way to represent side chains so as to optimize the fit of the model with less ambiguity.

3.2.1 The Cross Correlation Coefficient

The Pearson Correlation is a measure of the linear dependence between two variables. In EM-Fold, it is quantified by the Cross Correlation Coefficient (CCC) and is used to evaluate the fit of a model into a density map. At each Monte Carlo stage, a density map of each model is simulated using trilinear interpolation to the given resolution. Then, the experimental and simulated density maps are overlaid, and the CCC quantifies the fit.

$$CCC = \frac{k\sum_{x=0}^{x=k}\rho_S(x)\rho_E(x) - \sum_{x=0}^{x=k}\rho_S(x)\sum_{x=0}^{x=k}\rho_E(y)}{\sqrt{k\sum_{x=0}^{x=k}\rho_E(y)^2 - \left(\sum_{x=0}^{x=k}\rho_E(y)\right)^2}\cdot\sqrt{k\sum_{x=0}^{x=k}\rho_S(y)^2 - \left(\sum_{x=0}^{x=k}\rho_S(y)\right)^2}} \tag{3.1}$$

Here, ρ_S represents the density of the voxels of the simulated density map, while ρ_E represents the density of the voxels of the experimental density map.

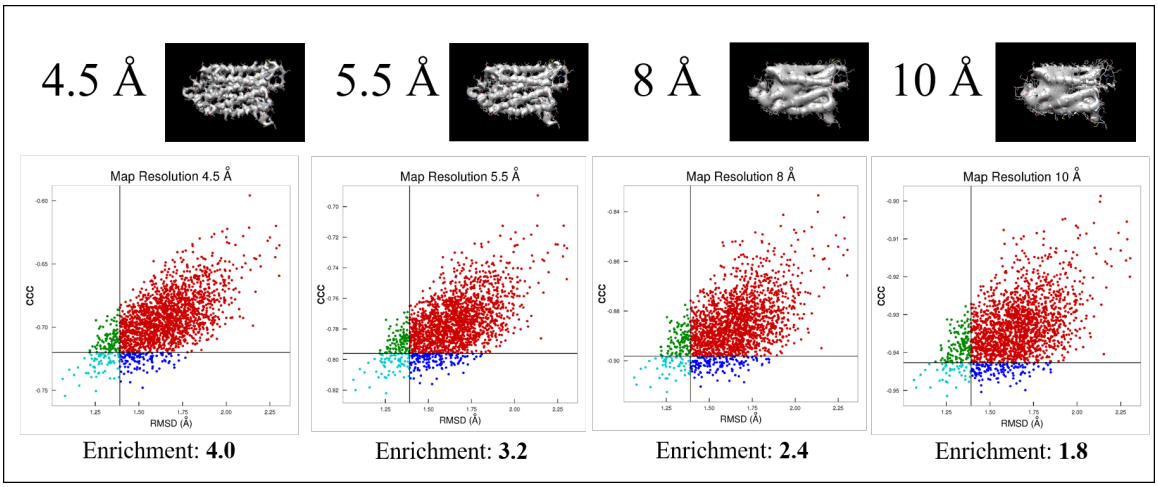

Figure 3.1: Enrichment achieved with CCC score. A) Density maps overlaid with atomic model of rhodopsin. Density maps simulated at 4.5 *Å* and 5.5 *Å* resolution, and maps blurred using Gaussian blur from experimental density map to 8 and 10 *Å* resolution. B) Enrichment of top 10% of CCC values for best quality models, i.e. models among top 10% by RMSD.

3.2.2 Relationship Between CCC and Model Quality

The CCC quantifies the agreement of each protein model with the provided experimental density map. The CCC correlates with model quality as quantified by root mean squared deviation (RMSD) normalized to 100 residues (Figure 3.1). As resolution improves, the enrichment of high quality models by prefiltering by CCC increases accordingly. We hypothesize that this is due to the extra level of information present at higher resolutions. Once side chain character starts to be seen in density maps, the CCC becomes increasingly relevant.

3.2.3 Side Chain Representation

Currently, atomic side chain coordinates are added using Rosetta to perform a monte carlo simulation following model refinement in the BCL, as BCL models have no explicit side chain representation. We plan to leverage the CCC score in the BCL at higher resolutions by simulating some side chain character in BCL models. The rotameric state of

each side chain is unknown, so modeling the atomic coordinates of each side chain is unfeasible. Modeling the entire side chain could also discount flexible amino acids, whose density will like be blurred. Rather, we plan to represent the side chain by modeling density corresponding to each side chain off each amino acid.

The masses of each side chain atom were summed for each amino acid to create a "superatom," which could be placed at any position off the protein backbone. The main option explored here were to place the superatom at the C_β position of each amino acid or at the center of mass of the average rotameric state of the amino acid. The average rotameric position was calculated using Roland Dunbrack's rotamer library [310].

3.2.4 Methods

In order to test the effects of adding side chain representation on overall model quality, the refinement stage of EM-Fold was run starting from the idealized crystal structure of a protein with loop regions removed. From this model, 50 models were created from 50 different seeds, resulting in a total of 2500 models. The Monte Carlo protocol was continued for either 2000 iterations or 400 unimproved iterations. The temperature varied over the course of minimization so that the fraction of accepted steps decreased linearly from 0.25 to 0.05.

A number of primarily α-helical proteins with experimental density maps available on the EM Data Bank (EMDB) [311] were used for testing the side chain addition method (Shown in Table 3.1). 1GAKA does not have an experimental density map, but is a small α-helical protein that was used in conjunction with a simulated density map for testing various parts of the algorithm. In addition to the experimental density maps, density maps were simulated from the crystal structure of each protein at both their experimental resolutions and at 4.5 Å.

Protein Name	Experimental Resolution	Length	Number of Helices	Number of Sheets
5A63B	467	3.4	9	1
5A63C	3.4	265	8	0
5A63D	3.4	101	6	0
3J6JA	3.6	97	8	0
3JAFA	3.8	342	6	13
3KTTB	4.0	513	21	2
1GZMA	6.9	331	7	0
1GAKA	-	141	6	0

Table 3.1: Proteins used

3.2.5 Results and Discussion

For all proteins with both experimental and simulated maps, adding either representation of side chains did not change overall model quality. Figure 3.2 shows the output for this procedure for 5A63B (a monomer of human Gamma-secretase) with a density map simulated at 4.5 Å. Even though there is a clear distinction between the CCC scores for models with side chains and without, this does not translate into a difference in model quality. One potential explanation for this is that the CCC is not weighted highly enough in the EM-Fold scoring function to actually have an effect.

In order to further examine whether adding side chains will actually have an effect on model quality, I examined one individual α-helix. I took one helix and individually applied each of the five EM-Fold refinement mutates to it. In Figure 3.3, the effects of rotating and translating the helix along its Z-axis can be seen. In both cases, the addition of side chains worsens the CCC of the model with the density map. The difference is greatest where the protein is farthest from its native state, i.e., has the highest RMSD. This indicates that the CCC can be a good metric for discriminating better quality models from lower quality ones.

However, the same principle was applied to the other three refinement mutates: bends, XY rotations, and Z flips, the results of which are shown in Figure 3.4. Unlike with Z rotations and translations, there is no clear relationship between model quality and the difference in CCC for bend and XY rotations. One flip of the helix along its Z axis reveals

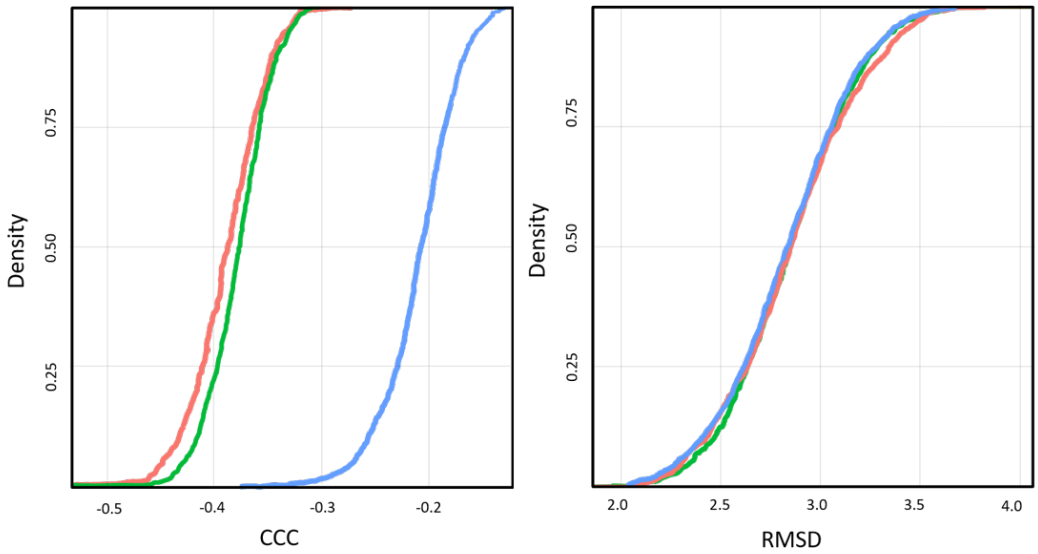

Figure 3.2: Effect of Side Chain on Model Quality. Blue represents models with no side chain information. The red curve represents models with all side chain density at C_β position, and the green curve represents models with all side chain density at the position of the average rotameric state of the amino acid. The y-axis represents the density of models with A) CCC and B) RMSD better than that value.

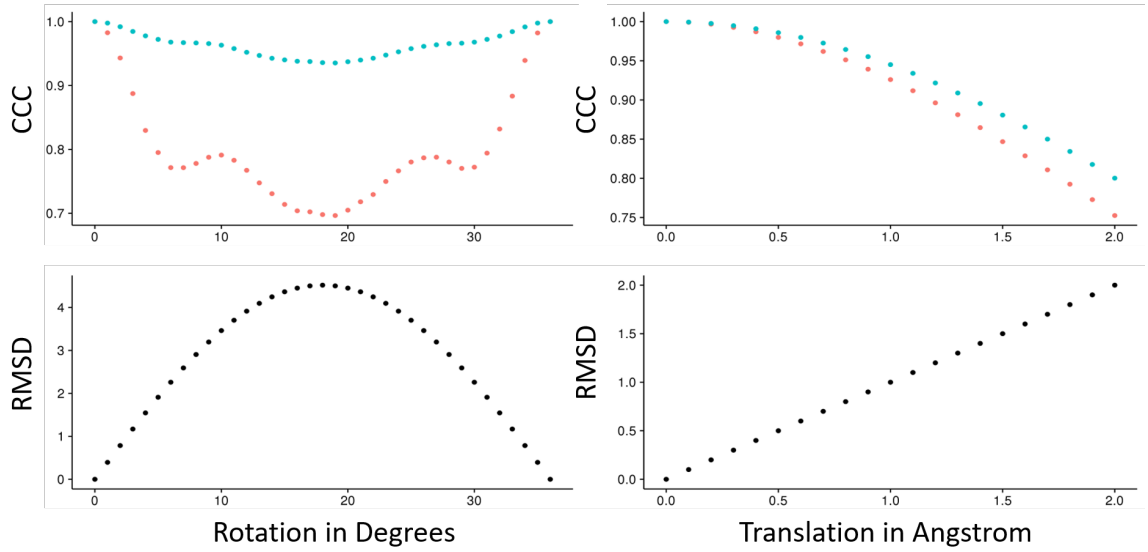

Figure 3.3: Effects of Side Chains on Rotation and Translation. Single α-helix A) rotated and B) translated from its native position. Top panel: blue curve represents models with no side chains, and the red curve represents models with side chain density at C_β position. Lower panel shows the corresponding RMSD for each model.

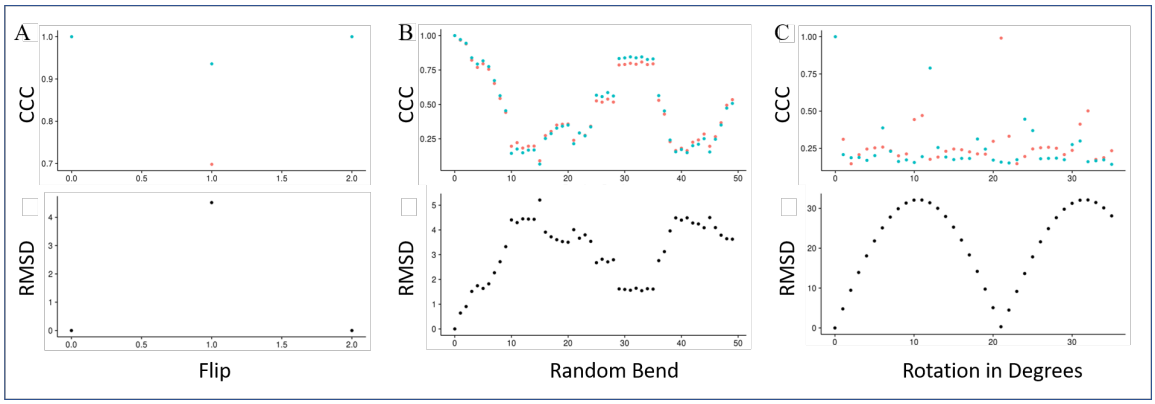

Figure 3.4: Effects of Side Chains on Other Refinement Mutates. A single helix was A) flipped along the Z axis, B) bent randomly up to 5 rad, and C) rotated along its XY-axes. Top panel shows the CCC with the density map, with the blue points representing models with no side chains, and the red points representing models with side chain density at the Cb position. Bottom panel plots the RMSD of the model.

the expected trend of a much lower CCC corresponding with poor model quality. These data indicate that although adding side chains may assist in creating high quality models when Z rotations and translations are used, this is not necessarily the case with all the mutates used in the refinement protocol, which may indicate that the current sampling procedure needs to be updated.

3.3 Implementation of Multistage Refinement

Based on the observation that adding side chains only improves model selection when Z rotations and translations are used, we propose a multistage refinement protocol. In the first stage, we aim to orient the general shape of a helix within a density rod by using all refinement mutates. The next stage will consist only of Z rotations and translations and will include side chain density at the C_β position. The purpose of the second stage is to properly orient the helices once they have been placed in the proper density rods.

3.3.1 Methods

The proteins 1GZMA, 1GAKA, and 5A63B (described in Table 3.1) were used to study the effects of the two stage refinement protcol. For each protein, a series of ten different starting models were produced by running the BCL assembly protocol with the protein's native pool. Each outputted model was inspected visually to ensure that the SSEs were in the correct general position with respect to one another, so they would be appropriate models to refine with EM refinement. Starting models had RMSDs varying from 3 - 11 Å. Different quality starting models were used to prevent the possiiblity of differing effects depending on this quality. For instance, it is possible that if a starting model is very close to the native position, then making these changes won't help improved the model in any measurable way.

Density maps of each protein at a resolution of 4.5 Å were simulated from the native model and used as restraints for EM refinement. Each starting model was put through two rounds of refinement. The first round was identical to the original EM Refinement: all five refinement mutates were used, and models had no side chain information. After the first stage, the model with the best RMSD was selected for further refinement in the second stage, which consisted of only rotations and translations. I tested this second stage both with and without any side chain representation.

3.3.2 Results and Discussion

An instance of the results of using the two stage refinement protocol can be seen in Figure 3.5. For each protein tested, 10 different starting models of various qualities were used to begin the refinement protocol. All starting models and proteins tested revealed the same general trend. The second stage of refinement increased model quality and the range of quality of models produced. However, there was no discernible difference between models in the second stage with and without side chains. This trend held true for all ten

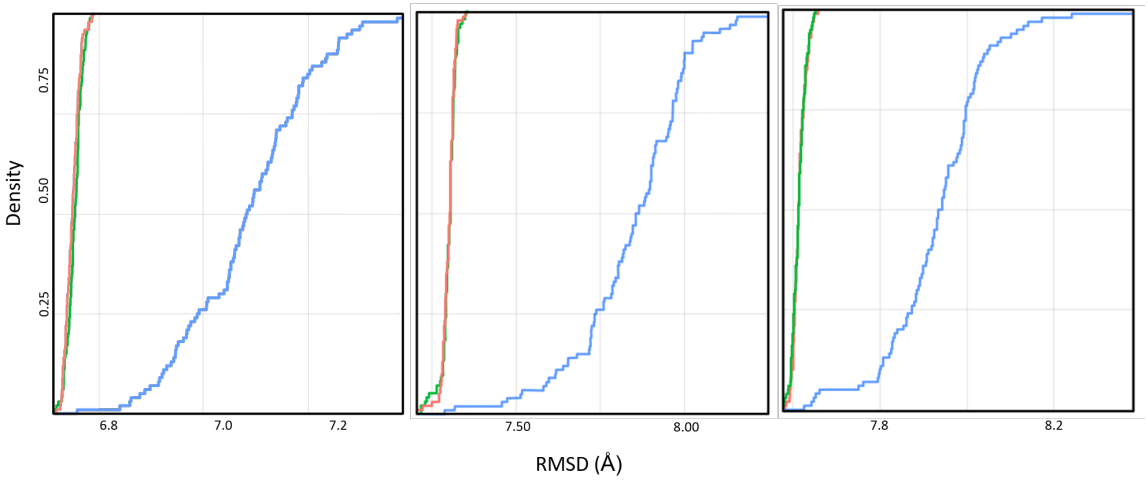

Figure 3.5: Two Stage Refinement Protocol. The two-stage refinement protocol was implemented for A) 1GAKA, B) 1GZMA, and C) 5A63B. In each graph, the blue curve represents the models produced by the original refinement protocol, the green curve represents models with two stages of refinement, and the red curve represents models produced by two stages of refinement with the addition of side chain information.

starting models for all three proteins.

The observation that addition of side chains has no effect on quality of the models produced by refinement using just rotations and translations was surprising due to the previous observation that adding these side chains to a single helix resulted in drastic differences in model quality. We hypothesized that this discrepancy was due to the low weight of the CCC score in each model's overall energy score.

3.4 Optimization of CCC Score

There are many different terms that go into computing the overall BCL energy of each model in EM Refinement; these terms and their weights are summarized in Table 3.2. The CCC has a weight of only 1. Excluding the penalty scores, this score makes up only about 1% of the total score, depending on the number of SSE prediction methods used. The hypothesis here is that increasing the weight of the CCC score will result in the models with side chains producing better quality models.

Scoring Term	Weight
*AA Clash	500
AA Distance	0.35
AA Neighbor Count	50
Loop Length	10
*Loop Closure Gradient	50000
Radius of Gyration	5
Contact Order	0.5
*SSE Clash	500
SSE Packing	8
Strand Packing	20
SSE Prediction Methods	1
CCC	1

Table 3.2: EM Refinement Scoring Terms. Terms denoted with an asterisk are considered penalty terms, and are weighted so that any models that do not adhere to these penalties will be exempt.

3.4.1 Methods

Here, only the second stage of refinement was studied, as this stage was where the discrepancy between models with and without side chains was minimal. The same proteins were used as were used for examining the effects of the two stage refinement (1GAKA, 1GZMA, 5A63B). The ten models that had the best RMSD after the first stage of refinement were used as starting models. Again, density maps were simulated from the native model at a resolution of 4.5 Å. Models were scored using three different scoring weights: the original scoring function with a CCC weight of 1, weighting the CCC at 1000, and using solely the CCC to evaluate model quality.

3.4.2 Results and Discussion

The results of adjusting the weight of the CCC score were varied both between models of the same protein and between different kinds of proteins. In some cases, increasing the weight of the CCC score improved the quality of outputted models with side chains; an example is shown in Figure 3.6 A-C. However, in other cases, increasing the weight of

the score made little difference (Figure 3.6 D-F). The fact that increasing the CCC weight makes a difference in some cases supports the hypothesis that adding side chains to BCL models can help improve model quality based on the fit of models within density maps. Further investigation is required as to why only some cases see an improvement in outputted model quality. In addition, further studies can be done to optimize how strongly the CCC score should be weighted.

3.5 Conclusion

Adding side chains to BCL models has shown some promise in improving models produced by EM fold in the 3-5 Årange. In particular, using a two stage refinement protocol leads to an improvement in model quality, where the first stage serves to orient a helix within a density rod, and the second optimizes the helix's orientation. Further work needs to be done to enhance both the sampling and scoring functions used in EM-Fold refinement. In addition, EM-Fold is currently based on positioning α-helices within density rods, but β-strands also start to appear at higher resolutions. An algorithm for positioning β-strands within EM density maps should eventually be developed as well.

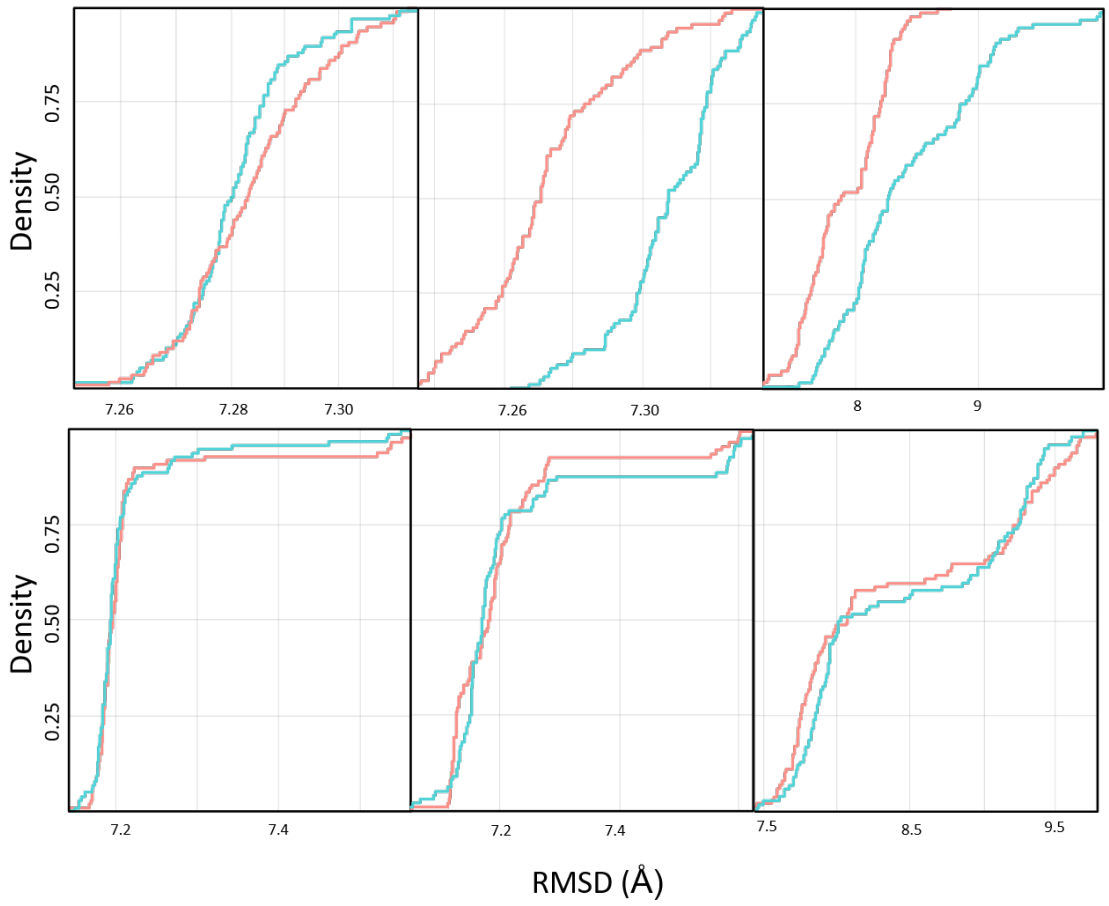

Figure 3.6: Optimizing Weight of CCC Score. An example of changing the weight of the CCC score while folding 1GAKA with EM-Fold refinement. Each row respresents a different starting model. Red curves represent models with side chains, whereas blue curves represent original BCL models. Models were scored using the original CCC weight (A,D), a weight of 1000 (B,E), or only with CCC (C,F). With the first starting model, increasing the weight of the CCC score leads to the improvement of models with side chains over original models.

Chapter 4

Development of Fragment-Based Topology Score

4.1 Introduction

The basic premise of the complete *de novo* protein folding pipeline with the BCL is shown in Figure 3.1. Clustering is currently done by first creating a distance matrix comparing each produced model by RMSD. Next, K-means clustering is performed based on this distance matrix, and the medoids of each cluster are selected as for further refinement. Each medoid is intended to be a representative model for the topology represented by all the models in each cluster.

4.1.1 K-Means Clustering

K-means clustering is arguably the most widely used tool for cluster analysis. Its purpose is to divide a set of data points $x = (x_1, \ldots, x_n)$ into sets $S = S_1, \ldots, S_k$ such that intra-class similarity is mazimized, but inter-class similarity is minimized. The Euclidean distance between each point and the cluster mean is used to measure this similarity, i.e. the term

$$\sum_{i=1}^{k} \sum_{x \in S} (x - \mu_i)^2 \tag{4.1}$$

where μ_i is the mean of each cluster k is minimized. Typically, k-means clustering is performed in an iterative manner. The basis of performing the clustering in this manner is that minimizing this term is asymptotic, that is, S will converge to a single paritioning [312]. Iterative k-means clustering begins with a random selecting of k points to represent cluster means. Next, each point is assigned to a cluster based on the mean to which it has the lowest Euclidean distance. The mean of each cluster is subsequently recalculated, and this process is repeated until the partitioning converges.

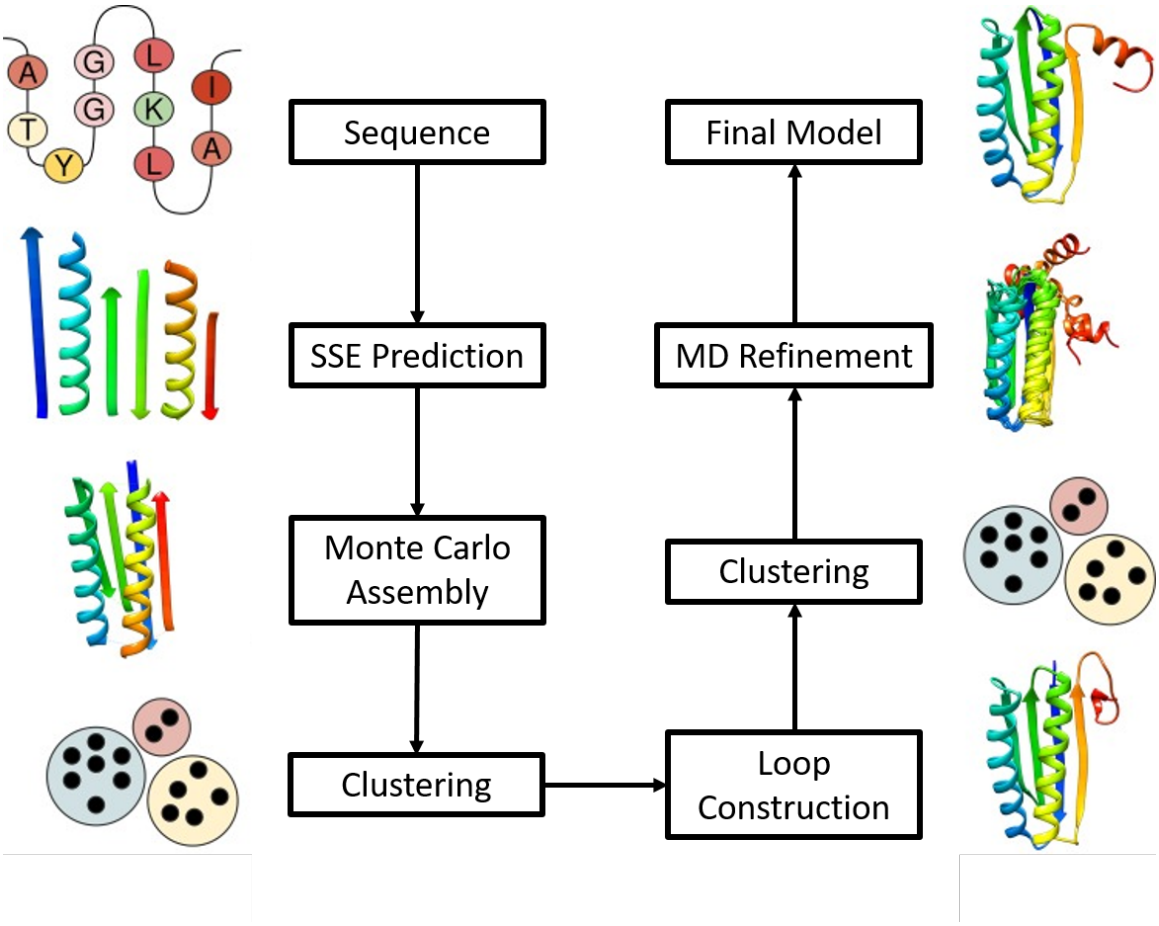

Figure 4.1: De Novo Folding Pipeline. The overall protocol for folding, from a protein sequence to a final three-dimensional model, is outlined here. Clustering steps follow the Monte Carlo assembly step and the loop construction step.

K-means clustering is intuitively a good way to approach partitioning data sets, as it is robust and easily progrmamable. However, k-means clustering is highly dependent on two parameters: the number of clusters, which is user-inputted, and the initial mean selection. Having a set number k means that the user must know or expect something about the data set prior to analysis. The fact that clustering depends on initial random selection means that multiple outcomes can occur. In addition, the presence of outliers can skew the partitioning due to the algorithm's reliance on calculating means.

4.1.2 Partitioning Around Medoids

An alternative to traditional k-means clustering is paritioning around medoids (PAM) clustering, which is more robust to outliers and noise in the data. PAM also accepts a dissimilarity matrix as input. PAM clustering is performed in essentially the same way as k-means, but the distance to the cluster medoid rather than mean is minimized. In addition, the distance metric used is the Manhattan distance, so the term

$$\sum_{i=1}^{k} \sum_{x \in S} \|x-]mu_i\| \tag{4.2}$$

is minimized. However, PAM clustering maintains the errors of k-means clustering in that the resulting clustering is dependent on both the user-inputted k value and the initial partitioning of the data. Two ways to get around these flaws are to run the clustering with many different values of k, and to run the clustering many times so all potential clusterings will be sampled.

4.1.3 Quantifying Clustering

There are many existing methods for comparing the structcures of two different protein models, all of which have different intentions and basic premise. Structure comparison methods are generally organized into two main classes based on whether or not they require

the initial superimposition of the two protein models. Of the superimposition-depenent methods, RMSD is likely the most popular. RMSD is calculated by the following equation

$$RMSD = \sqrt{\frac{1}{n} sum_{i=1}^{n} d_i^2} \tag{4.3}$$

where n represents the number of pairs of equivalent atoms, and d_i is the distance between the atoms in the i-th pair. RMSD is a standard metric for quantifying model similarity, but it is also extremely sensitive. RMSD calculation requires an all-atom alignment of the two models and compares each individual atom, so even if two models have the same overall topology, a small movement of one SSE could correspond with a relatively drastic increase in RMSD. Other popular methods that rely on structural superposition are the Global Distance Test (GDT) and Longest Continuous Segment (LCS). GDT looks solely at a protein's α carbon atoms, and it calculates the proportion of residues that can be aligned under various distance cutoffs., whereas LCS calculates the proportion of continuous residues that can be aligned under various RMSD cutoffs. GDT and LCS are more robust than RMSD to small perturbations in protein structure.

Superimposition-independent structure comparison methods often rely on quantifying the contacts made within a protein. These methods have differing ways of representing the points that are in contact as well as the actual definition of a contact. Points can either be individual atoms, fragments of each individual residue, or a coarse-grained model where a point represents an entire residue. Contacts can be defined either by various distance cutoffs or by physics-based contact definitions [313]

4.1.4 Clustering in BCL:Fold

The BCL *de novo* protein folding pipeline is outlined in Figure 3.1. There are two stages of clustering: after model assembly, and after loop construction. In both of these stages, PAM clustering is performed on the thousands of models that are outputted. Then,

the medoid of each cluster is selected as a representative topology of that cluster. The idea behind clustering in this manner is a reduction of sampling space. After assembly, thousands of models are outputted, but many will have the same overall topology that corresponds to a local energy well. Only one model of each topology is necessary to continue on in the pipeline, as more precise refinement is done in a later stage. A similar approach is taken after constructing loops for each representative model. Although the medoid of each cluster may not necessarily represent the cluster's lowest-energy model, it is enough that it represents a potential general topology. Here, topology is defined as the relative arrangement of SSEs.

Currently, a dissimilarity matrix is created at each of these clustering stage by comparing the RMSD of each model. Again, RMSD calculation requires an all-atom alignment of the two models and compares each individual atom, so its primary drawback is the amplification of minute differences. Rather, some sort of metric that can compare two topologies without counting each atom would be ideal to use at these stages to compare topologies.

Since a BCL model consists only of SSEs with no loops nor side chains, only the relative arrangement of SSEs needs to be looked at. One method would be to compare two SSEs of a model at their midpoints, from which you can glean the interaction distance, interaction strength, and twist of the SSEs with respect to one another. However, I propose to break each SSE into fragments. The comparison between the fragments of SSEs can provide a deeper layer of information of how the SSEs interact. Instead of only looking at midpoint interactions, looking at fragment interactions can show more dteail about how two SSEs interact, rather than only if they interact.

4.1.5 Interaction Weight

When quantifying the interaction between two SSEs, many different factors can be calculated, including the distance between the two, the twist angle, and the weight of the interaction. Here, the interaction weight is chosen as the metric that determines the pres-

ence of a contact between two SSEs. Interaction weight is essentially the the deviation of SSEs from being perfectly orthogonal. First, a line representing the shortest connection between the two SSEs is found. The angle between this line and the primary axis of each SSE is subseqeuently calculated.

4.2 All Amino Acid Matrices

4.2.1 Methods

In the BCL, SSEs are further represented by a vector of the overlapping fragments that make up the SSE. Alpha helices are broken down into 5 amino acid fragments, whereas Beta strand fragments are 3 amino acids long. In the first iteration of creating a topology scoring metric, a matrix was created for each of the two proteins being compared. The x and y axes of each matrix are identical; each consists of the entire amino acid sequence of the protein. First, a protein's matrix is filled in by iterating over its SSEs. Then, for each pair of SSEs, the algorithm proceeds by iterating over the fragments of each SSE. If the two fragments interact, the pair is given a score based on its interaction weight. This weight is filled in the appropriate spot in the matrix, where each fragment is represented by its central amino acid. Then, the actual topology similarity score between the two proteins is given by the Frobenius norm between the two matrices:

$$\| A - B \| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \left(a_{ij} - b_{ij} \right)^2} \tag{4.4}$$

where a_{ij} and b_{ij} represent all elements of matrices A and B, respectively. A protein-protein dissimilarity matrix can then be generated based on these scores, from which clustering can be performed.

Metric	Value
Rand index	0.905
Precision	0.58
Recall	0.78
F	0.67
Matthews Correlation Coefficient	0.62
Fowlkes-Mallows Index	0.67

4.2.2 Results

The primary goal of this topology metric is to improve model clustering. For this purpose, three different model proteins were used. These proteins were chosen because they are small and helical, so ideally all possible topologies will be explored as the models are created. For each protein, 1000 models were created. Incomplete models were filtered out. The initial clustering revealed different trends for each protein; in general, though, clustering with topology score did not show an improvement on clustering with RMSD (Figure 3.2). One hypothesis for this poor clustering trend is that certain topologies might not be similar to any others, and thus be in their own cluster. A filtering protocol was put into place to remove this possibility. A critical value was determined halfway between the maximum and minimum topology scores, and any models that had no value lower than this critical value were removed as they were determined to not be similar enough to any others. Filtering in this way gave marginally better clustering in some cases, but still did not reveal an overall improvement. Clustering was quantified by both silhouette score and Davies-Bouldin index.

An alternative to evaluating clustering accuracy by silhouette score or Davies-Bouldin index is by using internal clustering. 70 models of 1ULRA were manually clustered into 8 different groups based solely on visual inspection, and compared to how they were clustered based on k-means clustering by topology score. Several metrics were then calculated, which are summarized in Table 4.1. Each metric has a value between 0 and 1, where 1 represents clustering that matches perfectly with the manual clustering. Values greater than

0.5 indicate that there is some correlation between the two clustering definitions. Based on these values, this topology score gives moderately good clustering data, even though it is not necessarily good enough to cluster all models created.

4.3 Topology by SSE Mapping

Another hypothesis for the lack of success of topology scoring by creating all-amino acid matrices is that the SSEs of various models may not match up exactly. In this case, if SSE definitions are shifted by even one or two amino acids, then their interactions won't be correctly quantified. Rather, a more accurate way of calculating SSE interactions would be to map the SSEs of one model to another, and subsequently look at the fragment interactions of corresponding SSE pairs.

4.3.1 SSE Mapping

SSEs are mapped from one model to the next based on the Q3 score. Pools of SSEs are created from both the template and model protein and are compared pairwise. The Q3 score is calculated as

$$Q3 = \frac{100 \cdot \text{correct number}}{\text{total number}} \tag{4.5}$$

where correct number is the number of residues in both SSEs, and the total number represents the number of residues that are in the first SSE definition but are excluded from the second. When the SSE mapping is calculated, an SSE of a model is mapped to its counterpart SSE in another model based on to which SSE it has the highest Q3 score. If no maximum score is found, then the SSE is excluded from the mapping.

4.3.2 Methods

The SSEs of protein model A are first mapped to SSEs of protein model B using the previously described technique. Next, a matrix is created for each protein model. Each

matrix is a square matrix with dimensions the size of whichever model has more SSEs. An interaction score between two SSEs is then calculating by considering the fragments of each SSE as members of a disjoint set. The two SSEs then make up a bipartite graph, where each fragment represents a vertex, and two fragments are connected if they have an interaction score of at least 0.5. Then, the density of each bipartite graph is calculated by

$$d = \frac{e}{V_a} V_b \tag{4.6}$$

where e is the number of edges and V is the number of vertices in SSEs a and b.

The density score for each SSE pair is inputted into the appropriate matrix entry for each protein. Then, the two models are compared again by calculating the Frobenius distance between the two matrices. Models are then clustered based on these values.

4.3.3 Results and Discussion

Clustering based on SSE mapping in this way did not lead to much of an improvement over clustering based on RMSD (Figure 3.2). Even though the two topology metrics tested did not make a significant improvement on clustering, they were able to reach similar performance levels as attained by RMSD. However, these methods are based on much more intuitive metrics of comparing the overall topologies of two protein models. In addition, these methods are still in their basic stages, and they have room for many more improvements, whereas it is difficult to make RMSD a more accurate indicator of protein topology. Therefore, these results are promising that the topology score can be improved from the performance of RMSD to ever more accurate comparisons of topology.

Figure 4.2: Performance of Topology Metrics. The blue curve represents models clustered by RMSD, green represents those clustered by all amino acid matrices, and red represents those clustered based on SSE mapping. All analyses were performed on a filtered dataset. The x-axis shows the results for varying numbers of clusters k, with the corresponding silhouette score on the y-axis.

Chapter 5

Conclusion

The field of *de novo* protein structure prediction is a notoriously challenging task, as it is an immense challenge to both sample the vast conformational space of a protein and create scoring functions that emulate those of nature. The addition of experimental restraints helps to reduce conformational space, and the first portion of my research dealt with utilizing such restraints from cryo-EM at resolutions of 3 - 5 Åin conjunction with the BCL EM-Fold algorithm in order to improve the sampling and scoring functions of the algorithm. Next, I aimed to create a way to quantify similarities in protein topologies based on comparing fragments of SSEs.

The majority of currently known protein structures in the PDB have been determined by experimental techniques like X-ray crystallography, NMR, or EM. However, these techniques are not suitable for every protein, and it is also time- and labor-intensive to fully determine a protein structure using solely these techniques. Thus, there is a need for computational techniques to complement experimental techniques, and also ideally to predict structures completely *de novo* from the massive amount of sequence data available. BCL::Fold uses the unique sampling technique of Monte Carlo sampling of entire SSEs, which makes it a promising technique for proteins with higher contact order. Improving the clustering by topology of the BCL *de novo* folding pipeline will greatly improve its efficacy. In addition, many vital proteins, including the majority of drug targets, are membrane proteins, which are amenable to imaging by cryo-EM. As cryo-EM techniques and the resolution of the outputted images improve, more and more density maps of both MPs and soluble proteins will be available. EM-Fold is an automated technique to determine protein structures from these density maps, and it needs to be tailored for higher resolution structures.

The long-term goal of this project is *de novo* protein structure prediction, but the field of computational structural biology has a long way to go before attaining this ideal. Many different groups across the globe focus on their own methods of tertiary structure prediction, each with their own strengths and weaknesses. As each individual method develops, the collaborative nature of scientific research can ideally harness the strengths of each technique and continue the quest toward *de novo* tertiary structure prediction.

BIBLIOGRAPHY

[1] Westbrook J. Feng Z. Gilliland G. Bhat T.N. Weissig H. Shindyalov I.N. Berman, H.M. and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235 – 242, 2000.

[2] Goh K.I. Cusick M.E. Barabasi A.L. Yildirim, M.A. and M. Vidal. Drug-target network. *Nat Biotechnol*, 25(10):11119–26, 2007.

[3] Borwn A. Toots J. Scheres S.H.W. Amunts, A. and V. Ramakrishnan. The structure of the human mitochondrial ribosome. *Science*, 348:95–98, 2015.

[4] Myasnikov A.G. Natchiar S.K. Khatter, H. and B.P. Klaholz. Structure of the human 80s ribosome. *Nature*, 520:640–645, 2015.

[5] Cao E. Julius D. Liao, M. and Y. Cheng. Structure of the trpv1 ion channel determined by electron cryo-microscopy. *Nature*, 504:107–112, 2013.

[6] Yan C. Yang G. Lu P. Ma D. Sun L. Zhou R. Scheres S.H.W. Bai, X. and Y Shi. An atomic structure of human gamma-secretase. *Nature*, 525:212–217, 2015.

[7] K.A. Dill, S.B. Ozkan, M.S. Shell, and T.R. Weikl. The protein folding problem. *Annu Rev Biophys*, 37:289 – 316, 2008.

[8] H.S. Chan and K.A. Dill. The protein folding problem. *Physics Today*, 46:24 – 32, 1993.

[9] K.A. Dill and J.L. MacCallum. The protein folding problem, 50 years on. *Science*, 338:1042 – 6, 2012.

[10] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181, 1973.

[11] J.N. Onuchic and P.G. Wolynes. Theory of protein folding. *Curr Opin Struct Biol*, 14:70–5, 2004.

[12] C.M. Dobson and M. Karplus. The fundamentals of protein folding: Bringing together theory and experiment. *Curr Opin Struct Biol*, 9:92 – 101, 1999.

[13] C.M. Dobson, A. Sali, and M. Karplus. Protein folding: A perspective from theory and experiment. *Angew Chem Int Edit*, 37:868 – 893, 1998.

[14] G.R. Bowman, V.A. Voelz, and V.S. Pande. Taming the complexity of protein folding. *Curr Opin Struct Biol*, 21:4 – 11, 2011.

[15] A.I. Bartlett and S.E. Radford. An expanding arsenal of experimental methods yields and explosion of insights into protein folding mechanisms. *Nature Structual and Molecular Biology*, 16:582 – 588, 2009.

[16] S.W. Englander and L. Mayne. The nature of protein folding pathways. *Proceeding of the National Acedmy of Sciences of the United States of America*, 111:15873 – 15880, 2014.

[17] P.G Woynes. Evolution, energy landscapes, and the paradoxes of protein folding. *Biochimie*, 119:218 – 230, 2015.

[18] S.E. Radford. Protein folding: Progress made and promises ahead. *Trends in Biochemical Sciences*, 25:611 – 618, 2000.

[19] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem*, 48:545 – 600, 1997.

[20] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein folding – a perspective from simple exact models. *Protein Sci*, 4:561 – 602, 1995.

[21] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteski, A.E. Dawid, and A. Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116:7898 – 7936, 2016.

[22] C.H. Tai, H. Bai, T.J. Taylor, and B. Lee. Assessment of template-free modeling in casp10 and roll. *Proteins*, 82:57 – 83, 2014.

[23] J. Moult, J.T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23, 1995.

[24] C. Levinthal. Are there pathways for protein folding. *Journal de Chimie Physique*, 65:44 – 45, 1968.

[25] P.S. Kim and R.L. Baldwin. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem*, 51:459 – 89, 1982.

[26] P.S.. Kim and R.L. Baldwin. Intermediates in the folding reactions of small proteins. *Annu Rev Biochem*, 59:631 – 60, 1990.

[27] H. Frauenfelder, S.G. Sligar, and P.G. Woylnes. The energy landscapes and motions of proteins. *Science*, 254:1598 – 603, 1991.

[28] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21:167 – 195, 1995.

[29] K.A. Dill and H.S. Chain. From levinthal to pathways to funnels. *Nat Struct Biol*, 4:10 – 19, 1997.

[30] D.T. Jones. Progress in protein structure prediction. *Curr Opin Struct Biol*, 7:377 – 87, 1997.

[31] Y. Zhang. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol*, 18:342 – 8, 2008.

[32] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93 – 6, 2001.

[33] P. Bradley, K.M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309:1868 – 71, 2005.

[34] D.E. Kim, B. Blum, P. Bradley, and D. Baker. Sampling bottlenecks in de novo protein structure prediction. *J Mol Biol*, 393:249 – 260, 2009.

[35] M. Karplus and G.A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347:631–9, 1990.

[36] W.F. van Gunsteren and H.J.C. Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perpectives in chemistry. *Angewandte Chemie International Edition in English*, 29:992 – 1023, 1990.

[37] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646 – 652, 2002.

[38] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proc Natl Acad Sci USA*, 102:6679 – 85, 2005.

[39] J. Gumbart, Y. Wang, A. Aksimentiev, E. Tajkhorshid, and K. Schulten. Molecular dynamics simulations of proteins in lipid bilayers. *Curr Opin Struc Biol*, 15:423 – 431, 2005.

[40] E. Lindahl and M.S. Sansom. Membrane proteins: Molecular dynamics simulations. *Curr Opin Struct Biol*, 19:425 – 31, 2009.

[41] J.L. Klepeis, K. Lindorff-Larsen, R.O. Dror, and D.E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol*, 19:120 – 127, 2009.

[42] J.D. Durrant and J.A. McCammon. Molecular dynamics simulations and drug discovery. *BMC Biol*, 9:71, 2011.

[43] X. Periole. Interplay of g protein-coupled receptors with the membrane: Insights from supra-atomic coarse grain molecular dynamics simulations. *Chemical Reviews*, 117:156 – 185, 2017.

[44] A.R. Leach. Molecular modelling: Principles and applications. *Pearson Education*, 2001.

[45] D.C. Rapaport. The art of molecular dynamics simulation. *Cambridge University Press*, 2004.

[46] M. Gruebele. Protein folding: the free energy surface. *Curr Opin Struct Biol*, 12:161 – 8, 2002.

[47] Y. Duan and P.A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740 – 4, 1998.

[48] B. Zagrovic, C.D. Snow, M.R. Shirts, and V.S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol*, 323:927 – 37, 2002.

[49] T.J. Lane, D. Shukla, K.A. Beauchamp, and V.S. Pande. To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr Opin Struc Biol*, 23:58–65, 2013.

[50] D.E. Shaw, M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, and J.C. Chao. Anton, a special-purpose ma-

chine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News*, 35:1 – 12, 2007.

[51] D.E. Shaw, J. Grossman, J.A. Bank, B. Batson, J.A. Butts, J.C. Chao, M.M. Deneroff, R.O. Dror, A. Even, and C.H. Fenton. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, pages 41 – 53, 2014.

[52] K. Lindorff-Larsen, S. Piana, R.O. Dror, and D.E. Shaw. How fast-folding proteins fold. *Science*, 334:517 – 20, 2011.

[53] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341 – 6, 2010.

[54] H.S. Chung, S. Piana-Agostinetti, D.E. Shaw, and W.A. Eaton. Structural origin of slow diffusion in protein folding. *Science*, 349:1504 – 10, 2015.

[55] R.O. Dror, A.C. Pan, D.H. Arlow, D.W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D.E. Shaw. Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proc Natl Acad Sci USA*, 108:13118 – 23, 2011.

[56] Y. Shan, E.T. Kim, M.P. Eastwood, R.O. Dror, M.A. Seeliger, and D.E. Shaw. How does a drug molecular find its target binding site. *Journal of the American Chemical Society*, 133:9181 – 3, 2011.

[57] R.O. Dror, T.J. Mildorf, D. Hilger, A. Manglik, D.W. Borhani, D.H. Arlow, A. Philippsen, N. Villanueva, Z. Yang, M.T. Lerch, W.L. Hubbell, B.K. Kobilka, R.K. Sunahara, and D.E. Shaw. Signal transduction: Structural basis for nucleotide exchange in heterotrimeric g proteins. *Science*, 348:1361 – 5, 2015.

[58] V.S. Pande, K. Beauchamp, and G.R. Bowman. Everything you wanted to know about markov state models but were too afraid to ask. *Methods*, 52:99–105, 2010.

[59] J.H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J.D. Chodera, C. Schutte, and F. Noe. Markov models of molecular kinetics: Generation and validation. *J Chem Phys*, 134:174105, 2011.

[60] G.R. Bowman, D.L. Ensign, and V.S. Pande. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J Chem Theory Comput*, 6:787–94, 2010.

[61] J.K. Weber and V.S. Pande. Characterization and rapid sampling of protein folding markov state model topologies. *J Chem Theory Comput*, 7:3405–3411, 2011.

[62] S. Chowdhury, M.C. Lee, and Y. Duan. Characterizing the rate-limiting step of trp-cage folding by all-atom molecular dynamics simulations. *The Journal pf Physical Chemistry B*, 108:13855 – 65, 2004.

[63] K.A. Beauchamp, R. McGibbon, Y.S. Lin, and V.S. Pande. Simple few-state models reveal hidden complexity in protein folding. *Proc Natl Acad Sci USA*, 109:17807–13, 2012.

[64] Y. Okamoto. Generalized-ensemble algorithms: Enhanced sampling techniques for monte carlo and moleculary dynamics simulations. *J Mol Graph Model*, 22:425 – 39, 2004.

[65] R.C. Bernardi, M.C. Melo, and K. Schulten. Enhanced sampling techniques in molecular dynamicssimulations of biological systems. *Biochimica et Biophysica Acta*, 1850:872 – 7, 2015.

[66] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*, 314:141 – 151, 1999.

[67] U.H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett*, 281:140 – 150, 1997.

[68] T. Mori, N. Miyashita, W. Im, M. Feig, and Y. Sugita. Molecular dynamics simulations of biological membranes and membrane proteins using enhanced conformational sampling algorithms. *Bba-Biomembranes*, 1858:1635–1651, 2016.

[69] W. Zhang, J. Yang, B. He, S.E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H.B. Shen, and Y. Zhang. Integration of quark and i-tasser for ab initio protein structure prediction in casp11. *Proteins*, 84 Suppl 1:76 – 85, 2016.

[70] S. Piana and A. Laio. A bias-exchange approach to protein folding. *J Phys Chem*, 111:4553–9, 2007.

[71] M. Laio, A. ad Parrinello. Escapign free energy minima. *Proc Natl Acad Sci USA*, 99:12562–12566, 2002.

[72] O. Valsson, P. Tiwary, and M. Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annu Rev Phys Chem*, 67:159–84, 2016.

[73] A. Barducci, M. Bonomi, and M. Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1:826–843, 2011.

[74] Y.M. Rhee and V.S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophysical Journal*, 84:775–786, 2003.

[75] J.W. Pitera and W. Swope. Understanding folding and design: Replica exchange simulations of "trp-cage" fly miniproteins. *Proc Natl Acad Sci USA*, 100:7587–7592, 2003.

[76] F. Jiang and Y.D. Wu. Folding of fourteen small proteins with a residue-specific

force field and replica exchange molecular dynamics. *J Am Chem Soc*, 136:9536–9, 2014.

[77] H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *Journal of the American Chemical Society*, 136:13959–13962, 2014.

[78] K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268:209 – 25, 1997.

[79] P. Bradley, L. Malmstrom, B. Qian, J. Schonbrun, D. Chivian, D.E. Kim, J. Meiler, K.M. Misura, and D. Baker. Free modeling with rosetta in casp6. *Proteins*, 61 Suppl 7:128–34, 2005.

[80] D. Xu and Y. Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowedge-based force field. *Proteins*, 80:1715 – 35, 2012.

[81] M. Karakas, N. Woetzel, R. Staritzbichler, N. Alexander, B.E. Weiner, and J. Meiler. Bcl::fold - de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One*, 7, 2012.

[82] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087 – 1092, 1953.

[83] A. Vitalis and R.V. Pappu. Methods for monte carlo simulations of biomacromolecules. *Annu Rep Comput Chem*, 5:49–76, 2009.

[84] H.H. Xiangian Hu, David N. Beratan, and Weitao Yang. A gradient-directed monte carlo approach for protein design. *J Comp Chem*, 31, 2010.

[85] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein foldin - a perspective from simple exact models. *Protein Sci*, 4:561 – 602, 1995.

[86] X. Hu, D.N. Beratan, and W. Yang. A gradient-directed monte carlo method for global optimization in a discrete space: Application to protein sequence design and folding. *The Journal of Chemmical Physics*, 131:154117, 2009.

[87] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671 – 80, 1983.

[88] D.A.S. Constantino Tsallis. Generalized simulated annealing. *Physica A*, 233, 1996.

[89] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[90] S. Schulze-Kremer. Genetic algorithms and protein folding. *Protein Structure Prediction: Methods and Protocols*, pages 175 – 222, 2000.

[91] Y. Cui, R.S. Chen, and W.H. Wong. Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins: Structure, Function, and Bioinformatics*, 31:247 – 257, 1998.

[92] J.T. Pedersen and J. Moult. Genetic algorithms for protein structure prediction. *Curr Opin Struct Biol*, 6:227 – 31, 1996.

[93] R. Unger. The genetic algorithm approach to protein structure prediction. *Struct Bond*, 110:153 – 175, 2004.

[94] F.L. Custodio, H.J.C. Barbosa, and L.E. Dardenne. Investigation of the three-dimensional lattice hp protein folding model using a genetic algorithm. *Genet Mol Biol*, 27:611 – 615, 2004.

[95] F.L. Custodio, H.J.C. Barbosa, and L.E. Dardenne. A multiple minima genetic algorithm for protein structure prediction. *Appl Soft Comput*, 15:88 – 99, 2014.

[96] M.T. Hoque, M. Chetty, and A. Sattar. Genetic algorithm in ab initio protein structure prediction using low resolution model: A review. *Biomedical Data and Applications*, pages 317 – 342, 2009.

[97] X.L. Zhang, T. Wang, H.P. Luo, J.Y. Yang, Y.P. Deng, J.S. Tang, and M.Q. Yang. 3d protein structure prediction with genetic tabu search algorithm. *Bmc Syst Biol*, 4, 2010.

[98] B. Boskovic and J. Brest. Genetic algorithm with advanced mechanisms applied to the protein structure prediction in a hydrophobic-polar model and cubic lattice. *Appl Soft Comput*, 45:61 – 70, 2016.

[99] M.A. Rashid, S. Iqbal, F. Khatib, M.T. Hogue, and A. Sattar. Guided macro-mutation in a graded energy-based genetic algorithm for protein structure prediction. *Computational Biology and Chemistry*, 61:162 – 177, 2016.

[100] T. Dandekar and P. Argos. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Engineering*, 5:637 – 645, 1992.

[101] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231:75 – 81, 1993.

[102] K.F. Lau and K.A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22:3986 – 3997, 1989.

[103] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, pages 139–45, 2000.

[104] W. Wang, O. Donini, C.M. Reyes, and P.A. Kollman. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand,

protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct*, 30:211–43, 2001.

[105] J.W. Ponder and D.A Case. Force fields for protein simulations. *Adv Protein Chem*, 66:27–85, 2003.

[106] P.E. Lopes, O. Guvench, and A.D. MacKerell. Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol*, 1215:47–71, 2015.

[107] A.D. MacKerell. Empirical force fields for biological macromolecules: Overview and issues. *J Comput Chem*, 25:1584 – 604, 2004.

[108] A. Godzik. Knowledge-based potentials for protein folding: What can we learn from known protein structures? *Structure*, 4:363–6, 1996.

[109] M.J. Sippl. Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, 5:229 – 35, 1995.

[110] J. Skolnick. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol*, 16:166–71, 2006.

[111] S. Lifson and A. Warshel. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *The Journal of Chemical Physics*, 49:5116 – 5129, 1968.

[112] J. Behler. Perspective: Machine learning potentials for atomistic simularions. *J Chem Phys*, 145:170901, 2016.

[113] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187 – 217, 1983.

[114] B.R. Gelin and M. Karplus. Side-chain torsional potentials: Effect of dipeptide, protein, and solvent environment. *Biochemistry*, 18:1256 – 68, 1979.

[115] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, N. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102:3586 – 616, 1998.

[116] A.D. MacKerrell, M. Feig, and C.L. Brooks. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformation distributions in molecular dynamics simulations. *J Comput Chem*, 25:1400 – 15, 2004.

[117] R.B. Best, X. Zhu, J. Shim, P.E. Lopes, J. Mittal, and A.D. Feig, M. Mackerell. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi, and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput*, 8:3257 – 3273, 2012.

[118] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, D.M. Yang, W. ad York, and M. Karplus. Charmm: The biomolecular simulation program. *J Comput Chem*, 30:1545 – 614, 2009.

[119] P.K. Weiner and P.A. Kollman. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2:287 – 303, 1981.

[120] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106:765 – 784, 1984.

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

[122] D.W. Li and R. Bruschweiler. Nmr-based protein potentials. *Angew Chem Int Ed Engl*, 49:6778 – 80, 2010.

[123] W.L. Jorgensen and J. Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc*, 110:1657 – 66, 1988.

[124] M.J. Robertson, J. Tirado-Rives, and W.L. Jorgensen. Improved peptide and protein torsional energetics with the oplsaa force field. *J Chem Theory Comput*, 11:3499 – 509, 2015.

[125] W. Van Gunsteren and H. Berendsen. Groningen molecular simulation (gromos) library manual. *Biomos, Groningen*, 24:13, 1987.

[126] W.F. Van Gunsteren, X. Daura, and A.E. Mark. Gromos force field. *Encyclopedia of Computational Chemistry*, 1998.

[127] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, and A.H. de Vries. The martini force field: Coarse grained model for biomolecular simulations. *J Phys Chem B*, 111:7812 – 24, 2007.

[128] L. Monticelli, S.K. Kandasamy, X. Periole, R.G. Larson, D.P. Tieleman, and S.J. Marrink. The martini coarse-grained force field: Extension to proteins. *J Chem Theory Comput*, 4:819 – 34, 2008.

[129] C.A. Lopez, Z. Sovova, F.J. van Eerden, A.H. de Vries, and S.J. Marrink. Martini force field parameters for glycolipids. *J Chem Theory Comput*, 9:1694 – 708, 2013.

[130] J.A. McCammon, B.R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585 – 90, 1977.

[131] V.S. Pande, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, and E.J. Sorin. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68:91 – 109, 2003.

[132] C.M. Summa and M. Levitt. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci USA*, 104:3177 – 82, 2007.

[133] S. Piana, K. Lindorff-Larsen, and D.E. Shaw. Atomic-level description of ubiquitin folding. *Proc Natl Acad Sci USA*, 110:5915 – 20, 2013.

[134] J. Huang and A.D. MacKerell. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *J Comput Chem*, 34:2135 – 45, 2013.

[135] K. Lindorff-Larsen, P. Maragakis, S. Piana, M.P. Eastwood, R.O. Dror, and D.E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS One*, 7, 2012.

[136] K.K. Patapati and N.M. Glykos. Three force field' views of the 3(10) helix. *Biophys J*, 101:1766 – 71, 2011.

[137] J. Lee, S. Wu, and Y. Zhang. Ab initio protein structure prediction. *From Protein Structure to Function with Bioinformatics*, pages 3 – 25, 2009.

[138] S. Piana, J.L. Klepeis, and D.E. Shaw. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*, 24:98 – 105, 2014.

[139] S. Piana, K. Lindorff-Larsen, and D.E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc Natl Acad Sci USA*, 109:17845 – 50, 2012.

[140] M.J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding: An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7:473 – 501, 1993.

[141] S.J. Wodak. Protein structure and stability: Database-derived potentials and prediction. *Encyclopedia of Computational Chemistry*.

[142] P.D. Thomas and K.A. Dill. Statistical potentials extracted from protein structures: How accurate are they. *J Mol Biol*, 257:457 – 69, 1996.

[143] A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics*, 107:3698 – 3706, 1997.

[144] D. Shortle. Propensities, probabilities, and the boltzmann hypothesis. *Protein Sci*, 12:1298 – 302, 2003.

[145] A.V. Finkelstein, A.Y. Badretdinov, and A.M. Gutin. Why do protein architectures have boltzmann-like statistics. *Proteins-Structure Function and Genetics*, 23:142 – 150, 1995.

[146] J. Moult. Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol*, 7:194 – 9, 1997.

[147] T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, and J. Ferkinghoff-Borg. Potentials of mean force for protein structure prediction vindicated, formalized, and generalized. *PLoS One*, 5, 2010.

[148] M.J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and H. Flockner. Helmholtz free energies of atom pair interactions in proteins. *Fold Des*, 1:289 – 98, 1996.

[149] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44:223 – 32, 2011.

[150] M.Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15:2507 – 24, 2006.

[151] J. Kuszewski, A.M. Gronenborn, and G.M. Clore. Improving the quality of nmr and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci*, 5:1067–80, 1996.

[152] J.S. Yang, J.H. Kim, S. Oh, G. Han, S. Lee, and J. Lee. Stap refinement of the nmr database: a database of 2405 refined solution nmr structures. *Nucleic Acids Res*, 40:D525–30, 2012.

[153] P. Majek and R. Elber. A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins. *Proteins-Structure Function and Bioinformatics*, 76:822 – 836, 2009.

[154] J.P. Kocher, M.J. Rooman, and S.J. Wodak. Factors influencing the ability of knowledge-based potentials to idenitfy native sequence-structure matches. *J Mol Biol*, 235:1598 – 613, 1994.

[155] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, 295:337 – 56, 2000.

[156] S.Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. derivation of interaction potentials. *J Comput Chem*, 27:1866 – 75, 2006.

[157] S.Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: Ii. validation of the scoring function. *J Comput Chem*, 27:1876 – 82, 2006.

[158] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy fuction for protein-ligand, protein-protein, and protein-dna complexes. *J Med Chem*, 48:2325 – 35, 2005.

[159] A.M. Poole and R. Ranganathan. Knowledge-based potentials in protein design. *Curr Opin Struct Biol*, 16:508 – 13, 2006.

[160] N. Woetzel, M. Karakas, R. Staritzbichler, R. Muller, B.E. Weiner, and J. Meiler. Bcl::score - knowledge-based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS One*, 7, 2012.

[161] B.E. Weiner, N. Woetzel, M. Karakas, N. Alexander, and J. Meiler. Bcl::mp-fold: Folding membrane proteins through assembly of transmembrane helices. *Structure*, 21:1107 – 17, 2013.

[162] A.W. Fischer, S. Heinze, D.K. Putnam, B. Li, J.C. Pino, Y. Xia, C.F. Lopez, and J. Meiler. Casp11 - an evaluation of a modular bcl::fold-based protein structure prediction pipeline. *PLoS One*, 11, 2016.

[163] M.J. Sippl. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213:859 – 83, 1990.

[164] S. Tanaka and H.A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9:945 – 950, 1976.

[165] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18:534 – 552, 1985.

[166] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256:623 – 44, 1996.

[167] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediciton. *Protein Sci*, 11:2714 – 26, 2002.

[168] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M.J. Sippl. Identification of native protein folds amongst a large number of incorrect models: The calculation of low energy conformations from potentials of mean force. *J Mol Biol*, 216:167 – 80, 1990.

[169] I. Bahar and R.L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol*, 266:195 – 214, 1997.

[170] B. Park and M. Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258:367 – 92, 1996.

[171] B.H. Park, E.S. Huang, and M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol*, 266:831 – 46, 1997.

[172] B.A. Reva, A.V. Finkelstein, M.F. Sanner, and A.J. Olson. Residue-resdiue mean-force potentials for protein structure recognition. *Protein Eng*, 10:865 – 76, 1997.

[173] F. Melo and E. Feytmanns. Novel knowledge-based mean force potential at atomic level. *J Mol Biol*, 267:207 – 22, 1997.

[174]

[175] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275:895 – 916, 1998.

[176] M.R. Betancourt and D. Thirumalai. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*, 8:361 – 9, 1999.

[177] C.M. Summa, M. Levitt, and W.F. Degrado. An atomic environment potential for use in protein structure prediction. *J Mol Biol*, 352:986 – 1001, 2005.

[178] Y. Dehouck, D. Gilis, and M. Rooman. A new generation of statistical potentials for proteins. *Biophys J*, 90:4010 – 7, 2006.

[179] Q. Fang and D. Shortle. A consistent set of statistical potentials for quantifyinh local side-chain and backbone interactions. *Proteins*, 60:90 – 6, 2005.

[180] J. Qiu and R. Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, 61:44 – 55, 2005.

[181] M.J. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *J Mol Biol*, 260:644 – 8, 1996.

[182] J.U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science*, 253:164 – 70, 1991.

[183] S. DeLuca, B. Dorr, and J. Meiler. Design of native-like proteins through an exposure-dependent environment potential. *Biochemistry*, 50:8521 – 8, 2011.

[184] E. Durham, B. Dorr, N. Woetzel, R. Staritzbichler, and J. Meiler. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model*, 15:1093 – 108, 2009.

[185] B. Li, J. Mendenhall, E.D. Nguyen, B.E. Weiner, A.W. Fischer, and J. Meiler. Accurate prediction of contact numbers for multi-spanning helical membrane proteins. *J Chem Inf Model*, 56:423 – 34, 2016.

[186] A.W. Fischer, N.S. Alexander, N. Woetzel, M. Karakas, B.E. Weiner, and J. Meiler. Bcl::mp-fold: Membrane protein structure prediction guided by epr restraints. *Proteins*, 83:1947 – 62, 2015.

[187] T.R. Kim, J.S. Yang, S. Shin, and J. Lee. Statistical torsion angle potential energy functions for protein structure modeling: A bicubis interpolation approach. *Proteins*, 81:1156 – 65, 2013.

[188] E.A.D. Amir, N. Kalisman, and C. Keasar. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins-Structure Function and Bioinformatics*, 72:62 – 73, 2008.

[189] C. Ramakrishnan and G.N. Ramachandran. Stereochemical criteria for polypeptide and protein chain conformations. ii¿ allowed conformations for a pair of peptide units. *Biophys J*, 5:909–33, 1965.

[190] R.L. Dunbrack and M. KArplus. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol*, 230:543–74, 1993.

[191] R.A. Venters, C.-C. Huang, B.T. Farmer, R. Trolard, L.D. Spicer, and C.A. Fierke. High-level 2h/13c/15n labeling of proteins for nmr studies. *Journal of Biomolecular NMR*, 5:339 – 344, 1995.

[192] J.L. Battiste and G. Wagner. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry*, 39:5355 – 5365, 2000.

[193] D. Latek, D. Ekonomiuk, and A. Kolinski. Protein structure prediction: Combining de novo modeling with sparse experimental data. *Journal of Computational Chemistry*, 28:1668 – 1676, 2007.

[194] P.M. Bowers, C.E.M. Strauss, and D. Baker. De novo protein structure determination using sparse nmr data. *Journal of Biomolecular NMR*, 18:311 – 318, 2000.

[195] B.E. Weiner, N. Alexander, L.R. Akin, N. Woetzel, M. Karakas, and J. Meiler. Bcl::fold - protein topology determination from limited nmr restraints. *Proteins*, 82:587 – 595, 2014.

[196] Z.T. Farahbakhsh, C. Altenbach, and W.L. Hubbell. Spin labeled cysteines as sensors for protein-lipid interaction and conformation in rhodopsin. *Photochemistry and Photobiology*, 56:1019 – 1033, 1992.

[197] C. Altenbach, W. Froncisz, R. Hemker, H. McHaourab, and W.L. Hubbell. Accessibility of nitroxide side chains: Absolute heisenberg exchange rates from power saturation epr. *Biophysical Journal*, 89:2103 – 2112, 2005.

[198] M.D. Rabenstein and Y.K. Shin. Determination of the distance between two spin labels attached to a macromolecule. *Proceedings of the National Academy of Sciences*, 92:8239 – 8243, 1995.

[199] P.P. Borbat, H.S. McHaourab, and J.H. Freed. Protein structure determination using long-distance constraints from double quantum coherence esr: Study of t4 lysozyme. *Journal of the American Chemical Society*, 124:5304 – 5314, 2002.

[200] N. Alexander, M. Bortolus, A. Al-Mestarihi, H. McHaourab, and J. Meiler. De novo high-resolutioin protein structure determination from sparse spin labeling epr data. *Structure*, 16:181 – 195, 2008.

[201] S.J. Hirst, N. Alexander, H.S. McHaourab, and J. Meiler. Rosettaepr: An integrated tool for protein structure determination from sparse epr data. *Journal of Structural Biology*, 173:506 – 514, 2011.

[202] A.W. Fischer, N.S. Alexander, N. Woetzel, M. Karakas, B.E. Weiner, and J. Meiler. Bcl::mp-fold: Membrane protein structure prediction guided by epr restraints. *Proteins: Structure, Function, and Bioinformatics*, 83:1947 – 1962, 2015.

[203] K. Lasker, F. Forster, S. Bohn, T. Walztheoni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, and W. Baumeister. Molecular architecture of the 26s proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences of the United States of America*, 109:1380 – 1387, 2012.

[204] R.B. Jacobsen, K.L. Sale, M.J. Ayson, P. Novak, J. Hong, P. Lane, N.L. Wood, G.H. Kruppa, M.M. Young, and J.S. Schoeniger. Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linkng and high-resolution mass spectrometry. *Protein Science: A Publication of the Protein Society*, 15:1303 – 1317, 2006.

[205] S. Kalkhof, C. Ihling, K. Mechtler, and A. Sinz. Chemical cross-linking and high-performance fourier transform ion cyclotron resonance mass spectrometry for protein interaction analysis: Application to a calmodulin/target peptide complex. *Analytical Chemistry*, 77:495 – 503, 2005.

[206] J.W. Back, L. de Jong, A.O. Muijsers, and C.G. De Koster. Chemical cross-linking and mass spectrometry for protein structural modeling. *Journal of Molecular Biology*, 331:303 – 313, 2003.

[207] M.M. Young, N. Tang, J.C. Hempel, C.M. Oshiro, E.W. Taylor, I.D. Kuntz, B.W. Gibson, and G. Dollinger. High throughput protein fold identification by using ex-

perimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 97:5802 – 5806, 2000.

[208] A. Sinz. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *Journal of Mass Spectrometry*, 38:1225 – 1237, 2003.

[209] E. Bihnstein and M. Ohi. Cryo-electron microscopy and the amazing race to atomic resolution. *Biochemistry*, 54:3133 – 3141, 2015.

[210] W. Jiang, M.L. Baker, S.J. Ludtke, and W. Chiu. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *Journal of Molecular Biology*, 308:1033 – 1044, 2001.

[211] S. Abeysinghe, T. Ju, M.L. Baker, and W. Chiu. Shape modeling and matching in identifying 3d protein structures. *Computer-Aided Design*, 40:708 – 720, 2008.

[212] M.L. Baker, Z. Yu, W. Chiu, and C. Bajaj. Automated segmentation of molecular subunits in electron cryomicroscopy density maps. *Journal of Structural Biology*, 156:432 – 441, 2006.

[213] V.B.I. Burger and C. Chennubhotla. A hierarchical elastic network model for unsupervised em density map segmentation. 2, 2011.

[214] G.D. Pintilie, J. Zhang, T.D. Goddard, W. Chiu, and D.C. Gossard. Quantitative analysis of cryo-em density map segmentation by watershed and space-scale filtering and fitting of structures by alignment to regions. *Journal of Structural Biology*, 170:427 – 438, 2010.

[215] M.L. Baker, T. Ju, and W. Chiu. Identification of secondary structure elements in intermediate resolution density maps. *Structure*, 15:7 – 19, 2007.

[216] Y. Kong and J. Ma. A structural-informatics approach for mining beta-sheets: Locating sheets in intermediate-resolution density maps. *Journal of Molecular Biology*, 332:399 – 413, 2003.

[217] Y. Kong, X. Zhang, T.S. Baker, and J. Ma. A structural-informatics approach for tracing beta-sheets: Building psuedo-calpha traces for beta-strands in intermediate-resolution density maps. *Journal of Molecular Biology*, 339:117 – 130, 2004.

[218] M. Chen, P.R. Baldwin, S.J. Ludtke, and M.L. Baker. De novo modeling in cryo-em density maps with pathwalking. *Journal of Structural Biology*, 196:289 – 298, 2016.

[219] S. Lindert, R. Staritzbichler, N. Wotzel, M. Karakas, P.L. Stewart, and J. Meiler. Emfold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron density maps. *Structure*, 17:990 – 1003, 2009.

[220] S. Lindert, N. Alexander, N. Wotzel, M. Karakas, P.L. Stewart, and J. Meiler. Emfold: De novo atomic-detail protein structure determination from medium resolution density maps. *Structure*, 20:464 – 478, 2012.

[221] R.Y.R. Wang, M. Kudryashev, X. Li, E.H. Egelman, M. Basler, Y. Cheng, D. Baker, and F. DiMaio. De novo protein structure determination from near-atomic-resolution cryo-em maps. *Nat Meth*, 12:335 – 338, 2015.

[222] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292:195 – 202, 1999.

[223] R. Yan, D. Xu, J. Yang, S. Walker, and Y. Zhang. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports*, 3:2619, 2013.

[224] J.K. Leman, R. Mueller, N. Karakas, M. Woetzel, and J. Meiler. Simultaneous pre-

diction of protein secondary structure and transmembrane spans. *Proteins: Structure, Function, and Bioinformatics*, 81:1127 – 1140, 2013.

[225] B. Rost and C.J. Sander. *J Mol Biol*, 232:584 – 599, 1993.

[226] B. Rost, R. Schneider, and C. Sander. *Trends Biochem Sci*, 18:120 – 123, 1993.

[227] H. Viklund, A. Bernsel, M. Skwark, and A. Elofsson. Spoctopus: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, 24:2928 – 2929, 2008.

[228] H. Viklund and A. Elofsson. Octopus: Improving topology prediction by two-track ann-based preference scores and an extended topological gramar. *Bioinformatics*, 24:1662 – 1668, 2008.

[229] T. Nugent and D.t¿ Jones. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10:1 – 11, 2009.

[230] A. Krogh, B. Larsson, G. von Heijne, and E.L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *Journal of Molecular Biology*, 305:567 – 580, 2001.

[231] R.Y. Kahsay, G. Gao, and L. Liao. An improved hidden markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics*, 21:1853 – 1858, 2005.

[232] U. Gobel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18:309 – 317, 1994.

[233] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding and Design*, 2:S25 – S32, 1997.

[234] D.E. Kim, F. DiMaio, R. Yu-Ruei Wang, Y. Song, and D. Baker. One contact for every twelve residues allow robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82:208 – 218, 2014.

[235] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano. Critical assesment of methods of protein structure prediction (casp) - progress and new directions in round xi. *Proteins: Structure, Function, and Bioinformatics*, 2016.

[236] B. Monastyrskyy, D. D'Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych. New encouraging developments in contact prediction: Assessment of the casp11 results. *Proteins: Structure, Function, and Bioinformatics*, 2015.

[237] A.D. McLachlan. Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c551. *Journal of Molecular Biology*, 61:409 – 424, 1971.

[238] H. Ashkenazy and Y. Kliger. Reducing phylogenetic bias in correlated mutation analysis. *Protein Engineering Design and Selection*, 23:321 – 326, 2010.

[239] E. Neher. How frequent are correlated changes in families of protein sequences. *Proceedings of the National Academy of Sciences*, 91:98 – 102, 1994.

[240] D.D. Pollock and W.R. Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering*, 10:647 – 657, 1997.

[241] T. Hopf, C. Scharfe, J. Rodrigues, A. Green, O. Kohlbacher, C. Sander, A. Bonvin, and D. Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *eLife*, 3, 2014.

[242] L. Burger and E. van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol*, 6, 2010.

[243] D.T. Jones, D.W.A. Buchan, D. Cozzetto, and M. Pontil. Psicov: Precise strctural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28:184 – 190, 2012.

[244] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. HWa, and M. Weigt. Direcct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108:E1293 – E1301, 2011.

[245] H. Kamisetty, S. Ovchinnikov, and D. Baker. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110:15674 – 15679, 2013.

[246] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, 6:e28766, 2011.

[247] D.S. Marks, T.A. Hopf, and C. Sander. Protein structure prediction from sequence variation. *Nat Biotech*, 30:1072 – 1080, 2012.

[248] D.T. Jones, T. Singh, T. Kosciolek, and S. Tetchner. Metapsicov: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31:999 – 1006, 2015.

[249] S. Ovchinnikov, H. Kamisetty, and D. Baker. Robust and accurate prediction of residue-residue interactions across protein interfacecs using evolutionary information. *eLife*, 3:e02030, 2014.

[250] M. Ekeberg, T. Hartonen, and E. Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341 – 356, 2014.

[251] M. Michel, S. Hayat, M.J. Skwark, C. Sander, D.S. Marks, and A. Elofsson. Pconsfold: Improved contact predictions improve protein models. *Bioinformatics*, 30:i482 – i488, 2014.

[252] M.J. Skwark, A. Abdel-Rehim, and A. Elofsson. Pconsc: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, 29:1815 – 1816, 2013.

[253] M.J. Skawrk, D. Raimondi, M. Michel, and A. Elofsson. Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol*, 10:e1003889, 2014.

[254] L. Kajan, T.A. Hopf, M. Kalas, D.S. Marks, and B. Rost. Freecontact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, 15:1 – 6, 2014.

[255] A.N. Tegge, Z. Wang, J. Eickholt, and J. Cheng. Nncon: Improved protein contact map prediction using 2d-recursive neural network. *Nucleic Acids Research*, 37, 2009.

[256] B. Xue, E. Faraggi, and Y. Zhou. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins*, 76:176 – 183, 2009.

[257] G. Shackelford and K. Karplus. Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics*, 69:159 – 164, 2007.

[258] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12:15 – 21, 1999.

[259] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, 45:157 – 162, 2001.

[260] R.M. MacCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20:i224 – i231, 2004.

[261] P. Chen and J. Li. Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC Structural Biology*, 10, 2010.

[262] Y. Li, Y. Fang, and J. Fang. Predicting residue-residue contacts using random forest models. *Bioinformatics*, 27:3379 – 3384, 2011.

[263] P. Bjorkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson, and T.R. Hvidsten. Using multi-data hidden markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. *Bioinformatics*, 25:1264 – 1270, 2009.

[264] M. Lippi and P. Frasconi. Prediction of protein beta-residue contacts by markov logic networks with grounding-specific weights. *Bioinformatics*, 25:1264 – 1270, 2009.

[265] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113, 2007.

[266] S. Wu and Y. Zhang. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, 24:924 – 931, 2008.

[267] T. Kosciolek and D.T. Jones. Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics*, 2015.

[268] M. Stout, J. Bacardit, J.D. Hirst, R.E. Smith, and N. Krasnogor. Prediction of topological contacts in proteins using learning classifier systems. *Soft Computing*, 13:245 – 258, 2008.

[269] B. Wallner and A. Elofsson. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science*, 15:900 – 913, 2006.

[270] J. Ma, S. Wang, F. Zhao, and J. Xu. Protein threading using context-specific alignment potential. *Bioinformatics*, 29:i257 – i265, 2013.

[271] R. Bonneau and D. Baker. Ab initio protein structure prediction: Progress and prospects. *Annu Rev Biophys Biomol Struct*, 30:173 – 89, 2001.

[272] C. Hardin, T.V. Pogorelov, and Z. Luthey-Schulten. Ab initio protein structure prediction. *Curr Opin Struct Biol*, 12:176 – 81, 2002.

[273] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5:823 – 6, 1986.

[274] K. Illergard, D.H. Ardell, and Elofsson A. Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. *Proteins*, 77:499 – 508, 2009.

[275] A. Fiser, M. Feig, C.L. Brooks, and A. Sali. Evolution and physics in comparative protein structure modeling. *Acc Chem Res*, 35:413 – 21, 2002.

[276] D.T. Jones. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, Suppl 1:185 – 91, 1997.

[277] D.T. Jones. Predicting novel protein folds by using fragfold. *Proteins*, Suppl 5:127 – 32, 2001.

[278] D.T. Jones and L.J. McGuffin. Assembling novel protein folds from super-secondary structural fragments. *Proteins*, 53 Suppl 6:480 – 5, 2003.

[279] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292:195 – 202, 1999.

[280] B. Li, J. MEndenhall, E.D. Nguyen, B.E. Weiner, A.W. Fischer, and J. Meiler. Accurate prediction of contact numbers for multi-spanning helical membrane proteins. *J Chem Inf Model*, 2016.

[281] T. Kosciolek and D.T. Jones. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE*, 9:e92197, 2014.

[282] D.T. Jones, D.W. Buchan, D. Cozzetto, and M. Pontil. Psicov: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28:184 – 90, 2012.

[283] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, and D. Baker. Protein structure prediction using rosetta. *Method Enzymol*, 383:66, 2004.

[284] P. Bradley, D. Chivian, J. Meiler, K.M. Misura, C.A. Rohl, W.R. Schief, W.J. Wedemeyer, O. Schueer-Furman, P. Murphy, J. Schonbrun, C.E. Strauss, and D Baker. Rosetta predictions in casp5: Successes, failures, and prospects for complete automation. *Proteins*, 53 Suppl 6:457–68, 2003.

[285] R. Jauch, H.C. Yeo, P.R. Kolatkar, and N.D. Clarke. Assessment of casp7 structure predictions for template free targets. *Proteins-Structure Function and Bioinformatics*, 69:57–67, 2007.

[286] Ovchinnikov S., H. Park, N. Varghese, P.S. Huang, G.A. Pavlopoulos, D.E. Kim, H. Kamisetty, N.C. Kyrpides, and D. Baker. Protein structure determination using metagenome sequence data. *Science*, 355:294–297, 2017.

[287] T. Kortemme, A.V. Morozov, and D. Baker. An orientation-dependent hydrogen

bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326:1239 – 1259, 2003.

[288] J.N.D. Battey, J. Kopp, R.J. Bordoli, L. Read, N.D. Clarke, and T. Schwede. Automated server predictions in casp7. *Proteins-Structure Function and Bioinformatics*, 69:68–82, 2007.

[289] L.N. Kinch, W. Li, B. Monastryskyy, A. Kryshtafovych, and N.V. Grishin. Evaluation of free modeling targets in casp11 and roll. *Proteins*, 84 Suppl 1:51 – 66, 2016.

[290] L. Kinch, S.Y. Shi, Q. Cong, H. Cheng, Y.X. Liao, and N.V. Grishin. Casp9 assessment of free modeling target predictions. *Proteins-Structure Function and Bioinformatics*, 79:59–73, 2011.

[291] Y. Zhang. I-tasser: Fully automated protein structure prediction in casp8. *Proteins-Structure Function and Bioinformatics*, 77:100–113, 2009.

[292] J.Y. Yang, R.X. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The i-tasser suite: Protein structure and function prediction. *Nature MEthods*, 12:7–8, 2015.

[293] A. Roy, A. Kucukural, and Y. Zhang. I-tasser: A unified platform for automated protein structure and function prediction. *Nat Protoc*, 5:725–738, 2010.

[294] S.T. Wu, J. Skolnick, and Y. Zhang. Ab initio modeling of small proteins by iterative tasser simulations. *Bmc Biol*, 5, 2007.

[295] Y. Zhang and J. Skolnick. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal*, 87:2647–2655, 2004.

[296] Y. Zhang, A. Kolinski, and J. Skoklnick. Touchstone ii: A new approach to ab initio protein structure prediction. *Biophysical Journal*, 85:1145–1161, 2003.

[297] K.W. Plaxco, K.T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277:985 – 94, 1998.

[298] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker. Contact order and ab initio protein structure prediction. *Protein Sci*, 11:1937 – 44, 2002.

[299] D. Baker. A suprising simplicity to protein folding. *NAture*, 405:39–42, 2000.

[300] B. Li, J. Mendenhall, E.D. Nguyen, B.E. Weiner, A.W. Fischer, and J. Meiler. Improving prediction of helix-helix packing in membrane proteins using predicted contact numbers as restraints. *Proteins: Structure, Function, and Bioinformatics*, 2017.

[301] E. Paci, K. Lindorff-Larsen, C.M. Dobson, and M. Karplus, M. nd Vendruscolo. Transition state contact orders correlate with protein folding rates. *J Mol Biol*, 352:495 – 500, 2005.

[302] J. Moult. A decade of casp: Progress, bottlenecks, and prognosis in protein structure prediction. *Curr Opin Struc Biol*, 15:285 – 289, 2005.

[303] Woetzel N. Staritzbichler R. Alexander N. Weiner B. Karakas, M. and J. Meiler. Bcl::fold - de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One*, 7(11), 2012.

[304] He J. Pontelli E. Dal Dalu, A. and Y. Lu. Identification of alpha-helices from low resolution protein density maps. *Computation Systems Bioinformatics Conference*, pages 89–98, 2006.

[305] Baker M.L. Ludtke S.J. Jiang, W. and W. Chiu. Bridging the information gap: computational tools for intermediate resolution structure interpretation1. *Journal of Molecular Biology*, 308(5):1033 – 1044, 2001.

[306] Ju T. Baker, M.L. and W. Chiu. Identification of secondary structure elements in

intermediate resolution density maps. *Structure (London, England : 1993)*, 15(1):7–19, 2007.

[307] Zhang X. Baker T.S. Kong, Y. and J. Ma. A structural-informatics approach for tracing -sheets: Building pseudo-c traces for -strands in intermediate-resolution density maps. *Journal of Molecular Biology*, 339(1):117 – 130, 2004.

[308] Staritzbichler R. Wotzel N. Karakas M. Stewart P. L. Lindert, S. and J. Meiler. Em-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron density maps. *Structure*, 17:990–1003, 2009.

[309] Alexander N. Wtzel N. Karaka M. Stewart P.L. Lindert, S. and J. Meiler. Em-fold: De novo atomic-detail protein structure determination from medium resolution density maps. *Structure(London, England:1993)*, 20(3):464–478, 2012.

[310] M.S. Shapovalov and R.L. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19:844 – 858, 2011.

[311] Baker M.L. Best C. Bi C. Dougherty M. Feng P. van Ginkel G. Devkota B. Lagerstedt I. Ludtke S.J. Newman R.H. Oldfield T.J. Rees I. Sahni G. Sala R. Velankar S. Warren J. Westbrook J.D. Henrick K. Kleywegt G.J. Berman H.M. Lawson, C.L. and W. Chiu. Emdatabank.org: Unified data resource for cryoem. *Nucleic Acids Research*, 39:456 – 464, 2011.

[312] J.B. Macqueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281 – 297, 1967.

[313] Irina Kufareva and Ruben Abagyan. Methods of protein structure comparison. *Methods Mol Biol*, 857:231–257, 2015.