Using Protein-Likeness to Validate Conformational Alternatives

by

Daniel A. Keedy

Department of Biochemistry
Duke University

Date: _____

Approved:

_____
David C. Richardson, Co-Supervisor

_____
Jane S. Richardson, Co-Supervisor

_____
Bruce R. Donald

_____
Terrence G. Oas

_____
Pei Zhou

_____
Brian Kuhlman

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biochemistry
in the Graduate School of Duke University
2012

Abstract

Using Protein-Likeness to Validate Conformational Alternatives

by

Daniel A. Keedy

Department of Biochemistry
Duke University

Date: _____

Approved:

_____
David C. Richardson, Co-Supervisor

_____
Jane S. Richardson, Co-Supervisor

_____
Bruce R. Donald

_____
Terrence G. Oas

_____
Pei Zhou

_____
Brian Kuhlman

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry in the Graduate School of Duke University
2012

Abstract

Proteins are among the most complex entities known to science. Composed of just 20 fundamental building blocks arranged in simple linear strings, they nonetheless fold into a dizzying array of architectures that carry out the machinations of life at the molecular level.

Despite this central role in biology, we cannot reliably predict the structure of a protein from its sequence, and therefore rely on time-consuming and expensive experimental techniques to determine their structures. Although these methods can reveal equilibrium structures with great accuracy, they unfortunately mask much of the inherent molecular flexibility that enables proteins to dynamically perform biochemical tasks. As a result, much of the field of structural biology is mired in a static perspective; indeed, most attempts to naïvely model increased structural flexibility still end in failure.

This document details my work to validate alternative protein conformations beyond the primary or equilibrium conformation. The underlying hypothesis is that more realistic modeling of flexibility will enhance our understanding of how natural proteins function, and thereby improve our ability to design new proteins that perform desired novel functions.

During the course of my work, I used structure validation techniques to validate conformational alternatives in a variety of settings. First, I extended previous work introducing the backrub, a local, sidechain-coupled backbone motion, by demon-

strating that backrubs also accompany sequence changes and therefore are useful for modeling conformational changes associated with mutations in protein design. Second, I extensively studied a new local backbone motion, helix shear, by documenting its occurrence in both crystal and NMR structures and showing its suitability for expanding conformational search space in protein design. Third, I integrated many types of local alternate conformations in an ultra-high-resolution crystal structure and discovered the combinatorial complexity that arises when adjacent flexible segments combine into networks. Fourth, I used structural bioinformatics techniques to construct smoothed, multi-dimensional torsional distributions that can be used to validate trial conformations or to propose new ones. Fifth, I participated in judging a structure prediction competition by using validation of geometrical and all-atom contact criteria to help define correctness across thousands of submitted conformations. Sixth, using similar tools plus collation of multiple comparable structures from the public database, I determined that low-energy states identified by the popular structure modeling suite Rosetta sometimes are valid conformations likely to be populated in the cell, but more often are invalid conformations attributable to artifacts in the physical/statistical hybrid energy function.

Unified by the theme of validating conformational alternatives by reference to high-quality experimental structures, my cumulative work advances our fundamental understanding of protein structural variability, and will benefit future endeavors to design useful proteins for biomedicine or industrial chemistry.

I dedicate this thesis to Gramps, who inspired me to keep learning and laughing.

Contents

List of Tables

List of Figures

List of Abbreviations

abbrev	abbreviation (...)
BACPAC	Beyond Alpha Carbons Prediction Assessment for CASP
BRDEE	Backrub Dead-End Elimination
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CD	circular dichroism
CSD	Cambridge Structural Database
CS-Rosetta	chemical shift Rosetta
DEE	Dead-End Elimination
DEEPer	Dead-End Elimination with Perturbations
DER	dynamic ensemble refinement
DNA	deoxyribonucleic acid
EDS	Electron Density Server
EROS	ensemble refinement with orientational restraints
FM	free modeling
FPLC	fast protein liquid chromatography
GDC-sc	global distance calculation for sidechains
GDT-HA	global distance test, high accuracy
GDT-TS	global distance test, total score
GMEC	global minimum energy conformation
HBmc	mainchain hydrogen bond correctness

HBsc	sidechain hydrogen bond correctness
H-bond	hydrogen bond
LGA	local-global alignment
MCRS	mainchain reality score
MPscore	MolProbity score
MD	molecular dynamics
MUMO	minimal under-restraining minimal over-restraining
NESG	Northeast Structural Genomics center
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
OSPREY	Open-Source Protein Redesign for You
PDB	Protein Data Bank
PheA	phenylalanine adenylation domain of gramicidin S synthetase A
PHENIX	Python-based Hierarchical ENvironment for Integrated Xtallography
PSI	Protein Structure Initiative
QM	quantum mechanics
RDC	residual dipolar coupling
RMSD	root mean square deviation
RNA	ribonucleic acid
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SSE	secondary structural element
TBM	template-based modeling
VTF	wwPDB X-ray Validation Task Force
wwPDB	worldwide Protein Data Bank

Acknowledgements

When I stop to think about it, I remember how lucky I am. I've had the great fortune to grow up in an era enlightened by learning and technology, in the best country in recorded history. Aside from this long-term perspective, I'm *additionally* blessed to have been surrounded by encouraging influences essentially my entire life. Not everyone has these advantages. I'll now at least make an attempt to properly acknowledge the people and institutions that have helped put me where I am today.

First, I thank Rhodes College and the Ned McWherter Scholars Program for the financial aid they've given me. I also thank the National Institutes of Health for supporting the Structural Biology and Biophysics program and research in the Richardson laboratory.

I'm thankful to Chuck Stinemetz at Rhodes for convincing me to get involved in laboratory research in the first place (as opposed to medical school...), and also to Terry Hill, Brent Hoffmeister, and Larry Zwiebel: their encouragement and mentorship started me on the biophysics research career path I'm still on today. Among my friends from Rhodes I'd like to particularly acknowledge Jake Cremer for his much-needed generosity and support – including rides to and from the Memphis airport for interviews at Duke.

I've had great input (both in official annual meetings and informally) on many scientific and career issues from my thesis committee at Duke: Terry Oas, Bruce Donald, Pei Zhou, and Brian Kuhlman. I especially appreciate Terry for always

being excited to talk with me about proteins – particularly when conformational ensembles are involved. I'd like to thank Bruce even more vigorously: I very nearly joined his lab, yet despite my officially being a member of a different group, he's treated me as something more than just a collaborator. The associated privileges have been great: a stream of insights from Bruce's towering intellect, plus a seemingly endless supply of humorous anecdotes, robot-related or otherwise.

On a related note, I would be remiss not to thank John MacMaster and especially Cheng-Yu Chen for selflessly showing me the ropes in the Donald wet lab for several months.

Most importantly on the academic side, I've been very blessed to be an "apprentice" to Dave and Jane Richardson. They've certainly been the most influential figures in my life scientifically speaking, but I've also learned a lot from their philosophies on life in general: do a thorough job where it's important, but don't take things too seriously, and take time for enjoying the inconceivably beautiful natural world around us.

Another major perk of spending six years in the RLab has been the company. Bryan Arendall's perseverance and perspective on science have been inspiring, and Lizbeth Videau has always been there to collaborate/commiserate/congratulate. Much of my thesis builds pretty directly on science done previously by Ian Davis, whose balance of work vs. other aspects of life I respect wholeheartedly. Late-night shift changes with the contemplative Laura Murray were a memorable staple early on. Jeremy Block has been a near-constant fixture in the lab, and possesses impressive talent not only with protein structures but also with people (he had me pegged as a future Richardson acolyte during my rotation). Gary Kapral was always a great source for computer tips or random stories. Vincent Chen may have done the most to lure me to the lab – the one-two punch of his mellow good humor and mad coding skills made me feel at home from the start (luckily I never faced one of his *actual*

ninja punches). Jeff Headd was always in good spirits and rearing to collaborate with me and others. I'll never forget Chris Williams for his quirky jokes (and actions), steady thought process, and growth as a scientist. Swati Jain helped show me that computer scientists bring a lot to the table in structural biology, especially when they're super-friendly. Bradley Hintze was always good for a slightly inappropriate but hilarious online video, and is someone whose attitude toward life I genuinely respect. The latest addition to the lab, Lindsay Deis, has inspired us all by doing real wet lab work and, importantly, connecting it integrally to her computational research (also, her surprise greetings while I played racquetball or basketball will forever haunt me).

Penultimately, I'd like to thank "the gang" of good friends at Duke. Racquetball, Xbox, and camping trips have been welcome intermittent respites from the rigors of graduate level science. I particularly appreciate my long-time roommate Andreas "Dr. Dre" Pfenning for kindly alerting me when Paula Deen was on, providing a steady stream of bad puns, and keeping me on my toes with recurring practical jokes involving hidden junk mail (I almost went postal by the time he graduated). A special thank you goes to Ranjula Wijayatunge for lots of fun times and emotional support.

Last and farthest from least, I thank my family. I would never wish upon them having to read this tome... but Mom/Dad/Joel/Sarah, if you *are* actually reading this, know that I love you very much.

Thanks, y'all!

1

Introduction

The rise of high-speed, low-cost DNA sequencing technologies has allowed the elucidation of many organisms' genomes, including those of platypi (Warren et al., 2008) and, most notably, humans (Venter et al., 2001; Lander et al., 2001). These spectacular results mark the advent of an information age in biology. Many resources will continue to be invested in the analysis of biological function as determined by sequences on a genome-wide scale.

However, genes are merely the source code for the program of cellular life. In an information transfer so fundamental it is known as the "central dogma of molecular biology" (Crick et al., 1970), genes enact their biological effects by specifying RNA molecules that encode the amino acid sequences of proteins (though we now appreciate that noncoding RNA is also quite important). These molecular machines flit about the intracellular milieu, instantiating the genetic information in a physical-chemical sense. They are the workhorses of cellular metabolism, homeostasis, defense, and propagation. Indeed, "protein comes from the Greek word meaning 'of first importance' – and so it is, for without proteins, there would be no life" (I. Asimov, 1993).

Proteins perform these literally vital functions by virtue of their folded three-dimensional structures. Precise arrangement of a protein's constituent atoms is critical for function, and can only be achieved when the amino acid chain folds to the correct shape. Again from Asimov: "A protein's amino acid components have to be arranged correctly in order for the right doohickeys to be in the right place to do the right job. You can't have a nitrogen atom waggling off there when it should be here, up against something else" (I. Asimov, 1993). Thus, while invaluable information can be gleaned from genomic analysis, a detailed mechanistic understanding of biology – and, perhaps more importantly, an improved ability to rationally manipulate biological systems to our benefit – requires detailed knowledge and understanding of molecular structure.

Two approaches currently dominate the experimental elucidation of macromolecular structure: X-ray crystallography and NMR spectroscopy. These methods are quite mature and widely used today, especially X-ray crystallography. Indeed, thousands of structures are determined each year, leading to massive growth of the internationally recognized macromolecular structure repository, the Protein Data Bank (Berman et al., 2000). However, impressive progress in subfields such as ribosome crystallography (Dunkle and Cate, 2010) notwithstanding, it remains difficult to crystallize or otherwise determine the structures of many biologically important macromolecules and complexes, such as membrane proteins, which are estimated to comprise over a quarter of human proteins.

In light of these struggles, it is appropriate to recall observations by Christian Anfinsen 50 years ago, which demonstrated that a protein's native conformation corresponds to its "most stable conformation, thermodynamically speaking" (Haber et al., 1962). This so-called "thermodynamic hypothesis" earned Anfinsen a Nobel Prize for relating sequence and structure in such a fundamental way, and remains a central tenet of structural biology. In principle, then, it should be possible to

bypass experimental structure determination and instead predict the native state by identifying the lowest-energy conformation. Efforts in this area have recently seen some successes (Chapters 6 and 7), but consistently accurate prediction remains an unattained goal.

A complementary scientific thrust is protein design (Chapters 2, 3, and 7), which can be considered the inverse folding problem: the goal is to identify one or more sequences that adopt the desired conformation (and perhaps therefore the desired function), rather than to identify the conformation that the given sequence adopts.

The most successful approaches in both fields (protein structure prediction and design) rely on computation rooted in classical physics. Underlying this methodology lies a deterministic view of reality, in which a "perfect" energy function can optimally predict future states given initial conditions. Apart from quantum mechanics (QM), in which stochasticity plays a fundamental role, there is no substantive reason to contradict this philosophy.

Unfortunately, however, classical physics in the form of commonly used molecular mechanics force fields has deficiencies. One central failure is the simplified supposition that atoms interact pairwise from their centers. In actuality, atoms experience steric interactions when the electron clouds surrounding them contact each other. Furthermore, the distributions of electrons in these clouds for two interacting atoms can be shifted by the presence of a third nearby atom; this polarization effect gives rise to higher-order interactions, which are especially critical to model for long-range electrostatics. Also, most proteins are shrouded in water, which is perhaps the most fundamental building block for life, but is unfortunately difficult to model for chemically subtle reasons. Explicit solvation models still suffer from the pairwise-from-centers paradigm, and nevertheless are often too computationally expensive. Implicit solvation models, on the other hand, fail to capture the complex transition from continuous bulk solvent to discrete ordered water molecules playing

integral structural roles at the protein surface (essentially acting as chemical extensions of the protein). These fundamental limitations are likely a primary explanation for continuing failures in even simple macromolecular modeling tasks (Das, 2011).

Computation rooted in higher-level theory such as QM could address many of these deficiencies. However, at least MP2 level theory is necessary to reproduce what is seen empirically, such as van der Waals interactions, so computational expense precludes the consistent use of a more nearly perfect energy function (with modern resources, at least).

In the context of theoretical soundness but practical limitations, it is important to remember what we have at our fingertips: thousands of experimental structures that provide windows into structural "ground truth". These "gold standards" provide a rich source of information on what it means to be "protein-like". An important corollary is that filtering based on quality criteria is critical to ensure reliability, not just total numbers – "quality over quantity" is the operative phrase. Fortunately, with modern data availability, one can make stringent demands on data quality and still obtain more than sufficient quantity.

The primary thesis of this work is that macromolecular modeling can be aided by empirically motivated techniques. For example, local mutation-coupled backbone moves observed in real structures can be used for protein design (Chapters 2 and 3), and carefully curated distributions of protein geometry (Chapter 5) can improve both interpretation of experimental data during crystallographic refinement (Chapter 4) and discrimination of native-like vs. unrealistic predicted structures (Chapters 6 and 7).

Empirical observations can also augment molecular mechanics energy functions and hopefully improve performance in terms of precise energy estimation. Here, by contrast, the goal is to develop backbone moves, protein-like heuristics, etc. to better differentiate realistic from unrealistic conformations, without regard for precise

energetics. Future orthogonal work is necessary to develop downstream procedures for comparing realistic alternatives based on more precise energetic criteria.

An immediate goal of this research is to use the explosion of data in the current biological information age to improve and better understand macromolecular structures obtained by traditional means. A longer-term goal is to learn from Mother Nature the fundamental determinants of protein structure; we hope to eventually gain sufficient understanding to rationally manipulate proteins for our own benefit, effectively mimicking natural evolution to more directly benefit our own species. Overall, this work aims to advance our ability to productively engineer the complex and dynamic biological world in which we find ourselves.

2

Backrubs and Mutations

2.1 The biological importance of macromolecular flexibility

Anfinsen's dogma (Sela et al., 1957; Haber et al., 1962; Anfinsen, 1973) tells us that a protein's amino acid sequence alone is sufficient to determine its native structure. This theory revolutionized our understanding of intracellular information flow, from gene to RNA transcript to protein structure with associated function. However, its common formulation – "sequence determines structure" (singular) – is dangerous, because it can lead one to mistakenly infer that a given protein exists in a single, well-defined, unique conformation.

Counter-examples to this strict single-native-structure paradigm soon arose. For instance, an early crystal structure of myoglobin (which was previously the subject of the very first protein crystal structure, albeit at lower resolution (Kendrew et al., 1958)) showed that the oxygen ligand had no viable entry or exit pathway in the static structure (Takano, 1977). For its crystal structure to be reconciled with its known biological function, myoglobin must "breathe" (Branden et al., 1991).

Many more cases of protein dynamism have since been demonstrated. For exam-

ple, a state-of-the-art hybrid NMR/MD approach accounting for up to μs-timescale motions demonstrated that ubiquitin conformationally varies primarily along one dominant mode of pincer-like global motion, and that its numerous and diverse binding partners select out ubiquitin conformations that are sampled along this mode (Lange et al., 2008). This conformational selection strategy may have been favored by evolution because it reduces the entropic cost of binding and allows for multiple partners to bind specifically but differently. Another recent study used ambient-temperature crystallography to identify multiple specific conformations at the active site of a proline isomerase whose interconversion was required for enzymatic turnover (Fraser et al., 2009).

In hindsight, a literalist interpretation of Anfinsen's dogma was at odds from the beginning with the ensemble paradigm from statistical thermodynamics. In other words, we can't think of proteins as "solid rocks".

For proteins whose cellular role is primarily architectural, stochastic fluctuations around a relatively static energy landscape may be the dominant form of flexibility. Enzymes and signal transducers like those described above, however, must additionally exhibit flexibility in response to environmental stimuli such as binding by ligands or other macromolecules. The near-native fluctuation and induced change paradigms overlap to some extent, in that intermolecular interactions reshape molecular energy landscapes such that new conformations become stochastically accessible. The mechanisms of interplay between these phenomena remains a fundamental open question, as modern structural biology continues to explore the biological impact of structural plasticity.

Backbone flexibility of a different sort is important for computational protein design. A major goal is to learn how protein backbone shifts in response to sequence changes, so that new proteins with desired functions can be engineered by modifying existing proteins with known structures. One promising avenue for mastering

backbone flexibility in protein design is to directly observe and categorize common modes of backbone motion within single structures, and then employ those changes as putative adaptations to engineered mutations *in silico*. However, in order to validate their use for designing man-made proteins, it is important to first confirm that these modes of motion directly accompany sequence changes in natural proteins.

Here I discuss one experimentally well-supported mode of local, mutation-coupled backbone motion, the backrub; observations of backrub-coupled sequence changes; and application of the backrub to protein design.

2.2 The backrub model for local backbone motion

The backrub (Davis et al., 2006) is a highly localized backbone motion tightly coupled to sidechain rotamer jumps (Figure 2.1), initially characterized by examining alternate conformations in ultra-high-resolution crystal structures (Figure 2.2). A simple geometrical model of the backrub consists of a small ($< 15^{\circ}$) rotation of a dipeptide about the axis between the first and third Cα atoms. Resulting strain in the N-Cα-C bond angle τ of all three residues may be partially alleviated and backbone H-bonding maintained with small counter-rotations of the two individual peptides. Note that this Cα formulation is a simplified but very close approximation of the real molecular mechanism, which probably involves a computationally unwieldy set of small shifts in 6-10 backbone torsion angles, as previously discussed (Davis et al., 2006). Backrubs were seen for 3% of the total residues in that study, and for 2/3 of the alternate conformations with a change in Cβ position – far exceeding the next most common shifts, which are either peptide flips or local shear in a turn of helix (Chapter 3).

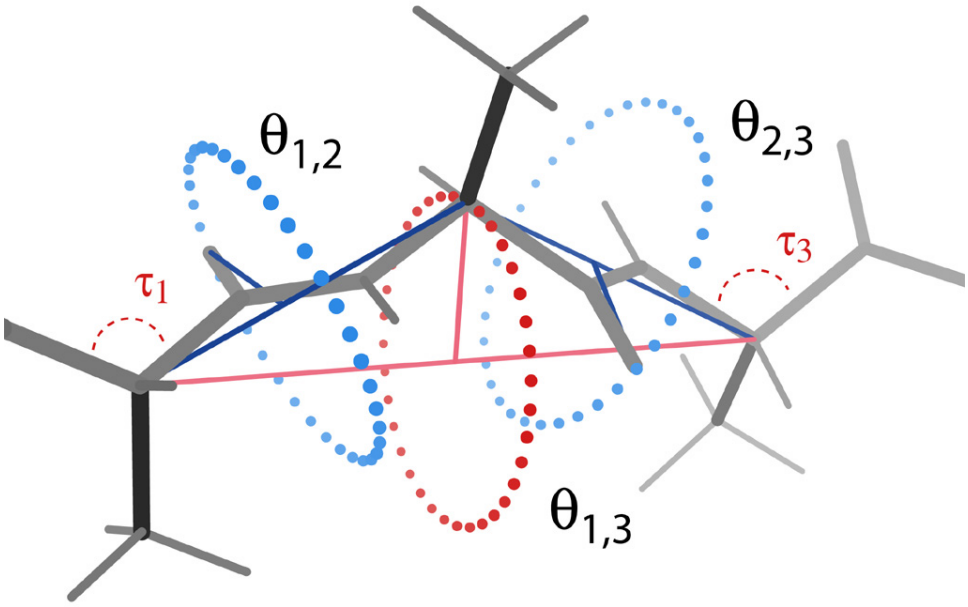

FIGURE 2.1: A schematic diagram of the backrub motion. The primary rotation ($\theta_{1,3}$) moves the central residue and its adjacent peptides around the red axis (Cα_{i-1} to Cα_{i+1}) as a rigid body, causing the central Cα to trace out the dotted circle. Secondary rotations ($\theta_{1,2}$ and $\theta_{2,3}$) move the individual peptides as rigid bodies around the blue Cα-Cα axes. A small amount of distortion is introduced into the τ angles (N-Cα-C), but they generally remain well within the range of values seen in typical crystal structures. Made for (Davis et al., 2006).

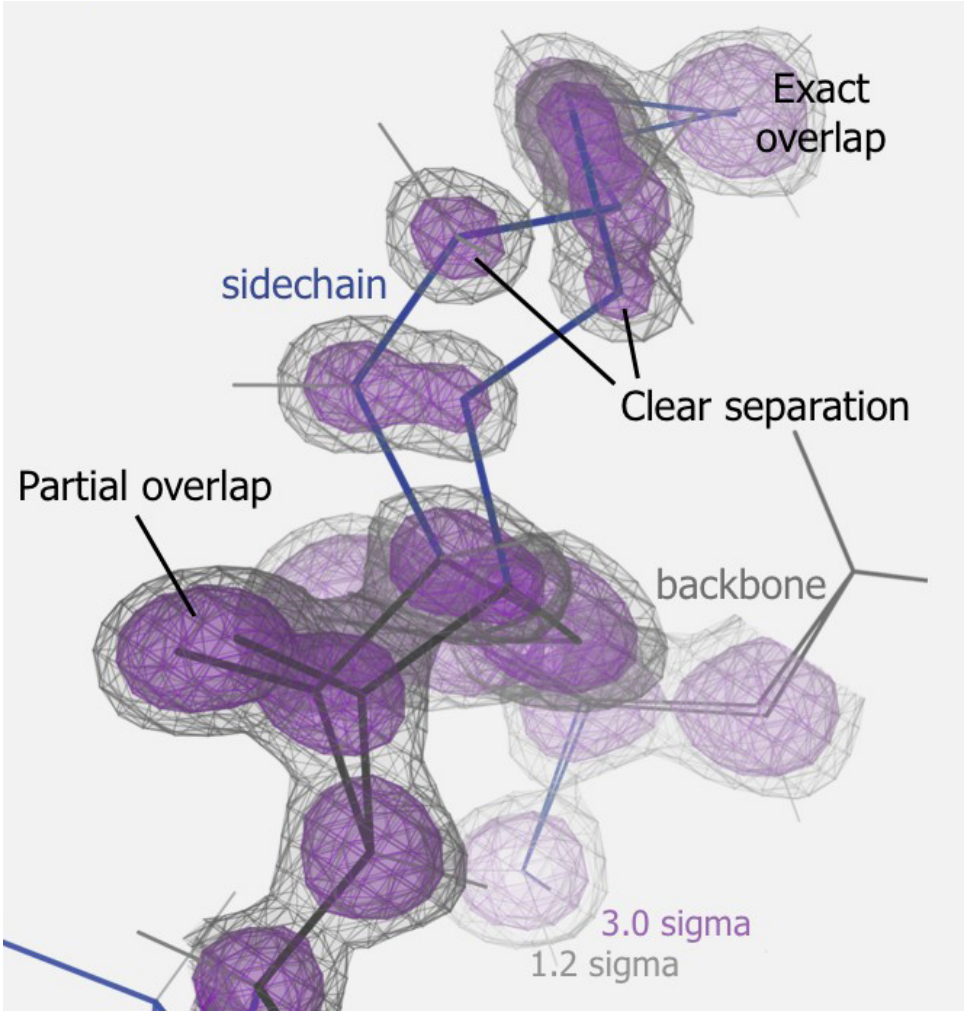

FIGURE 2.2: Example of a backrub at ultra-high resolution. 1us0 Lys100 has two different lysine rotamers that trace different routes, including different Cβ positions, to achieve the same placement for the terminal amino group. To accomplish this, the backbone must adjust for each rotamer; the difference is well modeled by a backrub motion. Credit: Ian Davis.

2.3 Natural backrub-coupled mutations

Several studies have successfully used the backrub approach to expand the search space of protein design efforts (Georgiev et al., 2008a; Smith and Kortemme, 2008; Donald, 2011) and improve agreement between computed sidechain dynamics and nuclear magnetic resonance (NMR) measurements (Friedland et al., 2008; Salmon et al., 2011). Recent work has shown that computational design of backbone structures generated by backrub sampling can recapitulate much of the sequence diversity found in the natural ubiquitin protein subfamily (Friedland et al., 2009) and by phage display experiments (Smith and Kortemme, 2011). However, the backrub has only been empirically demonstrated to accompany dynamic rotamer changes, not actual changes in amino acid identity. Importantly, no direct experimental evidence has been presented to support the assumption implicit in these studies that a dynamic, low-energy motion on the pico-to-nanosecond timescale is relevant on an evolutionary timescale. It would therefore be useful to confirm that backrubs accommodate real mutations in natural proteins in order to validate their as part of the repertoire of "moves" for protein design and other modeling efforts (Figure 2.3).

A naïve possibility is to directly compare wildtype and point mutant crystal structures and look for evidence of backrub-like backbone changes local to the mutation. However, backbone coordinate shifts due to backrubs are very small – on the order of the coordinate differences between crystal structures of the same protein (a few dÅ (deciangstroms)) (Kleywegt, 1999; Mowbray et al., 1999; DePristo et al., 2004), thus obscuring differences between genuine shifts and experimental noise. The initial description of the backrub bypassed this problem by comparing alternate conformations within single structures (Davis et al., 2006). However, this approach is not useful for comparing different molecules instead of different instances of the same molecule. Instead, I pursued a different approach, using the collective weight of

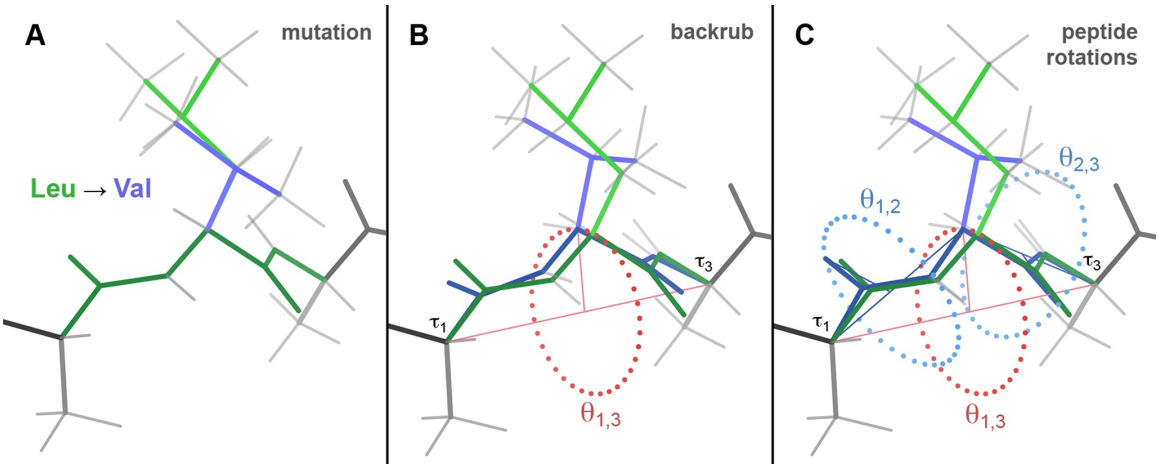

FIGURE 2.3: The backrub move for mutation-coupled local protein backbone adjustment. (A) A theoretical mutation in ideal β-sheet, from Leu in the *mt* rotamer (green) to Val in the *m* rotamer (blue) (Lovell et al., 2000), changes the interactions of the sidechain with its surroundings. Hydrogen atoms are shown in gray. (B) The primary backrub rotation angle θ1,3 (red dotted circle) rotates the dipeptide of interest around the Cα1-Cα3 axis (red line). As a result, the sidechain of residue 2 (the central residue) swings in a hinge-like manner. In the theoretical example shown, the space occupied by the new Val sidechain is now more similar to the space occupied by the Leu sidechain. (C) The secondary peptide rotation angles θ1,2 and θ2,3 (blue dotted lines) counter-rotate the individual peptides around the Cα1-Cα2 and Cα2-Cα3 axes (blue lines) to alleviate any strain introduced into the flanking τ1 and τ3 bond angles, respectively, and to restore H-bonding of the two peptides' amides and carbonyls, if necessary. The rotation angles, including the primary backrub angle θ1,3, define a motion, not a structure, and thus are meaningful only in reference to a pair of conformations (e.g. before vs. after or mutant vs. wildtype). Made for (Keedy et al., 2012).

many examples to ensure that observed local conformational differences were in fact genuine (Keedy et al., 2012). I focused on two very common, quite different, and representative structural motifs, each discussed in detail below.

2.3.1 Backrubs of α-helix N-caps

The N-cap or C-cap position of a helix is defined as the residue half-in and half-out of the helix: the peptide on one side of the cap makes standard helical backbone interactions, while the peptide on the other side has quite non-helical position and

interactions (Richardson and Richardson, 1988). α-helix N-cap residues can make several types of interactions that stabilize or specify the structural transition from loop into α helix, the most common and dominant of which is a sidechain-mainchain hydrogen-bond to the i+3 amide (Richardson and Richardson, 1988; Presta and Rose, 1988; Serrano and Fersht, 1989). (See Section 2.6 for a comparison between α-helix and 3_{10} N-caps.) N-cap H-bonds enhance proteins' stability by compensating for the loss of a mainchain H-bond at the helix start relative to the middle of a helix; as a modern philosopher put it, "micro machines ... strengthen with molecular bonds" (Deltron, 2000). Note that the sidechain cannot reach this H-bonding position if the residue has helical ϕ,ψ, so this interaction also specifies the exact helix start position and the direction from which the backbone can enter (Kapp et al., 2004).

Asn, Asp, Ser, and Thr are especially favored at N-caps because their sidechains have the proper chemical character and shape to mimic the helical backbone interactions (which Gln and Glu are too long to do). Notably, Asn/Asp sidechains are longer than Ser/Thr sidechains by one covalent bond, yet their H-bond distances (N-cap sidechain O to i+3 amide H) are only slightly shorter (2.01 $\pm$ 0.18 vs. 2.17 $\pm$ 0.18 Å) based on a survey of N-caps with i+3 H-bonds in the Top5200 database (see Section 5.2). This means the backbone must somehow slightly adjust to maintain similar H-bond geometry in both cases.

We first noticed that backrubs may explain such backbone adjustments upon mutation between short and long N-caps sidechains while examining N-caps in T4 lysozyme. Visual analysis using the backrub tool in KiNG revealed that a modest backrub (about 7°) nicely models the relationship between the Thr59 N-cap in the wildtype structure (2lzm) and the Thr59→Asn N-cap in the mutant structure (1lyg) (Bell et al., 1992) (Figure 2.4). Intriguingly, as well as being one of the most common N-caps, Ser is the most common amino-acid type for backrubs between al-

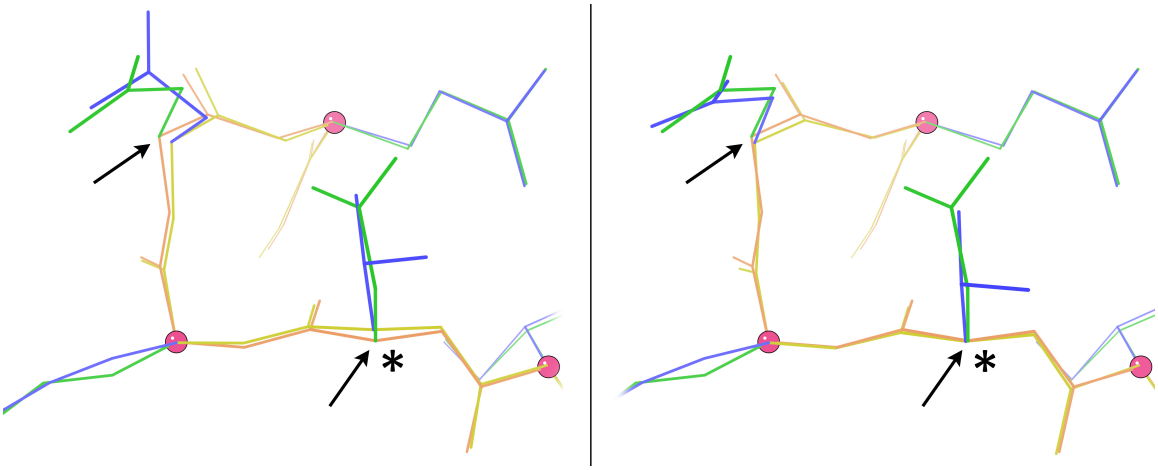

FIGURE 2.4: Motivation for exploring backrubs at N-caps. Left: The T49N mutant of T4 lysozyme (1lyg, orange backbone, green sidechains) differs from the wildtype (2lzm, yellow backbone, blue sidechains) by a mutation at the N-cap position (asterisk). Backbone differences that are suspiciously reminiscent of backrubs become apparent at the N-cap i and i+2 positions (arrows) after local superposition using the N-cap i-1, i+1, and i+3 Cαs (pink balls). Right: Manual backrub adjustments in KiNG interrelate the wildtype and mutant backbones nearly perfectly.

ternate conformations in crystal structures (Davis et al., 2006), perhaps because it has many distinct possibilities for sidechain-backbone H-bonding.

To ascertain whether or not backbone adjustments in response to N-cap mutations are backrub-esque more generally, I performed a stringent search for α-helix N-caps in the Top5200, resulting in identification of 429 Asn/Asp N-caps and a matching sample choice of 500 Ser/Thr N-caps (randomly selected from the 3208 total). The backbone conformations differ consistently: the longer Asn/Asp sidechains rotate the first turn's backbone away from residue i+3, while the shorter Ser/Thr sidechains pull the first turn's backbone toward i+3 in order to form the N-cap H-bond successfully (Figure 2.5; see also supplementary kinemage). When average Asn/Asp and Ser/Thr structures are superimposed using the Cαs surrounding the N-cap in the first turn (N-cap i-1 and i+1 to i+3), all Cαs match well ($<$ 0.05 Å) except the N-cap Cα itself (0.34 Å). The conformational difference at the N-cap position is

well modeled by a backrub rotation of about 11°, similar to shifts typical of rotamer-change backrubs. Furthermore, for both N/D and S/T, the Cβ deviations (Lovell et al., 2003) and the Cα-Cβ-Cγ bond angle distribution at N-caps are close to the general case distribution. This means the observed Cβ shifts and further leveraged sidechain shifts can be attributed primarily to backbone motion rather than altered covalent sidechain geometry.

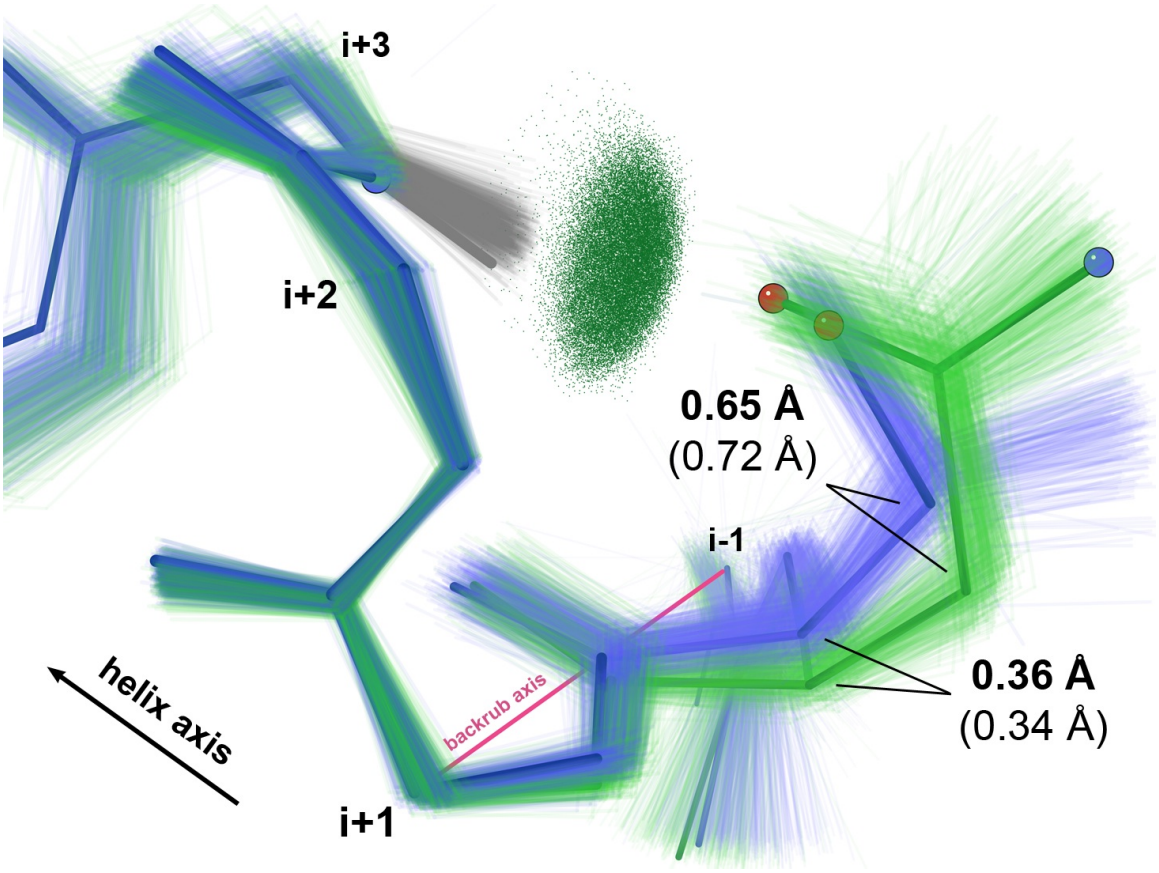

FIGURE 2.5: Backrubs at α-helix N-caps. Crystal structure ensembles for Asn/Asp (light green) and Ser/Thr (light blue) at α-helix N-termini are related by a backrub. Lowest-energy BRDEE conformations (see Section 2.4) for the N-terminus of an ideal α helix with Asn (dark green) or Ser (dark blue) at the N-cap position have a closely similar relationship. Cα and Cβ displacements between Asn/Asp and Ser/Thr for both average crystal structures (lighter, in parentheses) and low-energy BRDEE conformations (darker)evoke a hinge-like backrub operation. The ensemble i+3 sidechain-mainchain N-cap H-bonds are illustrated with "pillows" of green all-atom contact dots (Word et al., 1999b). Made for (Keedy et al., 2012).

I also examined two control cases with similar backbone geometry but different sidechain-mainchain interactions. First, I identified 538 α-helix N-caps with any amino acid type except Asn/Asp/Ser/Thr, in which case the i+3 sidechain-backbone H-bond is absent. Second, I chose 500 examples of mid-α-helix structure flanked by at least four α-helical residues in both directions, in which case the i+3 sidechain-backbone H-bond of an N-cap is satisfied by a usual i+4 backbone-backbone α-helical H-bond. The average Cα atoms for both control categories are in between the average Cα atoms for the Asn/Asp and Ser/Thr categories at the N-cap (or structurally equivalent) residue (see supplementary kinemage). This confirms that Asn/Asp and Ser/Thr N-caps are backrub-mediated excursions in opposite directions from equilibrium N-cap/helix structure.

2.3.2 Backrubs of aromatics in antiparallel β sheet

Aromatic residues often pair with glycine in antiparallel β sheet by adopting rotamers with $\chi 1 \approx +60^{\circ}$, which places the aromatic ring directly over a Gly on the adjacent strand across a narrow pair of backbone H-bonds (Richardson et al., 1992). Aromatic-glycine pairings in antiparallel β sheet have been demonstrated to yield a synergistic thermodynamic benefit (Merkel and Regan, 1998). If the opposite residue is changed to anything other than Gly, a sidechain including at least a Cβ atom is now present, which would sterically clash with the aromatic in its original conformation. However, the "plus $\chi 1$" aromatic rotamer will still be compatible with some rotamers of the opposite sidechain, provided that the aromatic may shift slightly to re-optimize packing of its ring against the opposite residue's Cβ hydrogens. Here we investigate whether backrubs enable this relaxation by excursions in both directions from a "neutral" β-sheet conformation. The leverage provided by such backbone motions could lean the aromatic residue forward/backward to maintain close interstrand contact when the identity of the opposite residue is changed to/from Gly.

A stringent structural motif search, similar to that described for N-caps above, identified 321 Phe/Tyr residues with "plus" $\chi 1$ rotamers in antiparallel β sheet. Aromatics are about three-fold as common in antiparallel vs. parallel β sheet, and are about twice as likely to adopt a plus $\chi 1$ rotamer when they do occur in antiparallel vs. parallel β sheet (data from Top5200), so we focused on antiparallel β sheet in this study. We took special care to avoid "frayed" examples for which the cross-strand pseudo-dihedral differed significantly in one vs. the other direction along the strand pair, i.e. where the sheet was beginning to "pull apart".

In 72 examples the amino acid on the opposite strand is a Gly, in which case the aromatic sidechain moves downward to contact the Gly Cα H. In the other 249 examples the Cβ H atoms of the amino acid on the opposite strand push the aromatic ring upward (Figure 2.6; see also supplementary kinemage). Pro cannot provide both β H-bonds, but all other non-Gly residues are equivalent in this role, since their sidechains must avoid the aromatic ring and present only Cβ H atoms toward it. When the two average conformations are superimposed onto each other using the aromatic i-2, i-1, i+1, and i+2 and opposite i Cαs, the surrounding Cαs match well (< 0.10 Å) but the central backrubbed Cα differs significantly (0.28 Å), as with the N-caps. The average Cβ deviation from ideality (0.05-0.06 Å) is far less than the outlier threshold (0.25 Å) (Lovell et al., 2003), and the average change in aromatic Cα-Cβ-Cγ bond angle is very small (0.6°, $< 1\sigma$), resulting in a < 0.05 Å shift of the Cζ contact point. These observations argue against the possibility that this large movement of the planar aromatic group is produced just by a bond-angle "hinge" with Cα or Cβ as the pivot. Rather, a dipeptide backrub rotation of about 11° (presumably 5-6° from neutral in each direction) almost perfectly interrelates the two average conformations.

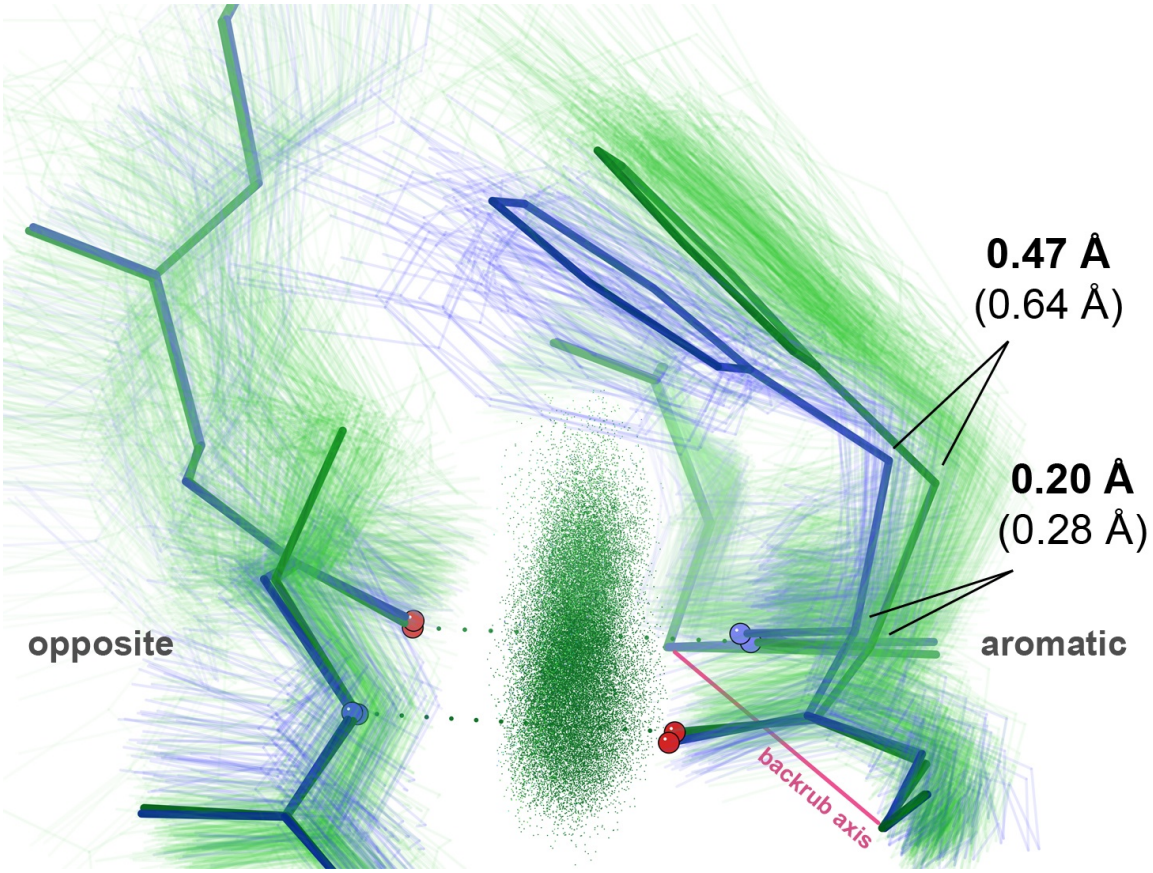

FIGURE 2.6: Backrubs at aromatic residues in antiparallel β-sheet. Crystal structure ensembles for Phe/Tyr across from Gly (light blue) vs. anything else (light green) undergo a backrub relative to each other. Lowest-energy BRDEE conformations for 1z84 Phe171 across from Gln188 (visually truncated at Cβ for clarity) (dark green) vs. Gln188 → Gly (dark blue) have a similar relationship. The aromatic Cα and Cβ displacements evoke a hinge-like backrub operation, both for average crystal structures (lighter, in parentheses) and for low-energy BRDEE conformations (darker). H-bonds are illustrated with "pillows" of green all-atom contact dots (Word et al., 1999b). Made for (Keedy et al., 2012).

2.4 BRDEE: Backrub Dead-End Elimination

Given this evidence, an enticing next step was to apply backrubs to protein design – essentially a computational analog of molecular evolution. Fortunately, I had willing and able collaborators in Bruce Donald in Computer Science and his student Ivelin Georgiev, purveyors and developers of the well-established dead-end elimination (DEE) algorithm.

The traditional DEE approach is to eliminate a "candidate rotamer" from further consideration if its energy is provably higher than that of a "competitor rotamer", i.e. if the following criterion holds:

$$E(i_r) + \sum_j \min_s E(i_r, j_s) > E(i_t) + \sum_j \max_s E(i_t, j_s) \tag{2.1}$$

where i_r is the candidate rotamer r at position i, i_t is the competitor rotamer t at position i, j_s is any other rotamer at any other position that contributes to pairwise energies, $E(i_r)$ is the self energy of i_r alone (with other self energy terms defined analogously), and $E(i_r, j_s)$ is the pairwise energy of i_r and j_s (with other pairwise energy terms defined analogously). This formalism essentially bins pairwise atom-atom interactions into pairwise residue-residue interactions, then iteratively prunes rotamers that provably cannot be part of the global minimum energy conformation (GMEC).

Unfortunately, minimization of models generated by this "rigid" version of DEE over any protein degrees of freedom, e.g. backrub rotations, destroys DEE's guarantee of identifying the GMEC. This is because after minimization, a model that was initially pruned may actually reach a lower energy than any model that was initially accepted. For example, if a model was initially pruned because one of its sidechains had a small steric clash, a backbone shift accomplished by backrub minimization could move the sidechain to alleviate that clash.

In previous related work, Bruce and Ivelin had devised DEE variants that enable provable DEE-style search while allowing some additional flexibility. For example, MinDEE allows sidechain χ dihedral minimization (Georgiev et al., 2008b) and BD allows global, empirically-based backbone motions based on combined ϕ,ψ changes that keep Cα atoms near their original positions (Georgiev and Donald, 2007). However, empirically motivated backbone motions such as backrubs had not yet been considered in a DEE framework.

I initially considered working with Bruce and Ivelin to (use inverse kinematics to) shoehorn backrubs into the only existing flexible-backbone flavor of DEE, BD (Georgiev and Donald, 2007), but this proved implausible due to the complexity of backrubs in ϕ,ψ space (Davis et al., 2006).

Instead, we selected for a new mutant of DEE, so to speak, this time with explicit (but discrete) backrub moves. The novel contribution of this algorithm, BRDEE (Georgiev et al., 2008a), is to extend traditional DEE to provably and exhaustively (Figure 2.7) search over backrub degrees of freedom in addition to sidechain rotamers. The BRDEE criterion (analogous to Equation 2.1) is as follows:

$$\begin{aligned} & E_{\ominus}(i_r) + \sum_j \min_s E_{\ominus}(i_r, j_s) - E_{t'_{\oslash}} - \sum_j \max_s E_{\oslash}(j_s) - \sum_j \sum_k \max_{s,u} E_{\oslash}(j_s, k_u) \\ > & E_{\oplus}(i_t) + \sum_j \max_s E_{\oplus}(i_t, j_s) \end{aligned} \tag{2.2}$$

where $E_{\ominus}(i_r)$ and $E_{\ominus}(i_r, j_s)$ are *lower* bounds on self and pairwise energies; $E_{\oplus}(i_t)$ and $E_{\oplus}(i_t, j_s)$ are analogously defined *upper* bounds; and $E_{\oslash}(j_s)$, $E_{\oslash}(j_s, k_u)$, and $E_{t'_{\oslash}}$ are *ranges* of possible self, pairwise, and template energies, respectively. Intuitively, Equation 2.2 is more stringent than the original DEE criterion (Equation 2.1) in that it requires the "best" (lowest-energy) conformation for the candidate rotamer i_r to be "better" (lower-energy) than the "worst" (highest-energy) conformation for

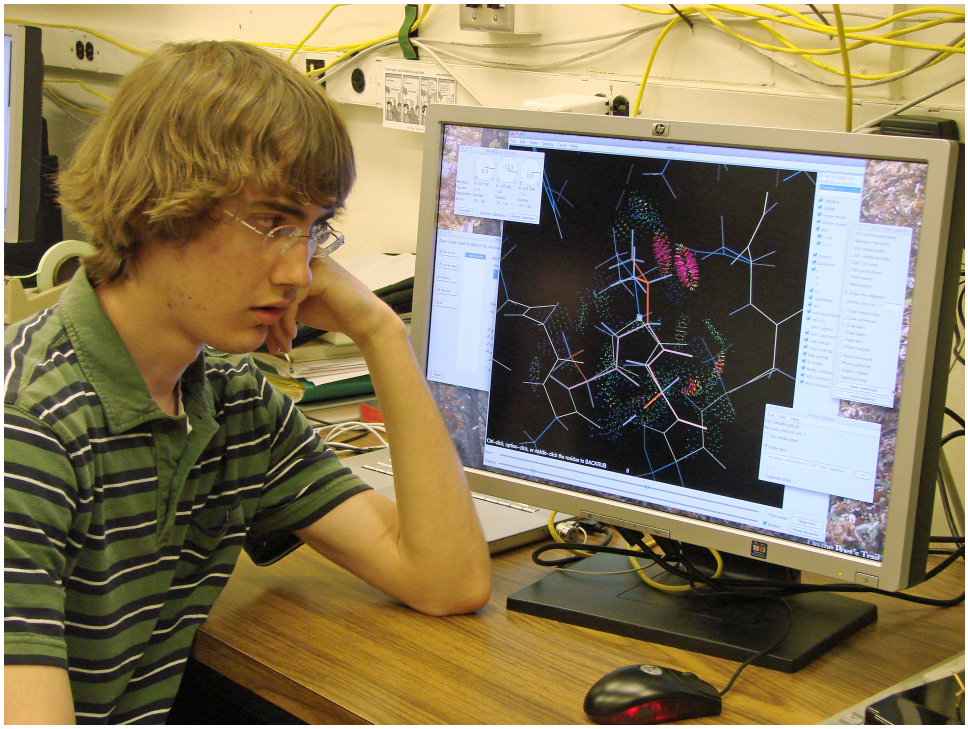

FIGURE 2.7: Before BRDEE, guaranteed identification of the global minimum energy conformation required exhaustive manual backrub sampling, e.g. in KiNG (Chen et al., 2009b).

the competitor rotamer i_t, even allowing for changes to the system introduced by backrub degrees of freedom.

With manual backrubs in KiNG (Davis et al., 2006; Chen et al., 2009b), single-peptide counter-rotations can be used to reestablish backbone carbonyl and amide vector orientations to maintain H-bonding interactions (see Section 2.2), but the (value of the) best counter-rotation is context-sensitive. For BRDEE, we decided to counter-rotate each peptide 70% of the way to complete restoration; this compromise was enough to often alleviate τ bond-angle distortions induced by the primary backrub rotation and effectively codified backrubs as single-parameter moves, at some cost in single-peptide variability.

I helped apply BRDEE to several systems, both to validate its ability to reproduce

natural backrubs and to determine its potential utility for protein design.

2.4.1 High-resolution alternate conformations

The original backrub paper (Davis et al., 2006) presented several convincing examples of alternate conformations visible in the electron density for ultra-high-resolution ($<$ 0.9 Å) crystal structures. These well-characterized backrubs present a useful test bed for BRDEE. The goal of these experiments was to recapitulate A-like and B-like conformations (i.e. conformations that match the rotamer as well as backrub direction and approximate magnitude of alternate A and B) and, equally importantly, avoid any other "decoy" conformations.

We investigated BRDEE's performance on four examples: 1muw A Val168, on the hydrophobic, buried surface of of a helix; 1n9b A Ile47, in the middle of a β sheet; 1gwe A Asp163, an α-helix N-cap (see Section 2.3.1); and 1dy5 B Met29, with a partially solvent-exposed, longer sidechain (Figure 2.8). After proper remodeling of 1n9b's backbone to reflect a split Cβ (Davis et al., 2006), A-like and B-like conformations were indeed recovered in every case. Moreover, if generated at all, decoys always scored worse than the crystallographically observed conformations. This was the case whether computations were initiated from the backbone and Cβ of alternate A or of alternate B, implying that backrubs alone are sufficient to nicely interrelate the two backbones. Note that the precise relative energies computed for the A-like and B-like conformations are not meaningful because the energy function used, a hybrid of molecular mechanics (Cornell et al., 1995) and a pairwise implicit solvation model (Lazaridis and Karplus, 1999), is simplified and inaccurate; however, the pruning of non-native-like conformations due to their unrealistically high energies is biophysically relevant and practically useful.

Another interesting observation came from comparison of BRDEE's top models for 1muw Val168. A-like and B-like conformations with m and t rotamers were the

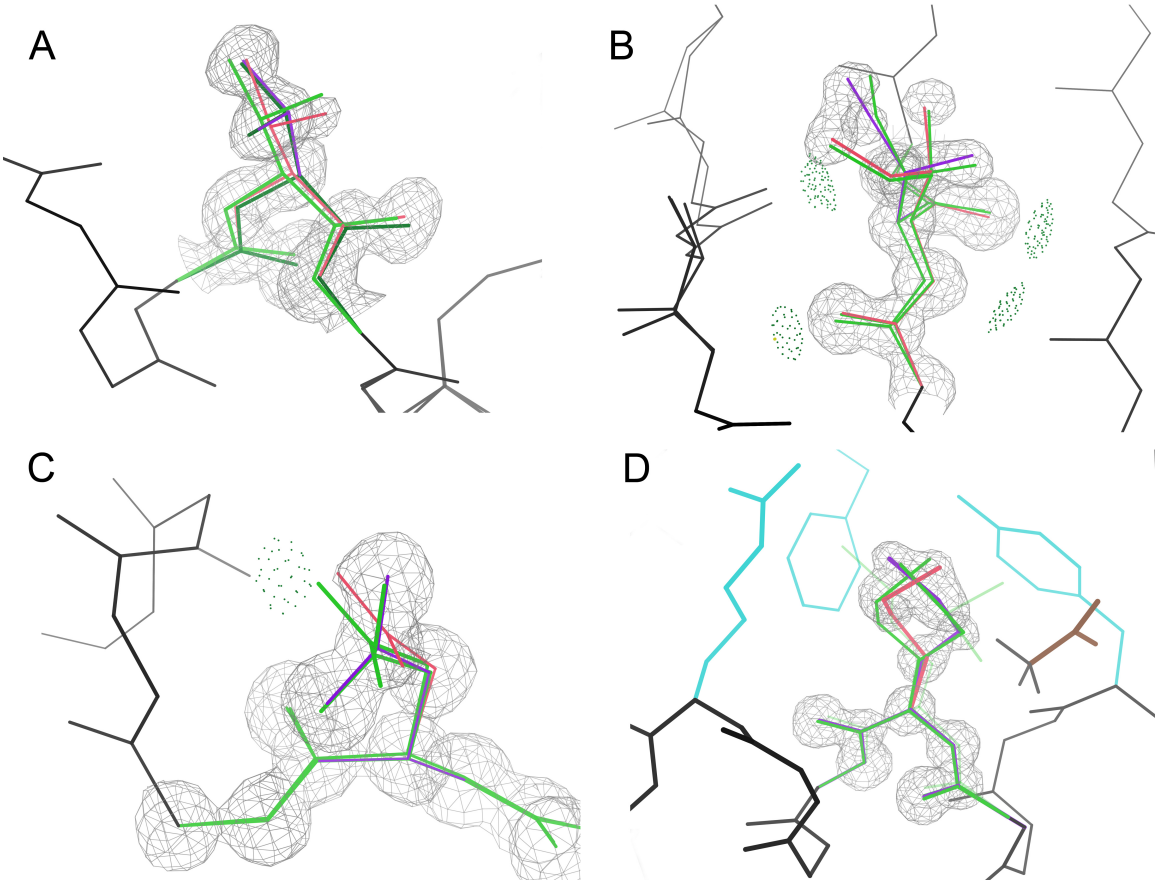

FIGURE 2.8: Alternate conformation backrubs recapitulated with BRDEE. The top two BRDEE conformations starting from the alternate A backbone and Cβ (green) match alternate A (purple) and alternate B (pink) of the deposited crystal structures. H-bonds are shown for two cases with green "pillows" (Word et al., 1999b). Chain A was used in each case except 1dy5, where the alternate sidechains of Met29 in chain B are better rotamers, have more closely matched occupancies, and fit the experimental electron density more clearly than in chain A. (A) 1muw Val168. The top two BRDEE conformations shown have been energy minimized with respect to their χ1 dihedral; see text for details. (B) 1n9b Ile47. The backbone and Cβs have been pre-split according to the original backrub study (Davis et al., 2006). (C) 1gwe Asp163. The i sidechain to i+3 mainchain H-bond is shown. (D) 1dy5 Met29. An acetate ligand (brown) was omitted from the BRDEE calculations, and would likely have sterically prevented the third and fourth best-scoring conformations (transparent green). The surrounding buttressing sidechains (cyan) cause the fifth best-scoring conformation (transparent green) to have significantly higher energy (by $>$ 13 kcal/mol) than the top two best-scoring conformations, which are native-like. The crystal structure has a single backbone, but the original backrub study (Davis et al., 2006) concluded that a small backrub better explains the electron density.

top conformations, as desired, but the third possible conformation with the p rotamer was also produced, albeit with considerably higher energy (5-13 kcal/mol worse). This energy gap between native-like and decoy BRDEE models is desirable in and of itself. Interestingly, though, *post hoc* sidechain dihedral minimization created an even wider energy gap, by lowering the native-like models' energies significantly (the sidechains became even more native-like in response to local packing constraints) but the decoy's energy less so. This result suggests that increased coverage of sidechain dihedral space in future algorithms could lead to better native/decoy discrimination. Therefore, one would like to have an algorithm that provably and simultaneously searches over both sidechain dihedrals and backbone degrees of freedom (such as backrubs and perhaps even other empirically motivated moves); I am continuing to collaborate with the Donald lab to define a version of DEE with precisely these capabilities (see Section 3.6).

Overall, BRDEE successfully reproduced experimentally observed rotamer-jump-coupled backrubs.

2.4.2 Natural backrub-coupled mutations

As a follow-up to Section 2.3, I wished to test whether a simple energy function based on molecular-mechanics terms from Amber (Cornell et al., 1995) and a solvation term from EEF1 (Lazaridis and Karplus, 1999) would recapitulate the empirically observed backbone changes, given the chance to access them via a backrub. This question is important for solidifying the connection between natural protein evolution and computational protein design, and can be addressed using BRDEE.

For the N-caps (Section 2.3.1), I created an idealized helix N-cap motif with a stretch of ideal helix (ϕ,ψ -60°,-40°) preceded by polyPro conformation (ϕ,ψ -80°,170°) for the N-cap and its previous residue. I prepared two versions of this motif as input to the algorithm, one with a short sidechain (Ser) and another with a long sidechain (Asn). I then used BRDEE to compute the lowest-energy model for each template, allowing backrubs up to 15° and rotamer changes at the N-cap as well as small i+3 peptide rotations. The lowest-energy Ser N-cap shifted "forward" whereas the lowest-energy Asn N-cap shifted "backward" in order to establish comparable hydrogen bonds in a manner remarkably similar to the empirically observed structures (Figure 2.5, supplemental kinemage and PDB files). In particular, the computed and observed Cα and Cβ shifts and inferred backrub angles are of similar magnitude and directionality.

For the β aromatics (Section 2.3.2), fewer examples, and more variation for β-sheet than for α-helix conformation, prevented the ideal-start calculation used for the N-cap case. Instead, low-energy conformations were computed by BRDEE for several examples judged to be appropriately representative of their respective type (across from Gly or across from other): 1gyh Gly122 and Gly122→Ala, 1khb Phe144 and Phe144→Gly, and 1z84 Phe171 and Phe171→Gly. In all cases, the lowest-energy conformation appears to match the average crystal structure very well, whether across

from Gly or from some other amino acid with a Cβ atom (Figure 2.6, supplemental kinemage and PDB files).

The BRDEE results recapitulate the average crystal structures, confirming the hypothesis that mutations between Ser/Thr and Asn/Asp at N-caps and between Gly and anything else across from aromatics in β sheet are well modeled by a backrub relationship. More generally, this implies that the backrub may reasonably accompany mutations during natural evolution or *in silico* protein engineering.

2.4.3 Core and active site redesign

To further establish the effect backrubs may have on sequence diversity, we used BRDEE to redesign two larger protein systems. The active site of the phenylalanine adenylation domain of the non-ribosomal peptide synthetase enzyme gramicidin S synthetase A (PheA for short), with the non-cognate substrate Leu instead of Phe, provided a solvent-exposed environment; the B1 domain of the immunoglobulin-binding protein G (GB1 for short) provided a contrasting hydrophobic environment (Georgiev et al., 2008a) (Figure 2.9).

Not unexpectedly, by accommodating mutant sidechains via backrubs, BRDEE made possible a greater range of sequences compared to traditional fixed-backbone DEE, as seen in Figure 2.10 for PheA (results were similar for GB1). For example, at position 301, only sequences including Gly were computed to be low-energy using traditional DEE, but sequences including Ala, Leu, Phe, Tyr, Gly, and Met were all computed to be low-energy using BRDEE (Figure 2.10). This greater variety of mutant sequencess is made possible because relatively subtle backrub adjustments can alleviate initial clashes created by modal rotamers instantiated on fixed backbones, which redounds to dramatically lower energies.

Altogether, these results bolster the idea that backrubs, which were originally demonstrated to apply only to single-sidechain dynamic rotamer hops, may in fact

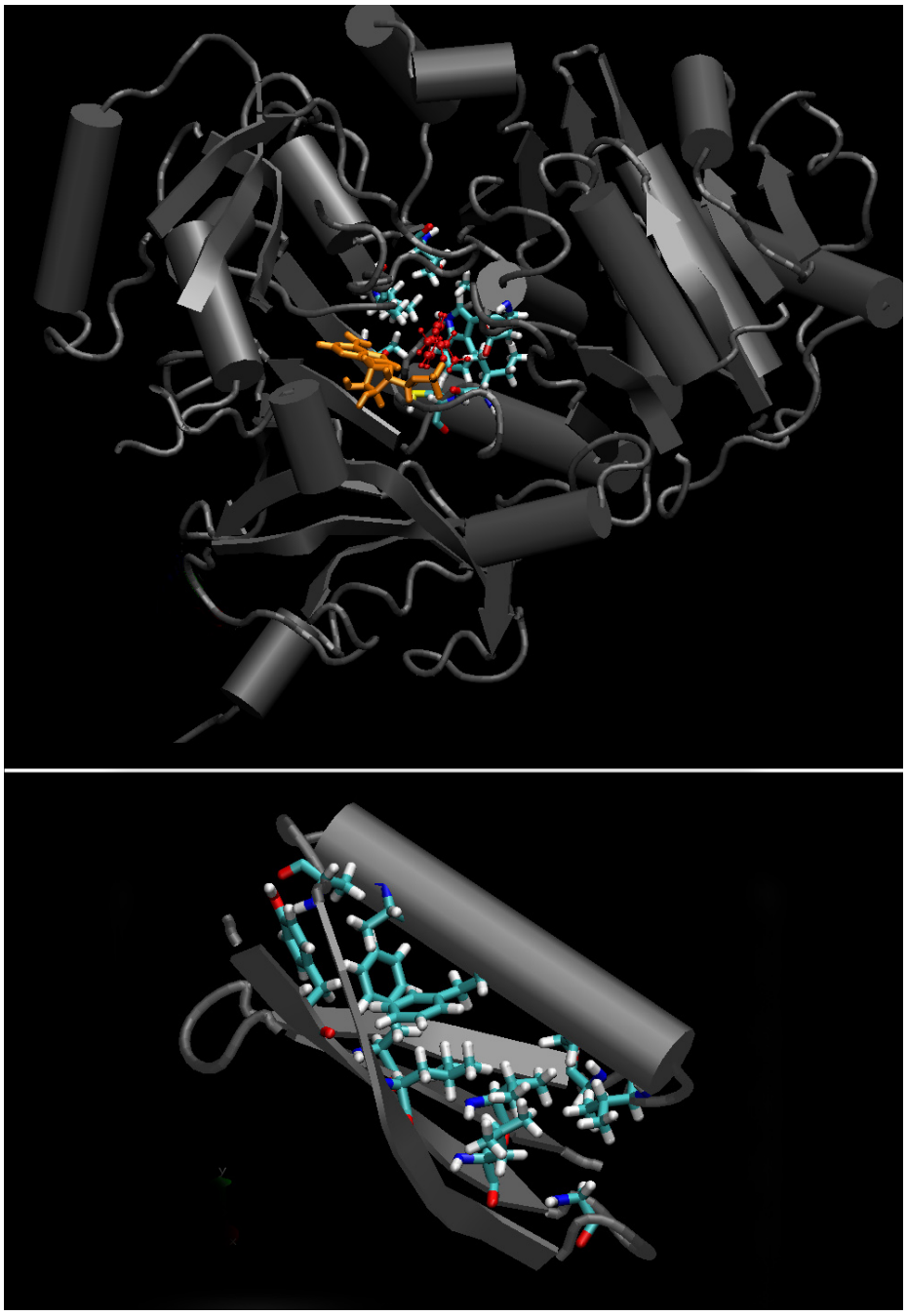

FIGURE 2.9: BRDEE redesign templates. Sidechains of residues subject to mutation and backrubs are shown in cyan. Top: active site of PheA (1amu), with the non-cognate substrate Leu in red and the PheA ATP cofactor in orange. Bottom: core of GB1 (1pga).

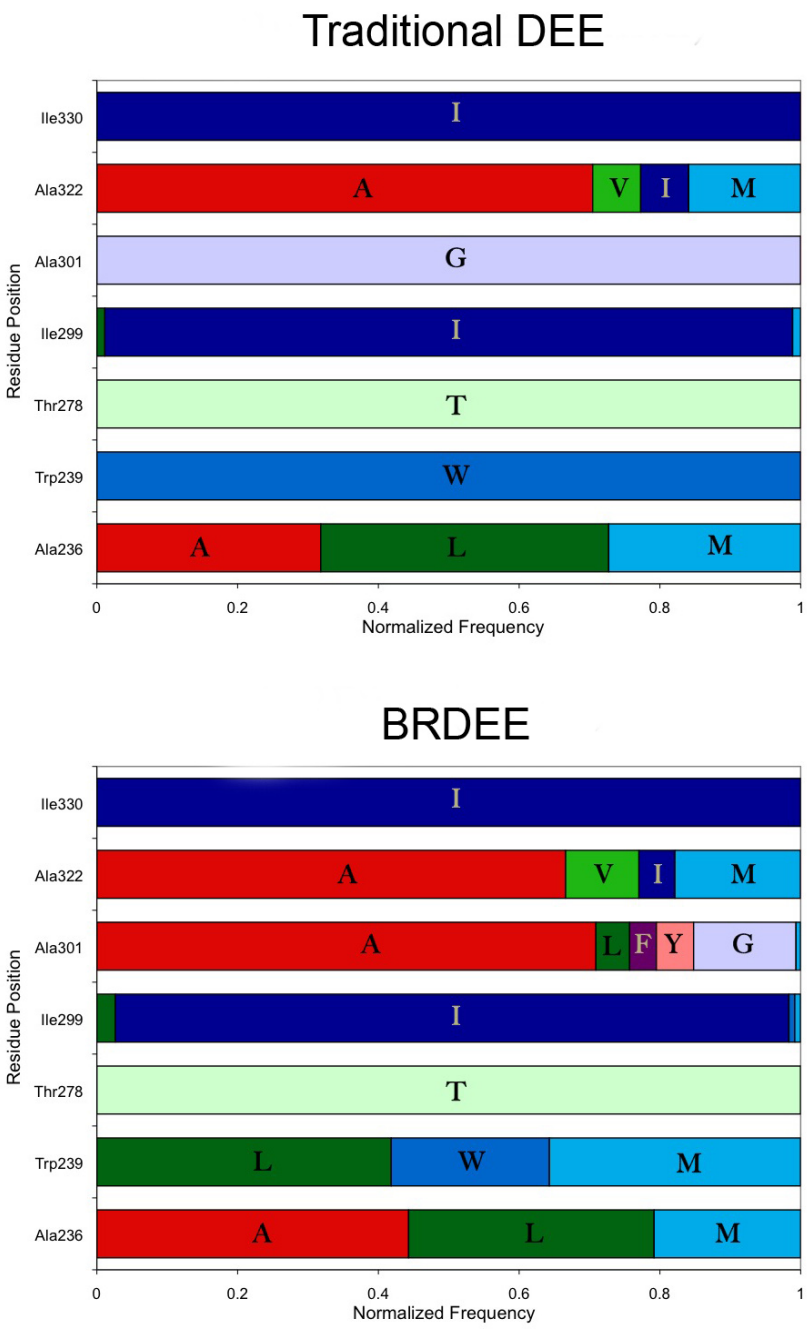

FIGURE 2.10: Backrub-enabled sequence diversity in the redesign of PheA for the non-cognate substrate Leu. Each row shows the distribution of mutations at a given active site residue position across the set of sequences whose energies are computed to be within 5 kcal/mol from the overall lowest BRDEE energy. A greater variety of sequences are possible with BRDEE (bottom) than with traditional DEE (top). Color coding of amino acid types is consistent across both panels and all residue positions. Adapted from (Georgiev et al., 2008a).

also describe backbone adjustments to fit mutated sidechains.

2.5 Stabilization of a redesigned PheA enzyme

Having helped to incorporate backrubs into a provably accurate design algorithm and established that they accompany natural mutations, I wished to use backrubs to tackle a prospective protein design problem. To that end, I considered previous work from the Richardson and Oas labs, which had showed that introduction of N-caps via mutagenesis can improve protein stability (Kapp et al., 2004). Because my earlier investigations also showed that backrubs can be important for accommodating mutations at N-caps, it seemed potentially productive to design N-cap motifs into a protein, using backrubs as necessary, to improve its stability.

PheA (Section 2.4.3) provided a useful test case. Its active site had been previously redesigned by the Donald lab to bind Leu instead of Phe; the resulting T278L/A301G double mutant displays a dramatic substrate specificity switch (Chen et al., 2009a). Subsequent "distal bolstering mutations" designed to be stabilizing had actually improved specificity for the non-cognate substrate Leu relative to the cognate substrate Phe in terms of the ratio of k_{cat}/K_M values, although it remains unclear whether this effect was due to global stabilization of the enzyme to compensate for destabilization by previously designed active site mutations or to some other mechanism (Chen et al., 2009a). Nevertheless, it seemed reasonable to hope that distal N-cap mutations designed to be stabilizing might also improve catalytic specificity.

With this goal in mind, Ivelin Georgiev and I designed N-caps in T278L/A301G PheA distal from the active site (Figure 2.11) using their protein design software. Preliminary calculations with a flexible-sidechain but fixed-backbone algorithm, MinDEE (Georgiev et al., 2008b), showed that N-cap mutations could potentially stabilize the protein. Subsequent calculations with BRDEE incorporating up to 10° backrubs at the N-cap and i+3 residues showed the promise of forming traditional

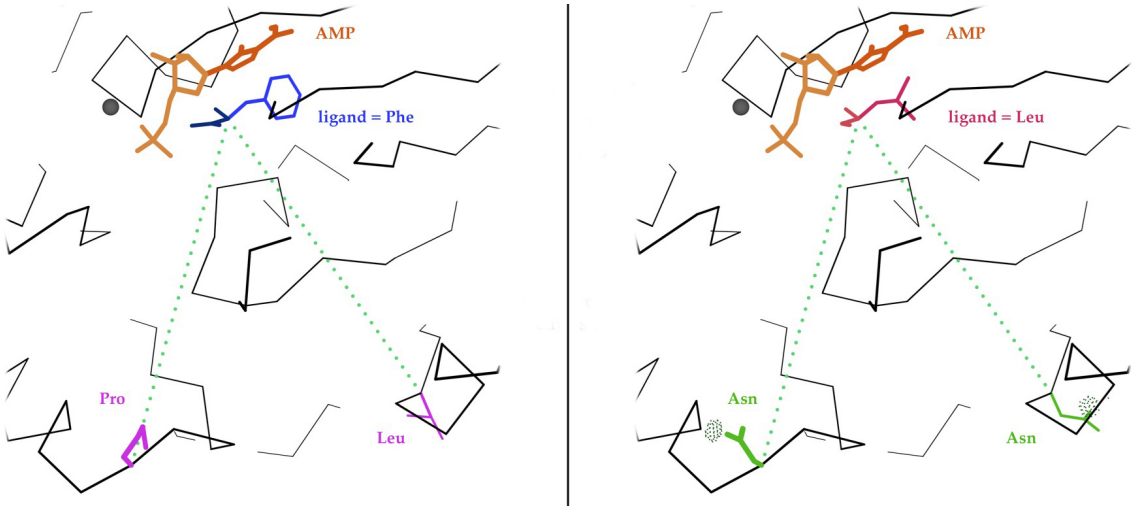

FIGURE 2.11: PheA N-cap mutations are distal from the active site. Left: Crystal structure of wildtype PheA enzyme (1amu) with the cognate substrate Phe (blue), cofactor Mg (gray), product-state cofactor AMP (orange), and distal Pro124 and Leu255 sidechains (magenta). Right: *In silico* model of the redesigned enzyme with new substrate Leu (pink), putatively stabilized by Pro124→Asn and Leu255→Asn (green). Green dotted lines indicate distances of 22 and 20 Å from the Cαs of residues 124 and 255 respectively to the Cα of the ligand. Green "pillows" represent modeled N-cap interactions: i sidechain to i+4 mainchain H-bonds (Word et al., 1999b).

N-cap i+2, N-cap i+3, and "capping box" H-bonds; notably, slight backrubs often helped to properly position the sidechains.

In order to experimentally test these predictions *in vitro*, I expressed, purified, and tested stability and catalysis of two of the best-scoring N-cap mutant proteins, P124N and L255N (on top of the T278L/A301G mutant background), in the Donald wet lab (with invaluable help from Cheng-Yu Chen). Cheng-Yu guided me through the steps of site-directed mutagenesis, bacterial transformation, induction of expression, and purification with a nickel column followed by FPLC chromatography. A Bradford assay (Bradford et al., 1976) showed that the final concentrations were 33.7 mg/mL for P124N and 46.9 mg/mL for L255N, and an SDS-PAGE gel showed that both both mutant proteins were very pure, possibly $>$ 90% (Figure 2.12).

The biochemical results relative to the starting T278L/A301G construct were

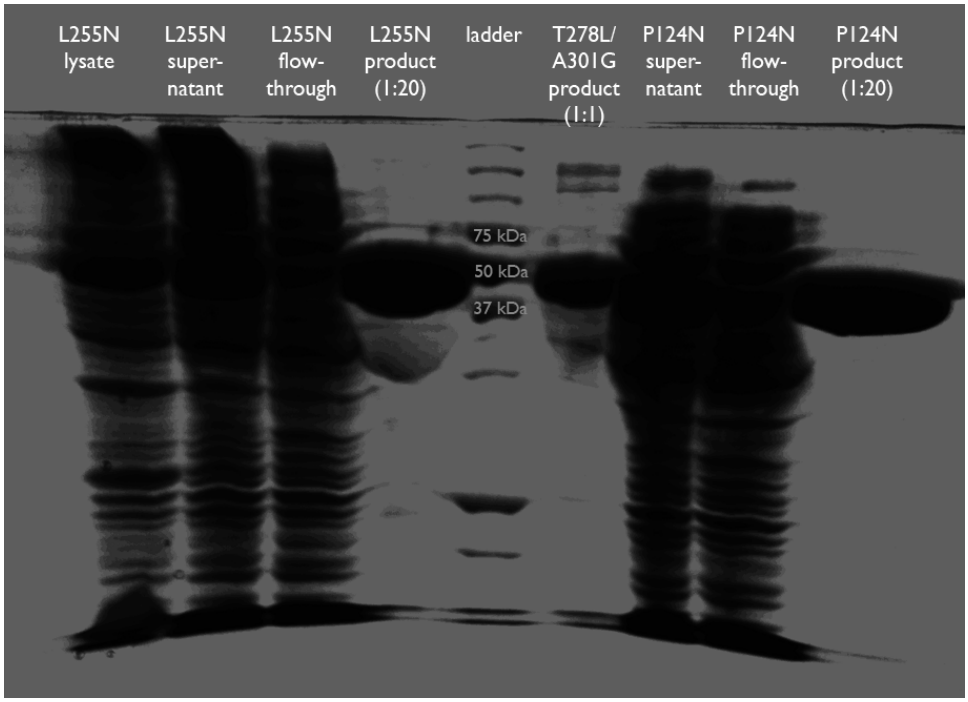

FIGURE 2.12: SDS gel with pure N-cap mutants of PheA. Byproducts of the purification process (whole-cell lysate, post-centrifugation supernatant, flow-through from nickel column) for both of the N-cap mutants (L255N and P124N in the T278L/A301G background) are quite diffuse on this SDS-PAGE gel, but the final products show sharp bands corresponding to both the proper molecular weight portion of the ladder (about 62 kDa) and a previously purified T278L/A301G provided by Cheng-Yu Chen. (The T278L/A301G/P124N lysate was omitted by accident.)

inconclusive, yet suggestive. P124N showed an approximately 20% reduction in k_{cat}/K_M (Table 2.1) based on a steady-state pyrophosphate-release assay used previously in the Donald lab (Stevens et al., 2006). It also showed no observable stabilization based on circular dichroism using chemical denaturation with guanidinium chloride (Figure 2.13). L255N, on the other hand, showed an approximately 80% increase in k_{cat}/K_M, due mostly to lower K_M (Table 2.1), and its CD curve hints at a slight stabilization (Figure 2.13) – a promising result. However, it is important to keep in mind that experimental error is significant with these experiments: the pyrophosphate-release assay for activity (Stevens et al., 2006) is somewhat unreli-

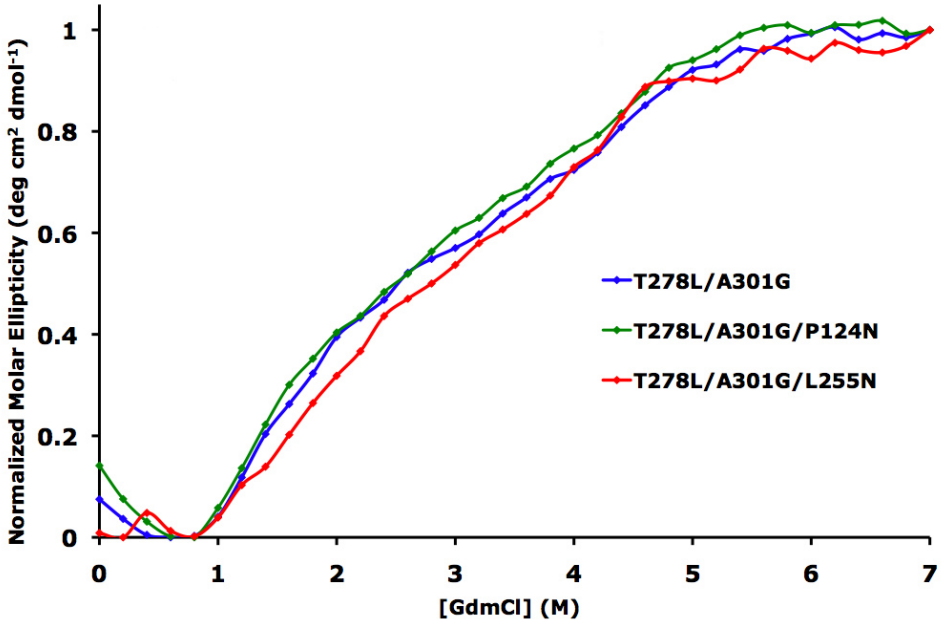

FIGURE 2.13: PheA N-cap mutant stability curves. The concentration of guanidinium chloride was slowly increased from 0-8 M to chemically denature the proteins. Circular dichroism molar ellipticity was measured at 222 nm wavelength for T278L/A301G and T278L/A301G/L255N and 230 nm wavelength for T278L/A301G/P124N. To normalize molar ellipticity for each of the three experiments, each value was expressed as a fraction of the observed ellipticity range across the guanidinium chloride range.

able, and CD curves are difficult to interpret for such a large (62 kDa) protein due to cooperatively folding substructures within the overall molecule. Further work, including repeating the experiments and combining several such "bolstering" mutations for a potentially additive effect, is required to ascertain the precise benefits (or lack thereof) of L255N.

Table 2.1: PheA N-cap mutant enzyme kinetics. Activities were measured by a steady-state pyrophosphate-release assay. *Data for comparison provided by Cheng-Yu Chen.

Enzyme	Substrate	k_{cat} (min^{-1})	K_M (mM)	k_{cat}/K_M (min^{-1} mM^{-1})
wildtype*	L-Phe	1.73 ± 0.29	0.0018 ± 0.0004	951.4 ± 111.2
wildtype*	L-Leu	28.74 ± 1.58	6.98 ± 1.00	4.15 ± 0.36
T278L/A301G*	L-Phe	3.37 ± 0.08	0.097 ± 0.013	34.94 ± 4.76
T278L/A301G*	L-Leu	1.16 ± 0.10	0.015 ± 0.002	79.49 ± 13.67
T278L/A301G/P124N	L-Leu	0.74	0.011	66.23
T278L/A301G/L255N	L-Leu	1.17	0.0085	137.75

2.6 α vs. 3_{10} N-caps

To understand why the N-cap mutations didn't yield more clear-cut results, we looked closer at the starting point for the designs, the PheA crystal structure. Most helices in PheA were already capped by a canonical Asn/Asp/Ser/Thr i to i+2 or i+3 hydrogen bond. The three N-caps we tried to design lacked such a traditional motif, but upon closer inspection, we realized all three actually had 3_{10} rather than α-helical backbone in the first helical turn. Our struggles to use these existing coordinates to straightforwardly design stabilizing N-caps suggested that nature has at least two ways to handle the start of a helix: an N-cap sidechain H-bond or 3_{10} backbone.

To further investigate this idea, I classified each N-cap in our Top5200 set. H-bond pseudo-energies were computed as in DSSP with the standard -0.5 kcal/mol cutoff (Kabsch and Sander, 1983). N-caps with an i+4 but not an i+3 mainchain-mainchain H-bond were labeled α, N-caps with an i+3 but not an i+4 mainchain-mainchain H-bond were labeled 3_{10}, and N-caps with both i+3 and i+4 mainchain-mainchain H-bonds were labeled bifurcated $\alpha/3_{10}$ and were henceforth ignored.

Asn/Asp/Ser/Thr were indeed found to be strongly preferred (by factors of 2.5-3) at α-helix N-caps relative to protein structure in general (Figure 2.14). Gly is next most common, but cannot form or be influenced by an N-cap H-bond.

On the other hand, Pro has a 2.5-fold spike of preference at 3_{10}-helix N-caps. Such motifs may be favorable due to van der Waals interactions between the Pro ring and the i+2 and i+3 sidechains of the first turn, as seen in Figure 2.15. The majority of 3_{10} Pro N-caps have β ϕ,ψ (86%, 403/471), which presents the Pro sidechain toward the first turn. Nearly all (96%) of those proline rings make van der Waals contact with first-turn sidechains.

Perhaps not coincidentally, Pro is also favored at the i+1 position immediately following an α-helix N-cap (Richardson and Richardson, 1988). Proline is good in the

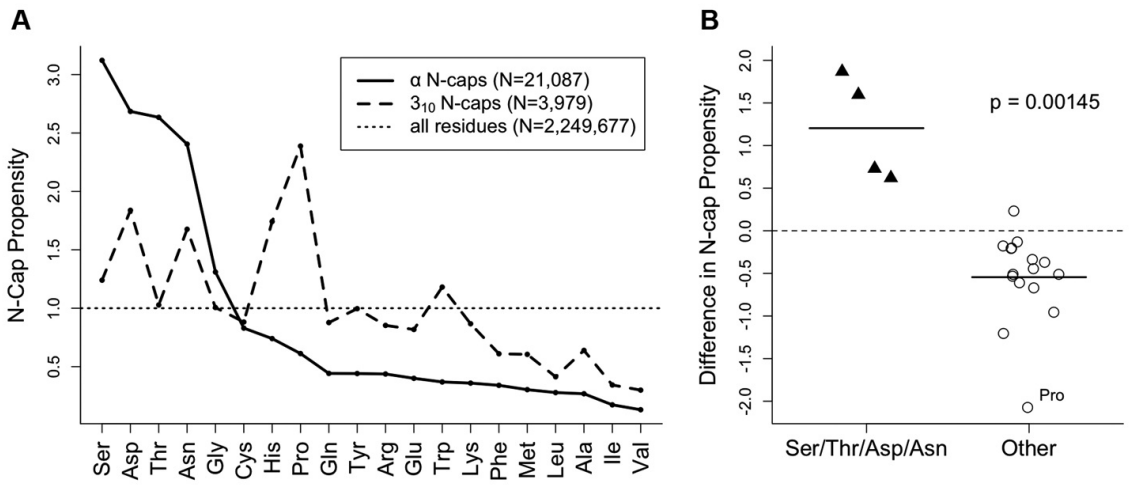

FIGURE 2.14: α vs. 3_{10} N-cap propensities. (A) The 20 amino acid types are shown ranked according to their α-helix N-cap propensity (solid line), defined as the fraction of α-helix N-cap residues of the given amino acid type, divided by the fraction of general case residues of that amino acid type (dotted line), all using the Top5200 data set. The correlation with the analogously defined 3_{10}-helix N-cap propensity (dashed line) is surprisingly weak, except for both slightly disfavoring hydrophobics. For example, Ser/Asp/Thr/Asn are the most common N-caps for α-helix but are not especially favored as N-caps for 3_{10}-helix. Some other hydrophobic amino acids like Ala/Ile/Val are uncommon as either type of N-cap. (B) The canonical α-helix N-caps Ser/Thr/Asp/Asn (triangles) are grouped separately from the other 16 amino acid types (circles); the two groups are compared based on the difference between α-helix N-cap propensity and 3_{10}-helix propensity. The horizontal dotted line at 0.0 indicates neither an increase nor a decrease in preference for α-helix N-caps instead of 3_{10}-helix N-caps. A one-tailed Mann-Whitney test shows with 95% confidence (p-value = 0.00145 $< \alpha = 0.05$) that Ser/Thr/Asp/Asn are statistically unique in terms of their specificity for α-helix N-caps.

first turn of any helix type, because the ring interferes sterically with any potential preceding helix turn, removes one unsatisfied NH, and entropically favors helical ϕ,ψ because it has fewer other possibilities than other residues. Pro slightly prefers a preceding residue that is not in helical conformation, so it is more common in N-cap i+1 than i+2 or i+3, but it cannot make either the N-cap sidechain-mainchain H-bond or the reciprocal "cap-box" i+3 sidechain H-bond to the N-cap NH (Kapp et al., 2004), so it is rare as an α-helix N-cap. That is presumably not a disadvantage for

3_{10}-helix where those interactions do not occur, so Pro is well suited as a 3_{10} N-cap.

This suggests the possibility that helix N-termini can transform from more tightly-wound 3_{10} (3 residues/turn) to looser α-helix (3.6 residues/turn), or vice versa, by deletions or insertions in the preceding loop. For example, a canonical α N-cap motif, with Asn/Asp/Ser/Thr at position i and Pro at position i+1, could be transformed to a 3_{10} Pro N-cap motif if deletion of residues in the preceding connection/loop pulled the first turn tighter. I found no evidence of an increased prevalence (relative to general-case protein structure) of vestigial Asn/Asp/Ser/Thr at the i-1 position immediately preceding extant 3_{10} Pro N-caps, suggesting that such residues may change identity once no longer useful. Conversely, a 3_{10} Pro N-cap could be the precursor to a canonical α N-cap motif if insertion of residues in the preceding connection/loop added "slack" to the first turn, in which case the Pro would become the N-cap i+1 and the preceding residue would become the new α N-cap. Either before or after the insertion and conformational change, this residue could mutate to Asn/Asp/Ser/Thr to solidify a new α N-cap motif.

Also note that Asn/Asp are preferred at 3_{10} N-caps but Ser/Thr are neutral, despite both being strongly preferred at α N-caps (Figure 2.14). As mentioned in the main text, with α N-cap conformation both categories of sidechains form strong i+3 sidechain-backbone H-bonds, but with 3_{10} N-cap conformation this interaction is not possible. Asn/Asp instead form strong i+2 sidechain-backbone H-bonds, and are therefore still preferred as 3_{10} N-caps (albeit slightly less so than as α N-caps). Ser/Thr, on the other hand, are slightly too short and form only weak i+2 sidechain-backbone H-bonds, and are therefore not preferred as 3_{10} N-caps. However, they are not disfavored either; these net neutral 3_{10} N-cap preferences suggest that any unfavorable energetics may be counteracted by the small benefit enacted by the weaker i+2 H-bonds.

In summary, even the relatively subtle and localized structural difference between

α and 3_{10} conformation at a helix N-terminus can have dramatic effects on the preferred sequence. This observation underscores the critical role backbone flexibility can play, both in the natural biological world and for rational protein design efforts. It also highlights the synergistic interplay possible between structural bioinformatics and interactive graphics: by investigating representative examples (and, sometimes even more interestingly, outliers) identified by statistical analysis, one can discover potentially fundamental evolutionary relationships (Figure 2.15) in a data-driven way.

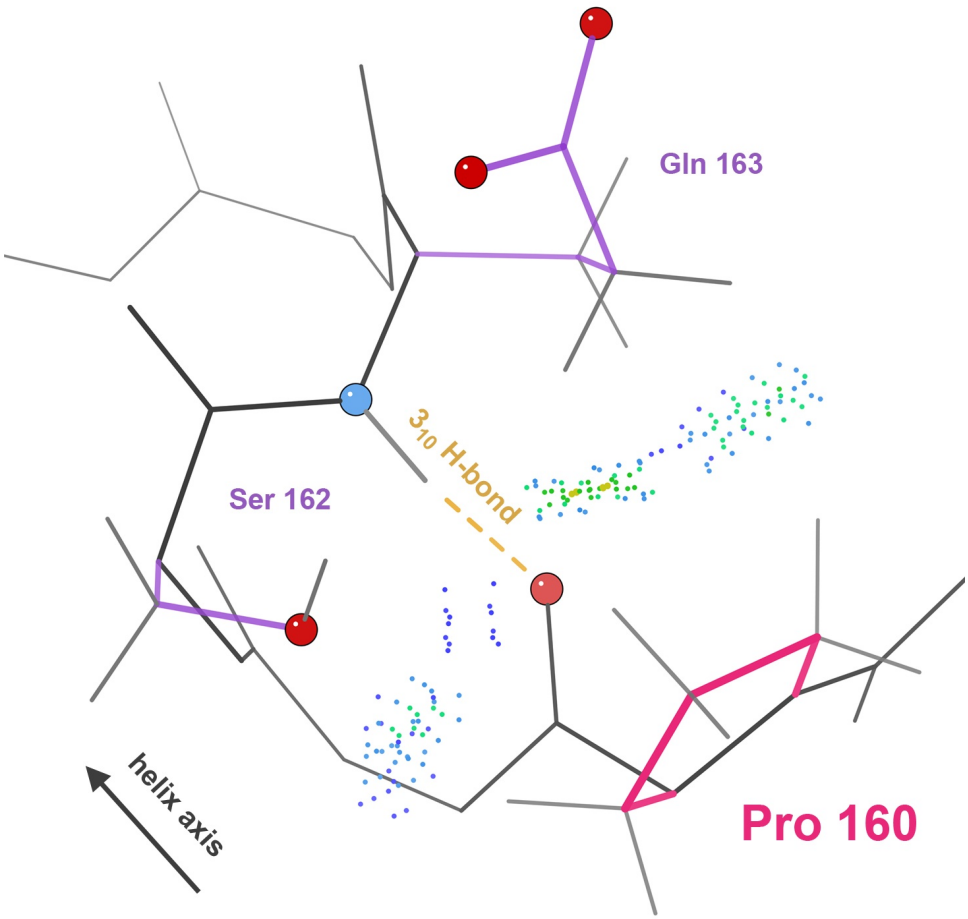

FIGURE 2.15: The ring of Pro160 (pink) in 1m79 chain A makes van der Waals interactions (green/blue dots) with the sidechains of Ser162 and Glu163 (purple). These close interactions are possible because the 3_{10}-helical i+3 mainchain-mainchain H-bond (gold dashed line) pulls the Pro close enough to the first-turn sidechains. Note that Ser at position i+2 and Glu at position i+3 are among the most common identities given a 3_{10} Pro N-cap.

2.7 Discussion

Before I began my studies of local protein backbone motion, it was known that backrubs allow dynamic transitions between rotamers (Davis et al., 2006), but there was no direct evidence that they allow transitions between amino acid types (i.e. mutations) and are therefore suitable for inclusion in protein design calculations.

To fill this gap in our knowledge, I first used ensembles of carefully curated, high-quality examples of local motifs to confirm that backrubs accompany at least some specific amino acid differences (Section 2.3). Next, I worked with collaborators to implement backrubs in a provably accurate protein design algorithm – the first of its type (Section 2.4). I then helped apply this procedure to recapitulate natural backrub-coupled rotamer jumps (Section 2.4.1) and sequence changes (Section 2.4.2), expand the portions of realistic sequence space accessible to active site and hydrophobic core designs (Section 2.4.3), and bolster a previously redesigned enzyme (Section 2.5). The partial failures of the latter experiment due to 3_{10} instead of α backbone further emphasized that small backbone changes can have amplified energetic effects (Section 2.6).

Given this evidence that backrubs accommodate sequence changes, one would suspect that they help proteins play host to accumulated mutations during the process of evolution. The N-cap and β aromatic studies do not directly address such true evolutionary relationships between proteins, but rather substantiate the idea that backrubs enable single amino acid changes at specific motifs. Nevertheless, those studies do indirectly support the hypothesis that backrubs accommodate individual mutations in specific proteins, thereby aiding actual evolution within protein families (Friedland et al., 2009); future work will be needed to further investigate this idea.

An interesting related question is the extent to which a mutant backbone may be

selected from the discrete backbone states possible for the wildtype residue (Tokuriki and Tawfik, 2009) as opposed to from a more continuous range of generic possibilities. For backrubs, backbone coordinate changes are very small, and it is likely that in most cases the mutant backbone can access the entire continuous $\pm$5-10° range, with the final choice dictated by the new sidechain's dealings with the (relatively fixed) local structural environment.

Rare exceptions may occur when that environment is exquisitely well packed (i.e. the local "structural memory" (Anfinsen, 1973) is strong), the wildtype residue has alternate rotamers related by a backrub, and at least one rotamer of the mutant amino acid is very chemically/structurally similar to at least one of the wildtype sidechain's alternate rotamers (e.g. Val-Thr, Ser-Thr, Val-Ile, etc.). In such a case, the mutant sidechain may fit in by approximating the pre-existing alternate wildtype rotamer to which it is most similar, in which case the mutant backbone would likely match the corresponding wildtype backrub state. For larger, more distributed/global/discrete backbone changes – such as peptide flips or other even larger crystallographically visible alternate backbone conformations – the notion that the mutant backbone is essentially selected from among a discrete set of pre-existing possibilities seems more plausible, since there are more interactions specifying a unique backbone conformation for long than for short backbone segments; a single mutation is unlikely to overwhelm this confluence of conspiring forces.

The situation is similar from the perspective of a rational protein engineer. For small backbone changes such as backrubs, it is desirable to completely explore the relevant backbone conformational space using an algorithm such as BRDEE, evaluating each possibility using a molecular mechanics energy function. For larger backbone changes, on the other hand, it may be more useful to propose possibilities based on pre-existing, experimentally observed structural heterogeneity; Section 3.6 of the next chapter offers guidance in this direction.

3

Shears

3.1 Fishing for new backbone motions

Backrubs are the most common type of local backbone motion, and are extremely well validated (see Chapter 2). However, despite their widespread occurrence, they offer only partial coverage of the conformational space available to protein backbone via local adjustments. Therefore, a pressing need is additional models of local backbone motion, equally well supported by empirical observations, to supplement backrubs.

To this end, I underwent an exploratory study to identify novel modes of backbone motion. I compared a number of motifs with distinct subsets of conformations, but unfortunately found no simple, easily describable backbone differences.

For example, I compared mid-helix Asn with the *m-80* rotamer (H-bond to *i*-4 carbonyl oxygen) vs. the *m-20* rotamer (van der Waals to *i*-4 carbonyl oxygen) (Lovell et al., 1999). It initially seemed plausible that interacting with the preceding helical turn in one way vs. another might require a small backbone adjustment to properly position the sidechain. Unfortunately (from the perspective of a prospective protein modeler requiring new backbone moves), ensembles of superim-

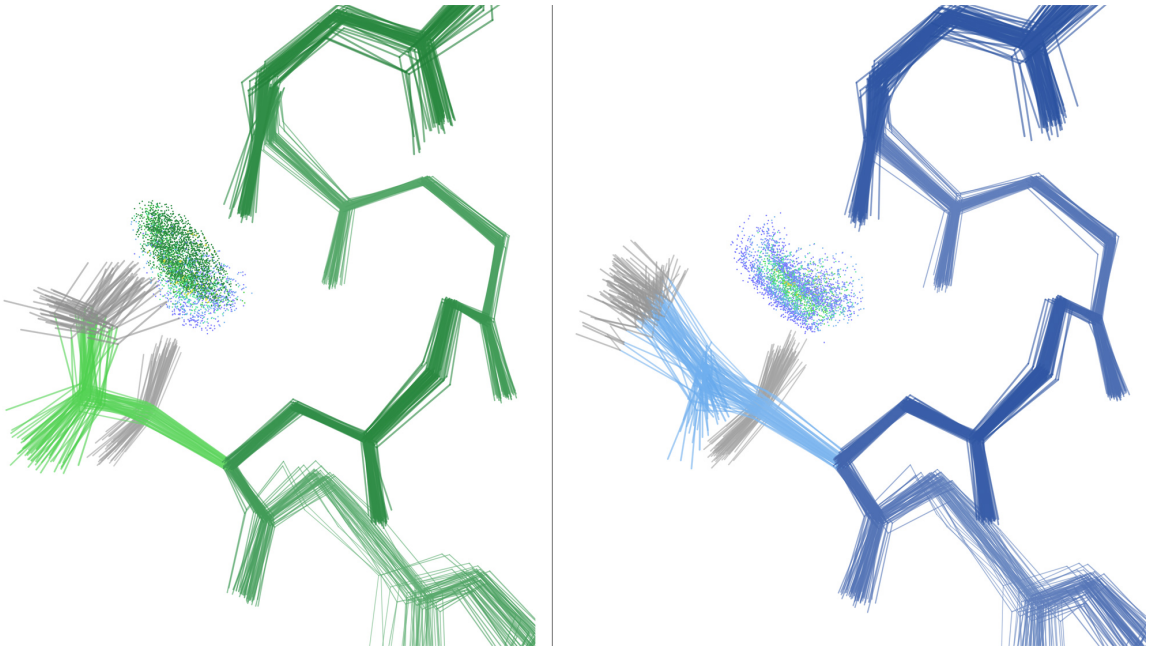

FIGURE 3.1: Mid-helix Asn *m-80* (left) and *m-20* (right) rotamers. The difference in $\chi 2$ results in contrasting modes of interaction with the preceding helical turn: an H-bond for *m-80* and van der Waals packing for *m-20*. However, these differences in non-covalent chemistry do not translate to differences in backbone coordinates: the backbones are indistinguishable. Examples were taken from the Top5200 data set (Chapter 4) and superimposed using the previously reported examples 1udc Asn99 and 1ab1 Asn12, respectively (Lovell et al., 1999), as reference structures.

posed examples extracted from otherwise unrelated crystal structures demonstrated that the backbones of the *m-80* and *m-20* subsets are almost identical (Figure 3.1): the Asn Cα to i-4 O distances are 3.84 $\pm$ 0.14 and 3.82 $\pm$ 0.16 Å, respectively, and the Asn Cβ to i-4 O distances are 3.51 $\pm$ 0.17 and 3.54 $\pm$ 0.19 Å, respectively.

I reached a similar conclusion examining Asn vs. Asp "pseudo-turns" (also known as "Asx turns") (Tainer et al., 1982; Rees et al., 1983). In this motif, an Asn/Asp sidechain Oδ1 at residue i in a pseudo-turn mimics the backbone O at residue i-1 in a true tight turn. However, the backbone does not seem to react appreciably to the slight difference in sidechain chemistry between Asn and Asp (Figure 3.2).

I also considered the possibility of sidechain-mainchain swaps as useful backbone

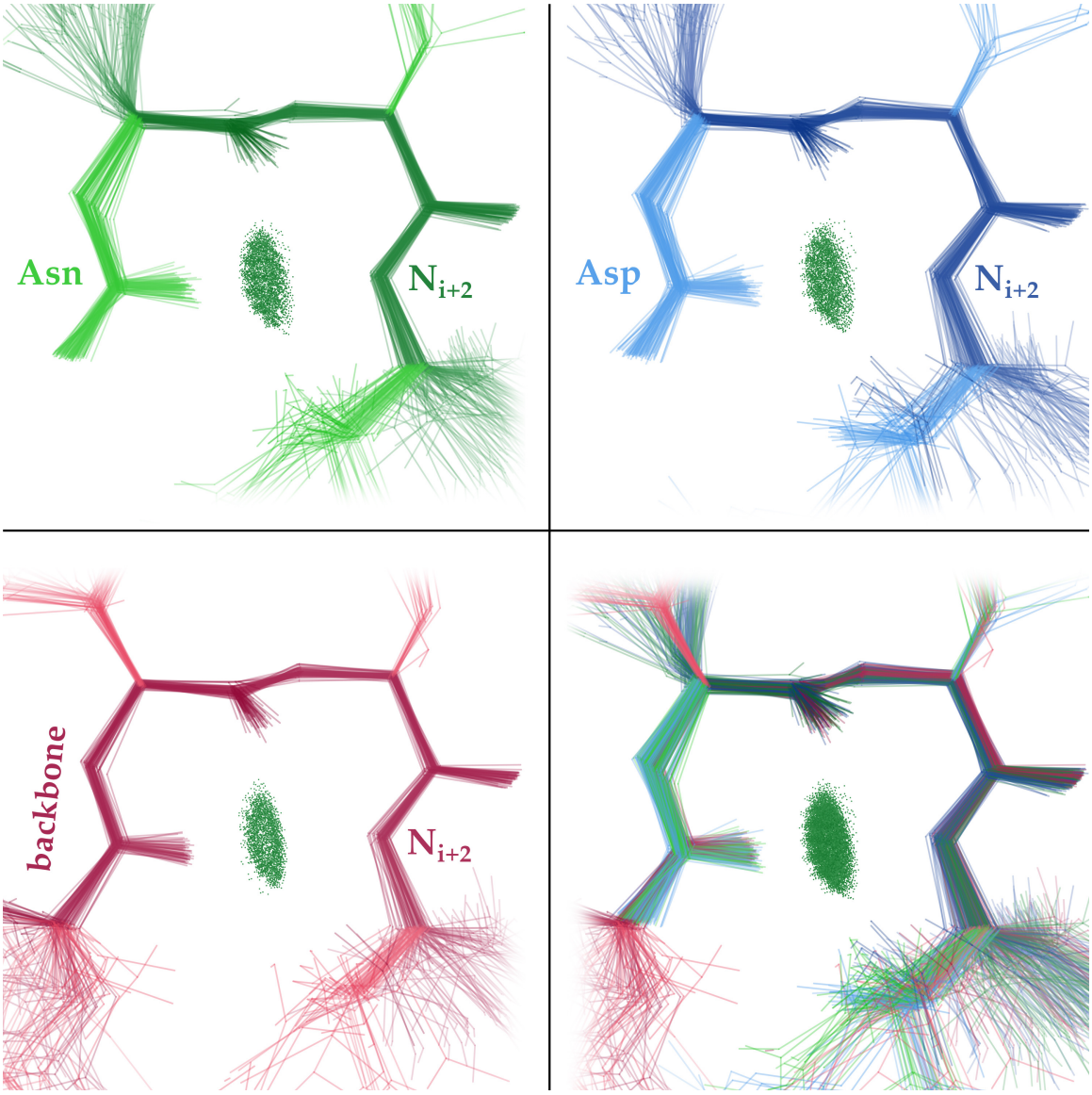

FIGURE 3.2: Asn or Asp pseudo-turns vs. tight turns. The Asn and Asp pseudo-turns are indistinguishable from each other (top left and right), and both match the true tight turns (bottom left) fairly well (bottom right). For each category, roughly 100 examples from the Top5200 were superimposed onto an arbitrary reference example with a $\leqslant$ 0.5 Å RMSD cutoff. (When a $\leqslant$ 1.0 Å RMSD cutoff was used instead, Asn and Asp pseudo-turns appeared still very similar to each other, but perhaps somewhat less similar to the tight turns.) The atoms for superposition were Cα i+2, i+1, and i and either Oδ1 i for pseudo-turns or Cα i-1 for tight turns. The examples shown are all type II (Venkatachalam, 1968), but results were along the same lines for types I, I', and II' in the sense that Asn and Asp pseudo-turns of a given type were very similar to each other and fairly similar to tight turns of that type.

moves. Indeed, mainchain mimicry by sidechains is observed in a number of contexts. The tight turn vs. pseudo-turn motif described above is one clear example. Another example is misfit regions of crystal structures where ambiguous electron density led the crystallographer to mistakenly model sidechain atoms where instead the backbone should continue, introducing geometric and steric errors that MolProbity flags as worrisome (Figure 3.3) (Arendall et al., 2005). (The sidechain-mainchain swap is a mistake in such cases, but it is nevertheless suggestive that certain atomic correspondences could be formulated as a useful *in silico* operator (see Section 7.4).) Furthermore, pairwise comparisons of some computationally predicted models reveal sidechain-mainchain swap relationships located precisely at critical folding nuclei where conformational changes must occur for *in silico* "folding" to proceed, as discussed in Section 7.4. Taken together, these observations suggest that a sidechain-mainchain swap operator may be useful as a discrete move with significant local impact. However, a loop closure protocol of some sort would be needed to address the resulting chain break, which may in general be rather large; this makes a swap operation significantly more complex than a backrub, which straightforwardly maintains chain connectivity. Therefore, sidechain-mainchain swaps remain a tantalizing and potentially powerful possibility for structural modeling, but more detailed investigation and careful benchmarking are required to establish their actual utility.

I additionally looked at aromatic residues across from Gly vs. any other residue type in parallel instead of anti-parallel (Chapter 2) β sheet, but because this motif is much more rare, I could not find a sufficient number of distinct examples for any meaningful analysis.

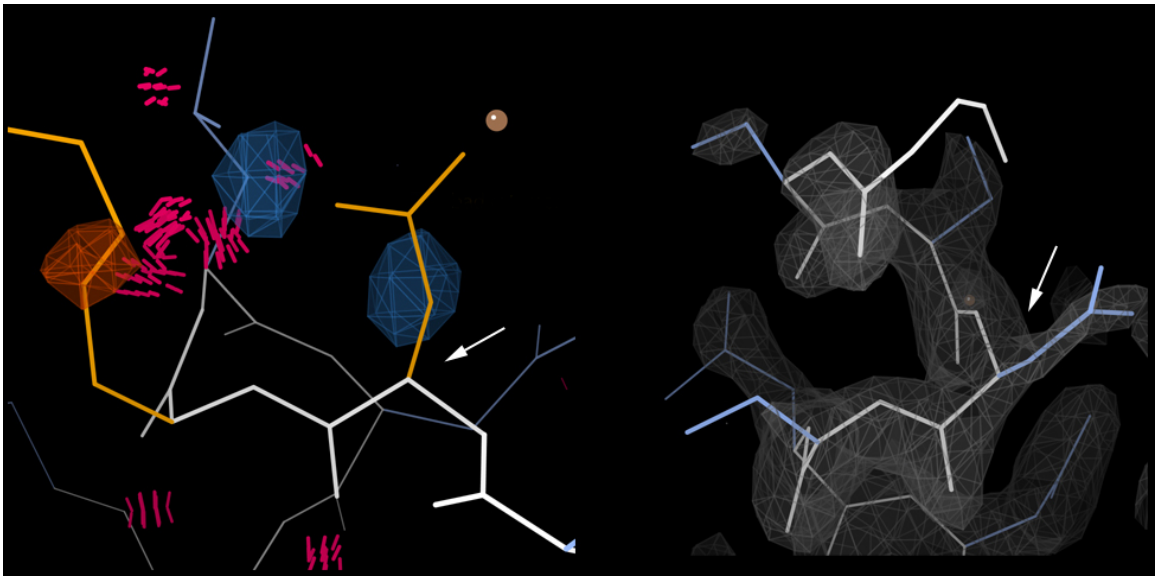

FIGURE 3.3: Erroneous sidechain-mainchain swap in crystal structure at N-terminus. Left: A sidechain is incorrectly assigned to mainchain density at the N-terminus of 1lpl. The error is flagged by multiple criteria including steric clashes (pink spikes), rotamer outliers (orange sidechains), and significant positive (blue mesh) and negative (orange mesh) Fo-Fc electron density peaks. Right: The error is corrected in the re-refined version, 1tov, by flipping out the sidechain and instead extending the helix N-ward for several more residues. The resulting model is a good fit to the 2Fo-Fc electron density (gray mesh). Credit: Jane Richardson and Ian Davis.

3.2 Shears: helical motions orthogonal to backrubs

In contrast to the partial failures described above, I did have success investigating an idea that our lab previously postulated (Davis et al., 2006) and that Tanja Kortemme's lab more recently discussed but did not study (Smith and Kortemme, 2008): helix combined winding/unwinding or "shear" (Figure 3.4). This movement affects three peptides to the backrub's two, and shifts the central peptide roughly parallel to its original orientation whereas the backrub effects a perpendicular motion. The 2D component of a shear move can be envisioned as a perturbation of a rectangle to form a parallelogram. A more concrete analogy might be a playground swing moving side-to-side instead of the usual back-and-forth. This swing metaphor emphasizes the geometric strain imparted to the "joints" (i.e. Cαs) for increasing shear magnitudes; thus shears are restricted to being subtle motions with local effects. (Note, however, that the actual shear occurs in 3D, with the joints offset in the z-direction relative to the x-y plane of the swing.)

To begin studying this new motion, I implemented shears in our Java code base much as Kortemme and colleagues imagined (Figure 3.4). I also built a tool in our graphics program KiNG (Chen et al., 2009b) to allow interactive shear modeling (Figure 3.5). Next, I searched for examples of shears in both crystal and NMR structures. Finally, I consulted with Mark Hallen of Bruce Donald's lab on incorporating shears into a new flavor of dead-end elimination called DEEPer.

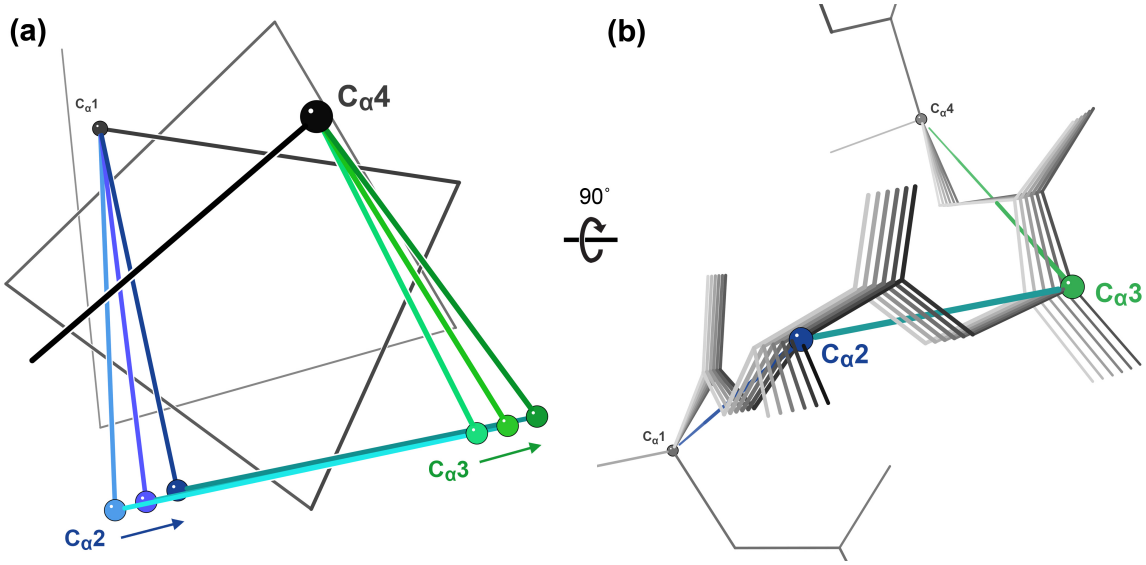

FIGURE 3.4: The shear backbone motion. (a) Simple Cα-only representation of shear, viewed down the axis of an ideal α helix (light colors). Shears of 5° (darker) and 10° (darkest) swing the central peptide Cα2-Cα3 peptide (cyan) sideways by coordinated rotations of the Cα1-Cα2 peptide (blue) and Cα3-Cα4 peptide (green). (b) All-atom representation of shear, viewed from the side of the ideal α helix (i.e. rotated 90° from (a)). Shears of 2° over a 10° range are shown. The central carbonyl is notably displaced parallel to the central peptide. One endpoint conformation is marked by balls and line segments colored as in (b). No individual peptide rotations are shown in this illustration.

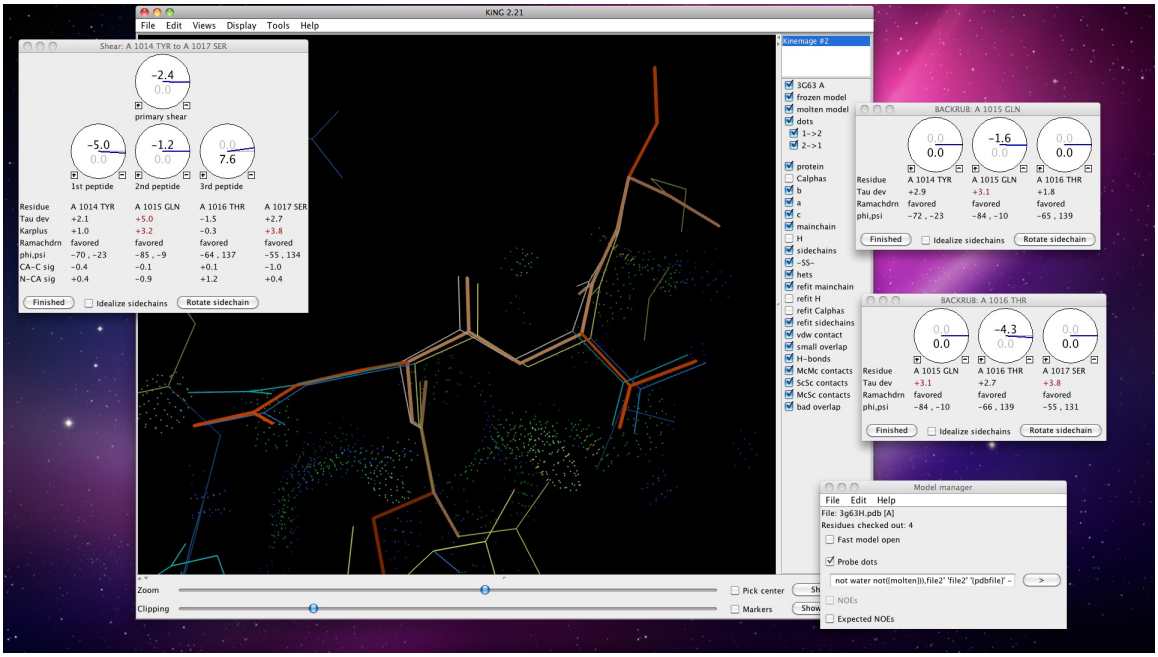

FIGURE 3.5: The shear tool in KiNG. Users can interactively remodel a region using the shear (top-left window), backrub (top-right and middle-right windows), and sidechain rotator (not shown) tools simultaneously, with all-atom contacts from Probe (Word et al., 1999b) illustrated in real time.

3.3 Characterizing shears in Cartesian and Ramachandran spaces

As mentioned above, Tanja Kortemme and colleagues proposed a reasonable parameterization of the shear motion (Smith and Kortemme, 2008), so I implemented a very similar model (Figure 3.4). First, the mainchain atoms from Cα i to Cα i+1 as well as the i+1 sidechain rotate about Cα i in the plane defined by Cα i, Cα i+1, and Cα i+2; this is the primary shear rotation. Then the i+2 sidechain and the mainchain atoms from Cα i+2 to Cα i+3 rotate about Cα i+3 in the plane defined by Cα i+1, Cα i+2, and Cα i+3. The angle of this rotation is calculated to keep the distance between Cα i+1 and Cα i+2 at its original value (near 3.8 Å); in practice, this is done not truly continuously or algebraically, but rather by sampling at 0.1° increments. Finally, the mainchain atoms between Cα i+1 and Cα i+2 are rotated about the axis defined by those Cαs to make the i+1 carbonyl C–O bond vector as close as possible to its unperturbed direction. Subsequent counter-rotations of the i to i+1 peptide and i+2 to i+3 peptide can optionally be used in a similar manner as with backrubs (Section 2.2).

I explored the ϕ,ψ behavior of shears, and found it to be quite different from that of backrubs (Figure 3.6). To be sure, both moves trace out swaths in Ramachandran space that are very context-sensitive, i.e. that differ greatly depending on whether the moves are initiated from α or β structure. However, shear ϕ,ψ traces appear to be more linear than backrub ϕ,ψ traces, suggesting shears are in some sense simpler. In addition, shears generally cannot be as large as backrubs (in terms of the magnitude of the primary rotation) without encountering distorted backbone geometry, although full counter-rotations of the flanking peptides appear to be possible more often with shears than with backrubs.

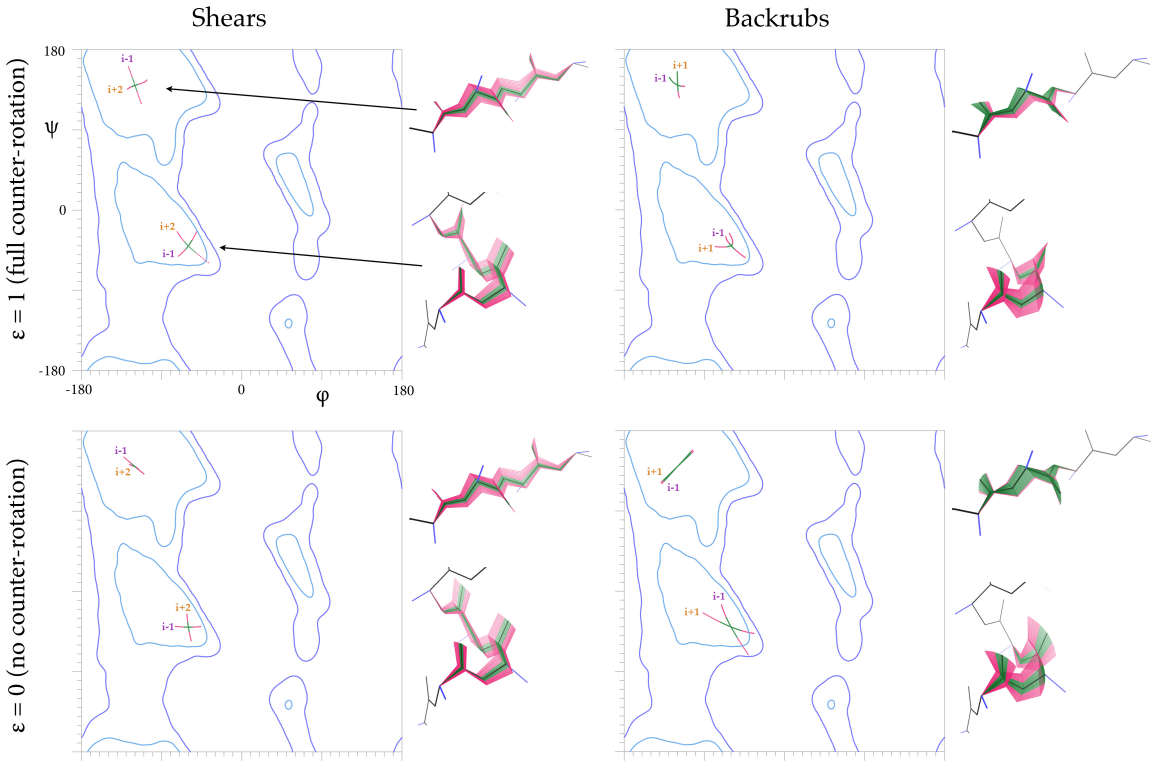

FIGURE 3.6: Shears vs. backrubs in Cartesian and Ramachandran spaces. Shears of ± 15° and backrubs of ± 30° were initiated from both ideal α helix (ϕ,ψ -60°,-40°) and ideal β strand (ϕ,ψ -120°,+140°) conformations with both full ($\epsilon = 1$) and no ($\epsilon = 0$) counter-rotations. Allowed conformations are in green; conformations disallowed by either Top8000 Ramachandran criteria (Chapter 5) or bad τ bond angles (> 5.5° from ideal) for any of the 4 (shears) or 3 (backrubs) constituent residues are in hot pink. The trace left by the N-terminal (i-1) residue is in purple; that left by the C-terminal (i+2 for shears, i+1 for backrubs) residue is in orange. The traces are complex, context-dependent, and different between shears and backrubs.

3.4 Shears in crystal structures

The original backrub analysis (Davis et al., 2006) involved extensive analysis of ultra-high-resolution crystal structures to provide experimentally derived evidence for the new motion. Similarly, here I present a broad survey of shears in ultra-high-resolution crystal structures.

3.4.1 Traversal between deposited alternate conformations

I first explored the ability of shears vs. backrubs to explain local protein backbone motions that were already modeled in crystal structures. To define an appropriate protein data set, I extracted the 54 protein-only (i.e. lacking DNA and RNA) structures in the PDB as of May 7, 2012 with $< 90\%$ sequence identity to each other, resolution $\leqslant 0.9$ Å, at least one chain with at least 34 residues, and electron density maps available from the EDS, and added hydrogens with Reduce (using the arcane `reduce -nobuild999` command to avoid Asn/Gln/His flips) (coordinates and maps available in supplement). The target set of local backbone regions included all "shear-like" regions, i.e. those with "anchor" Cα positions (Cα1 and Cα4) displaced by $\leqslant 0.01$ Å between declared alternate conformations. The pair of alternate conformations with the largest Cα2 and Cα3 displacements was chosen if more than the usual two (A and B) were available.

I then used a simple iterative algorithm to traverse from the first alternate conformation to the second. Each step consisted of the small ($\leqslant 1^\circ$) shear and/or backrub plus peptide rotations that most improved local backbone heavy atom RMSD over the window of 4 Cαs and 3 carbonyl oxygens, provided that ϕ and ψ remain in a favored or allowed Ramachandran region and τ remains within 5.5° of ideal. I declared convergence for a given region when the RMSD changes became very small (< 0.001 Å). This approach is deterministic in that it involves no stochasticity, but it does not

guarantee to find the combination or sequence of backbone changes that optimally interrelates the two alternates. Backrubs and shears are not precisely commutative – though the differences in final coordinates for different orders of operation are only a few cÅ (hundredths of an Ångstrom) – so an exact solution to this problem would be very computationally demanding (see also Section 3.6). My simple iterative approach sought to address this issue by limiting itself to a series of *small* backbone changes, thus approximating an equilibrium process, and ultimately achieved quite reasonable endpoints.

In the end, shears were more useful than backrubs in absolute terms for 22% of cases, and added value relative to using just backrubs for interrelating the two conformations for 89% of cases (Figure 3.7). Of course, shorter backrub-like two-peptide alternates are more common than the shear-like three-peptide alternates considered here – this is definitely true as modeled by crystallographers, and probably also true of structural dynamism in real molecules – so shears almost certainly account for less than 22% of alternates overall. Nevertheless, the results above suggest that the shear paradigm is in fact a reasonable explanation for a substantial fraction of local backbone changes.

3.4.2 Mining for shears in anisotropic electron density

To provide a lower bound on shears' prevalence to complement the upper bound described above, I conducted a manual search for shears using the data set of ultra-high-resolution crystal structures described above. The data set contained > 10,000 peptides which could conceivably be the central peptide in a three-peptide shear; I avoided examining all of them by narrowing my attention to candidates for which the electron density for the central carbonyl oxygen was anisotropic in a direction roughly parallel to the peptide plane. Many examples were deposited with a single conformation and were difficult to definitively categorize as shears, backrubs, or other

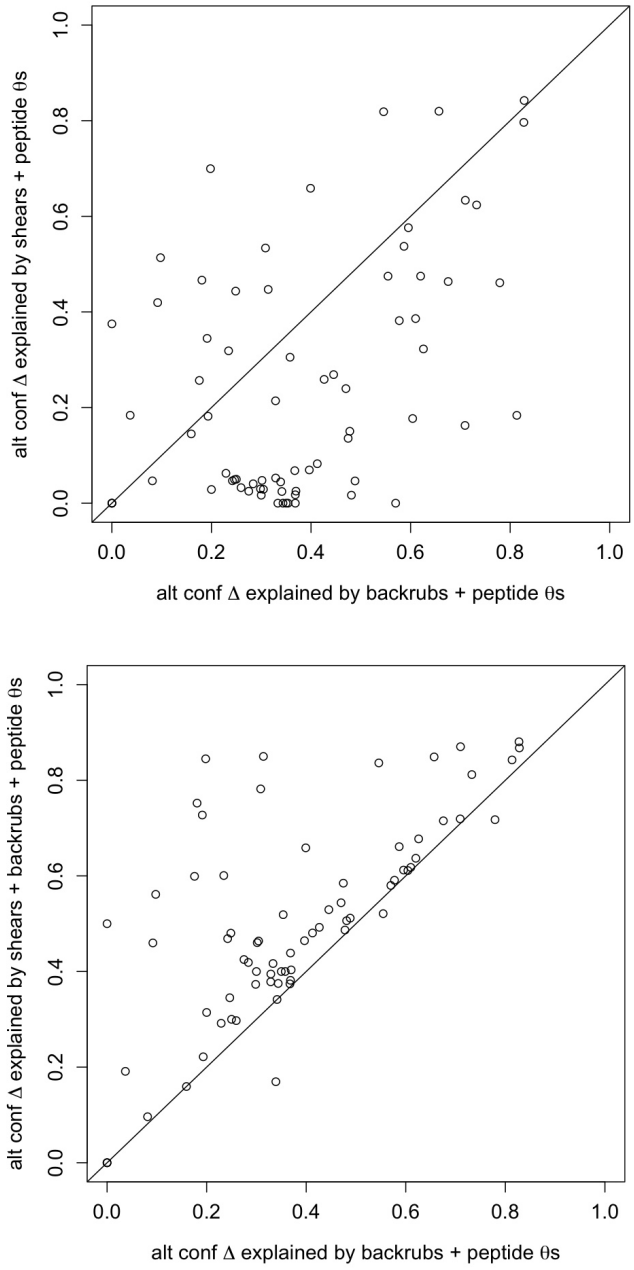

FIGURE 3.7: Shears vs. backrubs for interrelating 3-peptide alternate conformations in crystal structures. The x-axes describe the fraction of the original alternate conformation difference, expressed as RMSD over 4 Cαs and 3 carbonyl oxygens, that can be traversed by iterative backrubs plus rotations of the 3 interstitial peptides. The y-axes describe the same measure for shears plus peptide rotations (top) or for shears plus backrubs plus peptide rotations (bottom). Points above the diagonal lines indicate alternate conformation regions for which shears provided some value in absolute terms (top) or in addition to backrubs (bottom).

motions based on manual modeling with the KiNG shear, backrub, and sidechain rotator tools, but I was able to successfully identify 28 shear-like examples among the top 90 or so candidates (more persistent examination would have undoubtedly revealed more examples). This result translates to a lower bound of 0.3% of peptides undergoing shears, suggesting that shears are roughly an order of magnitude rarer than backrubs, which occur at about 3% of residues (Davis et al., 2006).

The confirmed shear examples were primarily localized to helical regions or helix-like loops. For example, 1muwA 63-66 is found in the first turn of a helix (Figure 3.8) and 2ykzA 116-119 is in the middle of a helix (Figure 3.9). About a third involve definitive rotamer jumps for at least one of the two central residues; this is likely a conservative estimate since sheared regions are often relatively solvent-exposed and therefore the central sidechains may have multiple low-occupancy conformations that are difficult to detect and explicitly model.

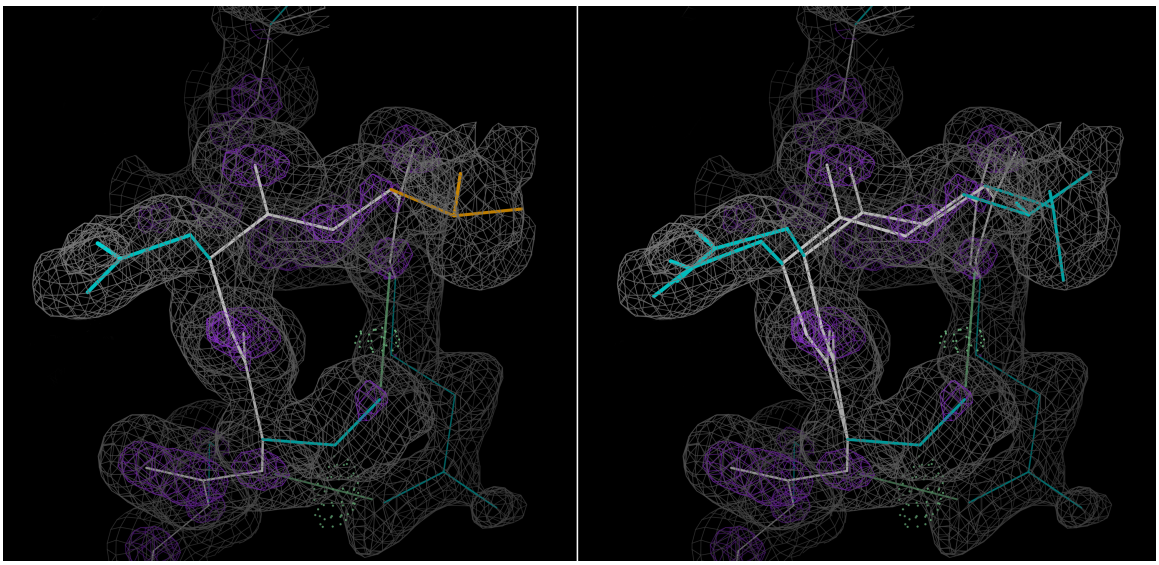

FIGURE 3.8: A shear in the first turn of a helix. Left: 1muwA (xylose isomerase) 63-66 is modeled with a single conformation, but the anisotropy (best seen for the central carbonyl oxygen) in the 2Fo-Fc electron density contoured at 0.7 σ (gray) and 3.0 σ (purple) and the rotamer outlier for Thr65 (orange) suggest a refitting is in order. Right: a manually refit model with a 7° shear (-4° and +3° from the original conformation), similarly small peptide rotations, and rotamer changes for Thr65 better explains the density and avoids the rotamer outlier. Note that Ser63 (bottom, foreground) is the N-cap and Glu66 (bottom, background) is the "capping box" (Harper and Rose, 1993) for this helix; corresponding sidechain-mainchain i to i+3 and i+3 to i H-bonds are shown.

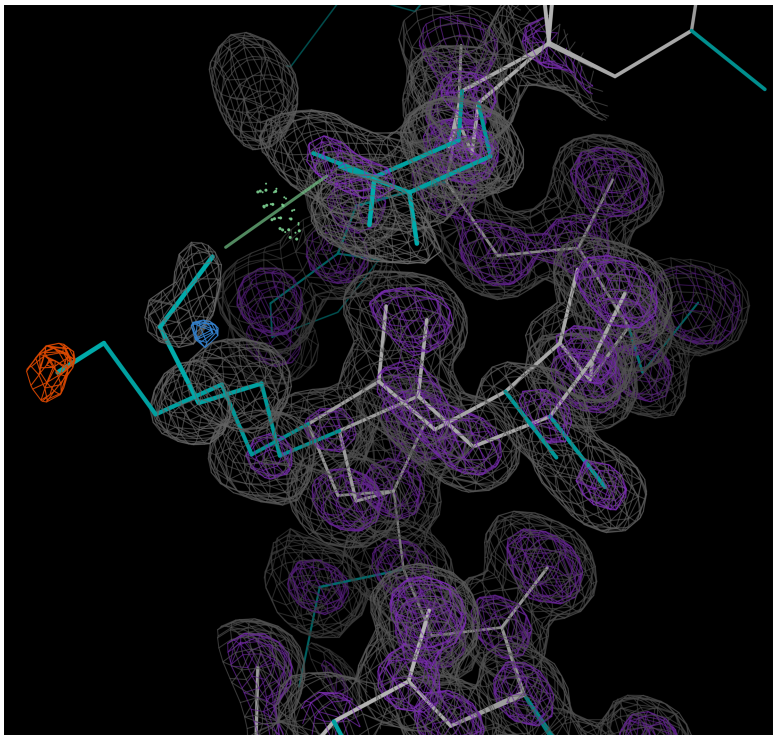

FIGURE 3.9: A shear in the middle of a helix. 2ykzA (cytochrome C prime) 116-119 is deposited with shear-like alternate backbone conformations which provide a good fit to the anisotropic 2Fo-Fc electron density contoured at 1.0 σ (gray) and 3.5 σ (purple); this is especially visible for the carbonyl oxygens. Lys117 is further modeled with alternate rotamers, one of which forms a sidechain-sidechain H-bond (light green "pillow" and line) to an Asp in the next turn and the other of which points out to solvent. The latter rotamer may have slightly too high occupancy judging by a -3.0 σ Fo-Fc peak (orange), and an additional rotamer may also be present based on a +3.0 σ Fo-Fc peak (blue).

3.4.3 Modeling shears into anisotropic electron density

With help from Jeehyun Lee, I went on to re-refine one such case (2jfr 135-138) in PHENIX (Adams et al., 2010). Refinements were performed both with and without a split into alternates using the KiNG shear tool. In each case, first occupancies were refined with *xyz* coordinates and B-factors fixed, then *xyz* coordinates and anisotropic B-factors were refined with occupancies fixed. As expected for such small coordinate changes, the final R-factors were nearly identical, with R_{work} 0.139 $\pm$ 0.001 and R_{free} 0.153 $\pm$ 0.001. Nevertheless, the fit to the density was improved visually (Figure 3.10).

Conclusion: Shears in crystal structures

There is reason to believe that the estimate given above for the prevalence of shears and the original estimate for the prevalence of backrubs (Davis et al., 2006) are conservative, i.e. that they underestimate how common these motions truly are. Namely, we now know that low-occupancy sidechain conformers lurk in electron density typically considered noise (Lang et al., 2010), especially with data collected at room temperature (Fraser et al., 2011). It is possible that these hidden sidechain conformers "drive" subtle backbone motions; indeed, the original backrub analysis (Davis et al., 2006) and the shear/rotamer-jump coupling statistics presented above implicate backbone motion quite generally when sidechains switch rotamers.

I recently identified an example of a heretofore hidden backbone motion in a multi-conformer, room-temperature crystal structure of CypA that James Fraser and Henry van den Bedem generated using qFit, a new algorithm for invoking alternate conformations only where they are locally necessary to aggregately explain the electron density (van den Bedem et al., 2009). Residues 142-145 were modeled with a single conformer in the original deposited structure (PDB code 3k0n), but interestingly qFit suggests a set of shear-like backbone alternate conformations

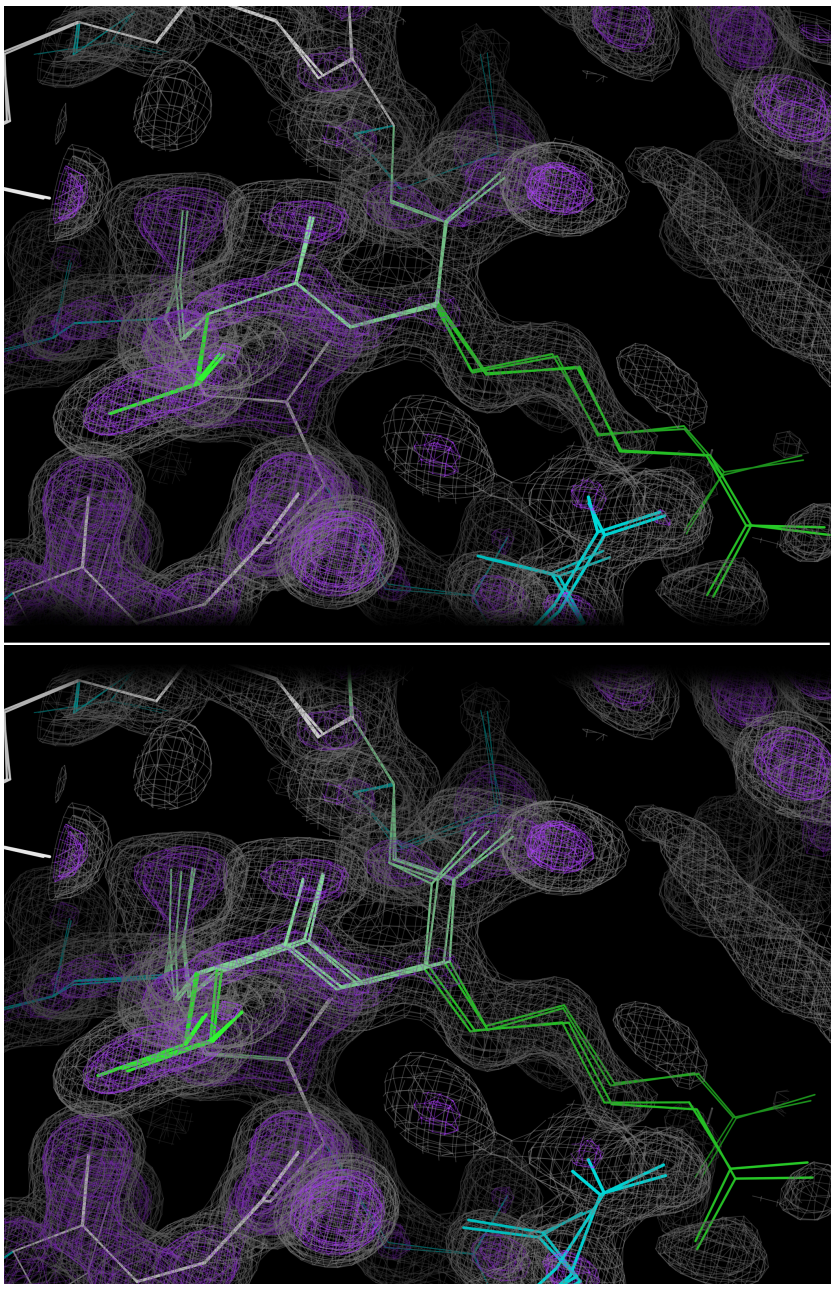

FIGURE 3.10: Remodeling and re-refinement of a candidate shear region. In both panels, each pair of extremely similar coordinates and electron density maps is from before vs. after re-refinement, demonstrating that refinement changes the model very little. Top: 2jfr 135-138 is deposited with a single conformation for the backbone (light green), but has alternates for the sidechain of Arg137 (dark green) to explain lower-contour electron density peaks (not shown). Bottom: This region appears to better fit the electron density after being manually remodeled with the KiNG shear tool to have alternate backbone conformations separated by a shear.

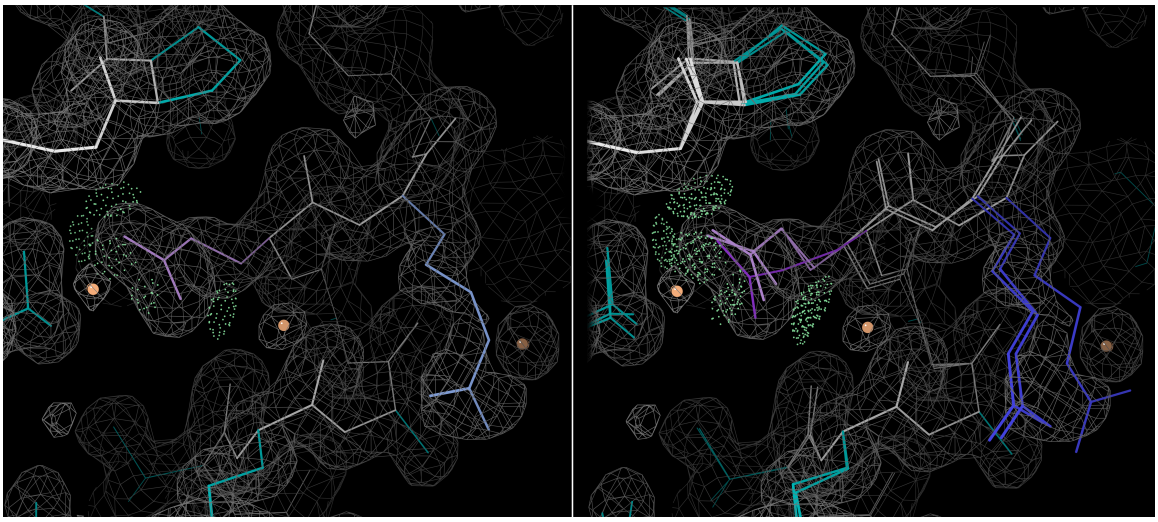

FIGURE 3.11: A shear in an room-temperature multi-conformer crystal structure. Left: Residues 142-145 in CypA are modeled with a single conformation in the "traditional" cryogenic structure (3k0n). The model is a reasonable fit to the 2Fo-Fc electron density contoured at 1.0 σ (gray), which is slightly anisotropic for the central carbonyl oxygen. Right: The room-temperature multi-conformer qFit model, on the other hand, includes 3 alternates with backbones related by a shear-like motion to explain the electron density. Each shear end-state is allocated about 50% occupancy. The multi-conformer model adds a second rotamer (dark purple) in addition to the original rotamer (light purple) for Glu133. It also selects a rotamer for Arg144 (dark blue) that is displaced in a swath-like fashion by the backbone shear, instead of the single original rotamer (light blue).

coupled to a rotamer change at Glu143 (Figure 3.11). Importantly, shears are not hard-coded into qFit, so the result of a shear "emerging" from better interpretation of the experimental data can be counted as validation of the shear concept.

3.5 Shears in NMR ensembles

I also looked for shears within three relatively recently published ubiquitin ensembles, each of which was created by imposing NMR dynamics measurements as restraints in molecular dynamics simulations.

The DER (dynamic ensemble refinement) (Lindorff-Larsen et al., 2005) ensemble was produced by imposing NOE and backbone ^{15}N relaxation S^2 restraints on 16 parallel molecular dynamics simulations, such that the replicas together explained the structural and dynamics data. This process was repeated for 8 cycles and the resulting models were pooled. The MUMO (minimal under-restraining minimal over-restraining) (Richter et al., 2007) ensemble was created similarly, but with 2 replicas for the NOE restraints and 8 replicas for the relaxation S^2 restraints; this arrangement was empirically found to better reproduce a simulated (and therefore exactly known) reference ensemble based on back-calculated NMR data. By contrast, the EROS (ensemble refinement with orientational restraints) (Lange et al., 2008) ensemble was generated by a less intuitive series of steps involving 2-replica simulations to satisfy NOE restraints (as with DER and MUMO) but with pooling of replicas that matched the backbone RDC S^2 restraints. The similarities and differences between the ensembles are summarized in Table 3.1.

Previously, Ian Davis searched for backrubs in the DER ensemble (the others hadn't been published yet) by superimposing each local 5-residue window onto the equivalent window from each other model in the ensemble. Backrub-like pairs were indicated by low RMSD for the "anchor" Cαs (1, 2, 4, 5) and large (i.e. farther from 0°) backrub Cα pseudo-dihedral (3-2-4-3', where 3 is Cα3 from one model and 3' is Cα3 from the other model).

To search for shears, I modified his method to instead superimpose each local 6-residue window. In this case, shear-like pairs had low RMSD for the new anchor

Table 3.1: Three ubiquitin ensembles reflecting structure and dynamics, created using different experimental restraints. *(Lindorff-Larsen et al., 2005) **(Richter et al., 2007) ***(Lange et al., 2008)

Ensemble	PDB code	Models	NOEs	^{15}N S^2	RDC S^2	Overfit?
DER*	1xqq	128	yes	yes	no	probably
MUMO**	2nr2	144	yes	yes	no	no
EROS***	2k39	116	yes	no	yes	probably

Table 3.2: Candidate shears common to all three ubiquitin ensembles. *Central two residues of six-residue window, flanked by two residues on each end. **Illustrated in Figure 3.12.

Central peptide*	# DER model pairs	# MUMO model pairs	# EROS model pairs
Ala28-Lys29	4	4	1
Lys29-Ile30	3	4	1
Ile30-Gln31**	3	6	2
Thr66-Leu77	2	1	1

Cαs (1, 2, 5, 6), small shear Cα pseudo-dihedral (maximum absolute value of either 3-2-4-3' or 4-3-5-4'), and large central Cα displacement (minimum of either 3-3' or 4-4'). Intuitively, regions for which some model pairs matched these criteria were shear-like because they were firmly anchored on both ends and had central Cαs that displaced significantly in the peptide plane.

I manually determined that values of $\leqslant$ 0.05 Å anchor RMSD, $\leqslant 4^\circ$ pseudo-dihedral, and $\geqslant$ 0.3 Å displacement produced similar numbers of interesting model pairs as did Ian's manually selected backrub parameter values. There was significant variability in the identities of the windows flagged as shear-like across the three ensembles, but there was unanimous agreement on a few regions (Table 3.2). The most prevalent region identified is in the heart of the only major helix in ubiquitin (Figure 3.12).

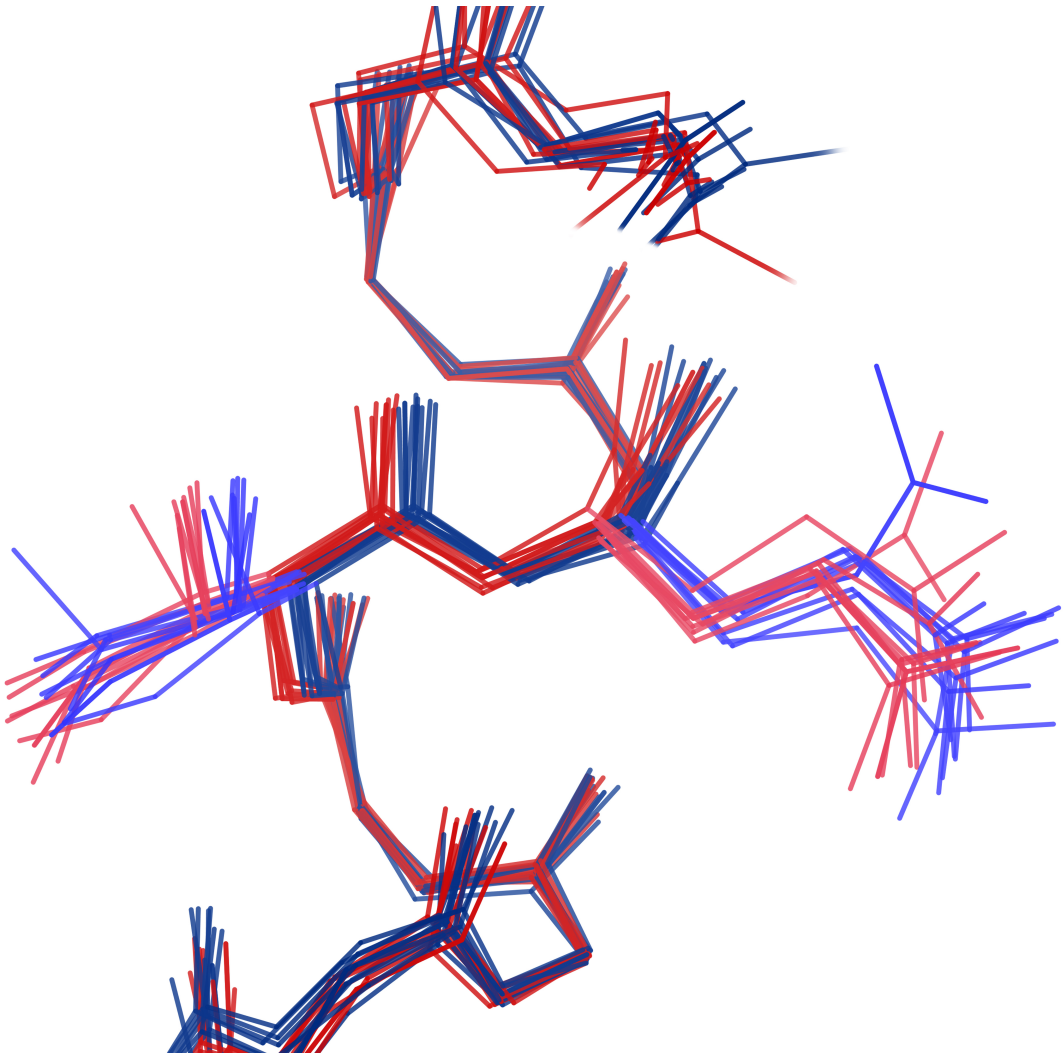

FIGURE 3.12: Ile30-Gln31, the most prominent shear common to all three ubiquitin ensembles. All models from the pairs implicated as undergoing a shear in this region (Table 3.2) are superimposed using the Cαs of residues 28, 29, 32, and 33 (i.e. the same Cαs used for superposition for the exploratory stage). Each model in each pair was manually assigned to a shear direction (red for minus, blue for plus) based on visual inspection. The backbone is clearly separated between the two categories for the central tripeptide but matches well for the flanking regions, indicating a shear-like relationship. The central sidechains, however, do not reveal any obvious coupling to the backbone.

By using more generous values of $\leqslant 0.10$ Å anchor RMSD, $\leqslant 5^\circ$ pseudo-dihedral, and $\geqslant 0.25$ Å displacement, I was able to plot ensemble-wide "shearness" – defined as the percent of all possible model pairs that meet the looser shear-like criteria – as a function of sequence (Figure 3.13). The most significant peaks are again in the long main helix, with a secondary peak for the short mini-helix.

Notably, the DER and MUMO ensembles have higher peaks than does the EROS ensemble; this is not due to their slightly larger sizes (128 and 144 models instead of 116) since "shearness" is a normalized quantity, namely percentage of model pairs suggestive of a shear. Thus one might imagine that shears are primarily fast-timescale motions, at least in ubiquitin. However, ^{15}N relaxation S^2 values are high for these regions, which one would naïvely interpret as meaning those regions are relatively rigid. Furthermore, RDC S^2 values are perhaps a bit low for the mini-helix, suggesting it could exhibit motion of some sort on a longer timescale. Based on these discrepancies, the fact that shear-like model pairs are more prevalent in the DER and MUMO ensembles may be due more to methodological differences between the DER/MUMO protocol and the EROS protocol (see above and Table 3.2) than to influences from faster-timescale experimental NMR data. Ultimately, a controlled comparison involving the same protocol but different experimental restraints would be necessary to fully disentangle the effects of timescale and methodology.

These results all corroborate my conclusion from observations of crystal structures (Section 3.4) that shears are indeed more common in helical regions.

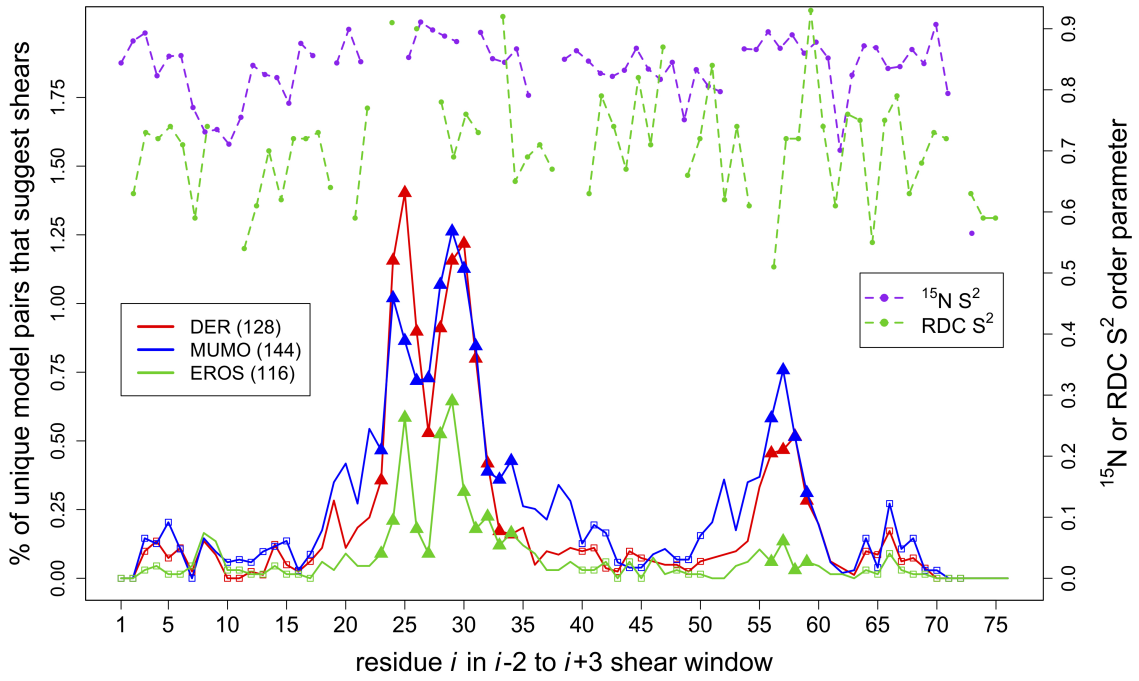

FIGURE 3.13: Shearness along sequence for three ubiquitin ensembles. The percent of all model pairs that have a shear-like local relationship (see main text) is plotted against sequence for three ubiquitin ensembles (red, blue, green) generated by different hybrid NMR/MD methods. For all three ensembles, the major peak of "shearness" is for the main helix (residues 23-35), although a minor peak also appears for the shorter mini-helix (residues 55-60) (filled triangles: α helix, open squares: β sheet). NMR order parameters based on shorter-timescale ^{15}N-relaxation data (Tjandra et al., 1995) (purple) and longer-timescale RDC data (Lakomek et al., 2006) (green) are relatively high in these regions.

3.6 DEEPer: protein design with shears, etc.

I have not yet fully investigated the propensity for sequence changes to induce shears, as I did with backrubs (Section 2.3). However, shears are similar to backrubs in many structural and geometric respects: they are small-scale, local backbone motions that often accompany rotamer jumps (in at least one third of cases – see Section 3.4). These observations suggest that shears could likewise be useful for accommodating mutations in design efforts.

To that end, I collaborated with Mark Hallen of the Donald lab on properly integrating shears alongside other backbone moves including backrubs and continuous sidechain minimization *a la* MinDEE (Georgiev et al., 2008b). The resulting algorithm, termed DEEPer (**D**ead **E**nd **E**limination with **Per**turbations), is the first provably accurate, deterministic protein design algorithm to incorporate both continuous backbone flexibility and continuous sidechain flexibility (Figure 3.14) (Hallen et al., 2012).

Assigning rotamer identities at different residue positions is commutative, so order of operations can be conveniently neglected in the usual DEE framework. The non-commutativity of shears and backrubs (Section 3.4), which I noticed early on in DEEPer's development, thus appeared to pose a problem: the conformation described by an s° shear and a b° backrub could in principle correspond to either the conformation created by an s° shear then a b° backrub or the conformation created by a b° backrub then an s° shear. Ivelin Georgiev and I skirted a similar issue for multiple backrubs in BRDEE by never allowing overlapping backrubs (Georgiev et al., 2008a). For DEEPer, by contrast, Mark Hallen implemented moves in separate "layers", such that final models could be created consistently by applying large moves first and small moves later.

Results of DEEPer calculations performed by Mark Hallen showed the promise

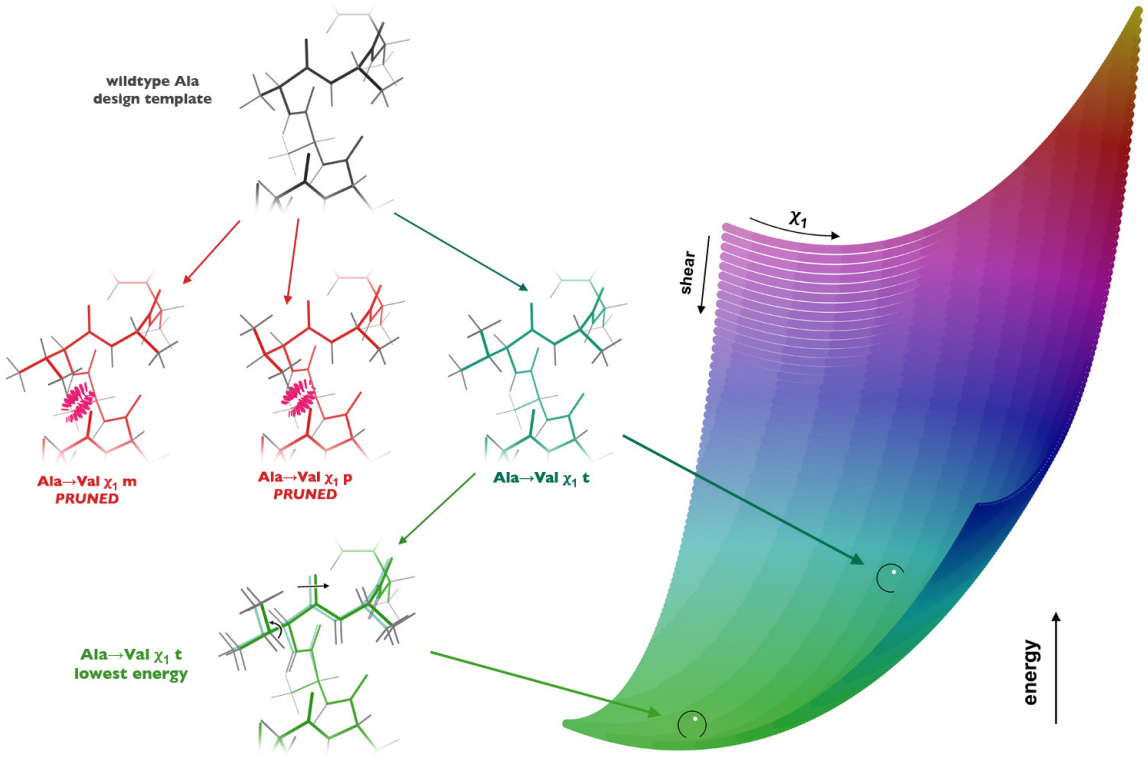

FIGURE 3.14: Cartoon of simultaneous backbone and sidechain flexibility in DEEPer. In an imagined scenario starting from an ideal helix (top left), an Ala→Val mutant can adopt one of three rotamers (middle left). DEEPer is able to prune two of them ($\chi 1$ m and p) due to clashes (red), but the third ($\chi 1$ t) (teal) is acceptable. Energy minimization simultaneously over shear and $\chi 1$ degrees of freedom results in the final lowest-energy conformation (green) (bottom left). These conformations can be mapped onto a simple theoretical energy surface (right) corresponding to $E = 2(s-4)^2 + 0.5(c-180)^2$, where s is the shear angle and c is $\chi 1$. Because the starting values for s and c are 0 and 175, respectively, the lowest-energy conformation corresponds to a 4° shear and a 5° $\chi 1$ rotation. The surface represents the multidimensional space that is provably completely searched by DEEPer for a given torsional well (in this case $\chi 1$ t). Made with help from Mark Hallen and Bruce Donald.

of combining two orthogonal move types – shears and backrubs – alongside continuous sidechain dihedral flexibility to identify realistic low-energy sequences and conformational ensembles.

For example, DEEPer was used to identify low-energy sequences for a region of 2bgx. Shears and backrubs of 0 or $\pm$ 5° were allowed for residues 126-131, and sidechain flexibility was allowed at adjacent residues, for a total of 19 flexible residues. Unlike most DEEPer runs, this particular search was discrete instead of continuous in order to better visualize the backbone conformational space being searched (Figure 3.15). The resulting lowest-energy conformation had three sequence changes and significant backbone displacement relative to wildtype.

DEEPer was also used to generate biophysically reasonable ensembles of low-energy conformations given fixed sequences, in the spirit of K* (Lilien et al., 2005). For example, residues 157-160 of 2ixt have alternate conformations related approximately by a shear motion in the crystal structure, likely indicating increased backbone dynamics compared to other parts of the structure. Correspondingly, the DEEPer ensemble generated using the native sequence showed more diversity of backbone conformations than in tests on other systems, sampling the conformational space around and between the alternates (Figure 3.16).

Low-energy states such as those described above are less likely to be discovered by algorithms allowing less flexibility, or by algorithms allowing DEEPer's flexibility but lacking guarantees of solution optimality. Notably, across 67 sequence-design runs on 64 protein systems, the lowest-energy conformations calculated by DEEPer were lower in energy than those calculated by a faster version of MinDEE by an average of 1.9 kcal/mol (ranging from 0 to 14.1 kcal/mol). Furthermore, the backbone motions in DEEPer induced sidechain motions: one or more rotamer changes were observed in 46% of tests, and up to four rotamer changes per test were observed. These results highlight the value of combined continuous backbone and sidechain flexibility, as

opposed to just continuous sidechain flexibility with fixed backbone.

DEEPer is implemented as part of the Donald lab's OSPREY (**O**pen **S**ource **P**rotein **Re**design for **Y**ou) software package (Gainza et al., 2012).

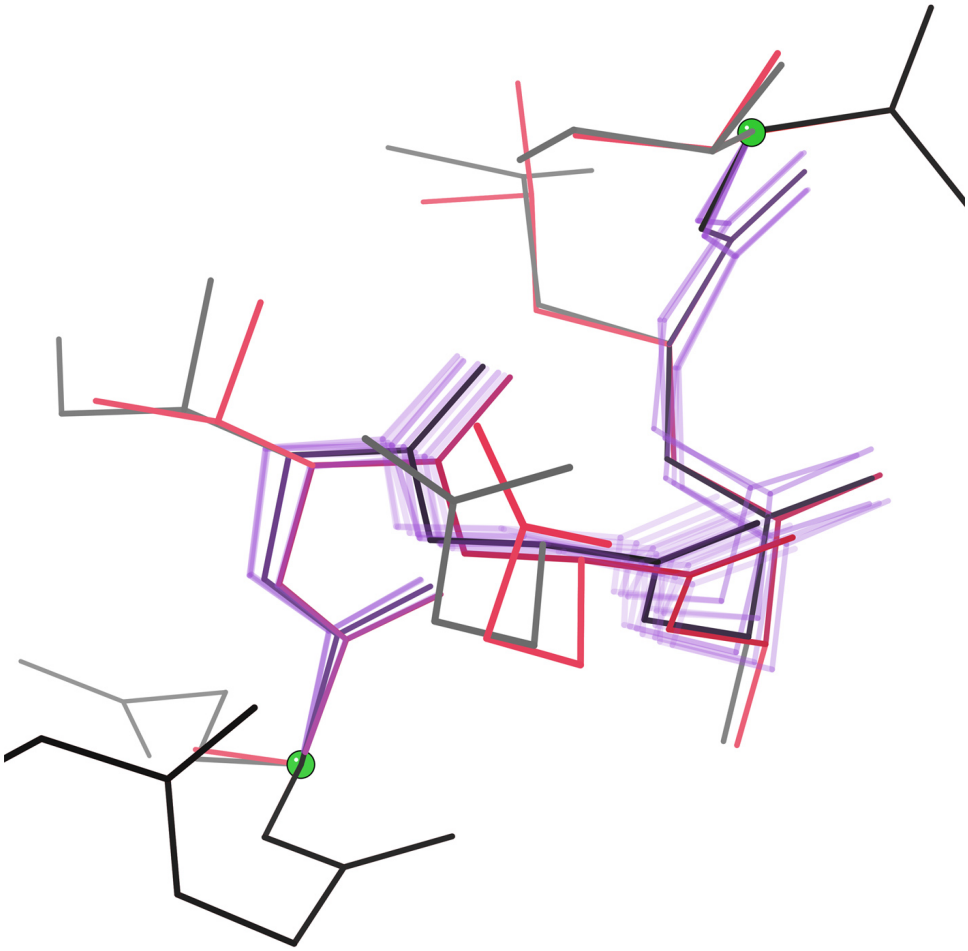

FIGURE 3.15: Sequence-design run on 2bgx (*E. coli* AmiD). The backbone of the lowest-energy conformation moved away from the starting conformation for residues 126-131. The lack of continuous flexibility in this run allows display of all searched backbone conformations. Starting structure, black/gray; complete searched ensemble, purple; GMEC, pink. Here and in Figure 3.16, green balls demarcate flexible-backbone regions; sidechains outside these regions are omitted for visual clarity. Made for (Hallen et al., 2012).

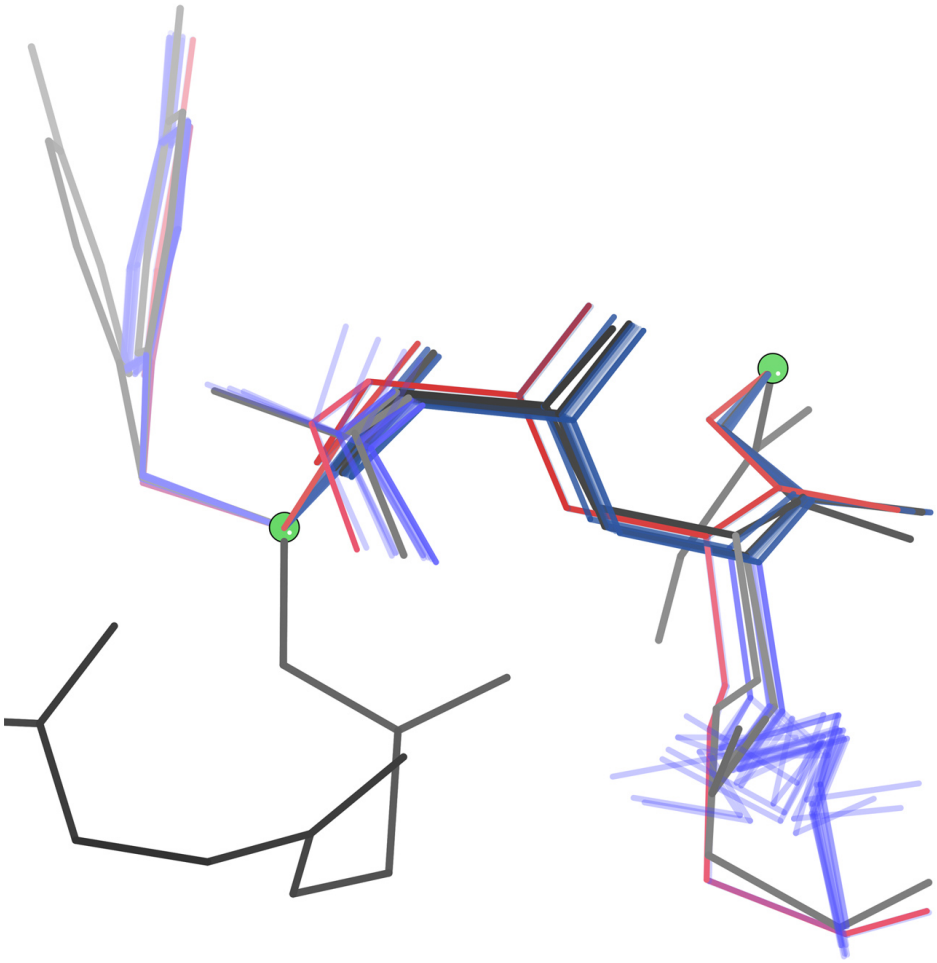

FIGURE 3.16: Ensemble-generation run on 2ixt (*L. sphaericus* sphericase). The low-energy ensemble of computed models was fairly wide at residues 157-160 and spanned the crystallographic alternates. The GMEC was on the fringe of the ensemble. Starting structure, black/gray; low-energy ensemble, blue; GMEC, pink. Made for (Hallen et al., 2012).

3.7 Discussion

I have described the shear, a primarily helical local backbone motion, and compiled ample and diverse evidence for its existence in experimental crystal and NMR structures. In collaborative work, I have also demonstrated its ability to help discover lower-energy sequences and conformations in protein design.

Both shears and backrubs are geometric simplifications of the actual underlying mechanisms, with the virtue of being more suitable for *in silico* manipulations. However, compared to backrubs, shears describe a more complex molecular motion that affects a larger local region (three peptides instead of just two). Shears are correspondingly less common in deposited structures because it is currently more difficult for crystallographers to model alternate conformations for larger regions. They are likely also rarer in real proteins because, if one assumes each residue may move roughly independently, larger regions are statistically less likely to move in a correlated fashion (clearly this assumption is not entirely valid because secondary structural elements, including helices, are inherently cooperative, but it nonetheless holds some value).

Nevertheless, I have established a conservative lower bound on the prevalence of shears, suggesting they are not all that much rarer than backrubs. Because shears swing backbone parallelly and backrubs push it perpendicularly, these two motions together form a convenient "basis set" for local backbone movement.

Future work will be needed to ascertain whether shears demonstrably accommodate mutations in natural protein evolution and to identify further heretofore undiscovered modes of local backbone motion.

4

Frustrations and Improvements at High Resolution

4.1 Crystallography at high resolution isn't always easy

As shown in the preceding chapters, proteins retain a significant measure of dynamism even in crystals. The evidence lies in high-resolution crystal structures, which allow one to discern precise atom positions and thereby surmise the existence of multiple conformations in many cases.

Along with this extra information, however, comes the burden of coalescing it into self-consistent, physically realistic multi-conformer models. For example, the crystallographic experiment provides no direct evidence as to the correlations and anti-correlations between observed alternate conformations, so external information in the form of steric contacts must be introduced. Furthermore, atomic occupancies must be set carefully to avoid implying illogical atomic overlaps (Figure 4.1). The problem is exacerbated in larger structures, where larger, more difficult-to-disentangle networks of coupled alternate conformations become possible (Figure 4.2).

Unfortunately, most existing refinement packages and related tools are insufficient to optimize networks of adjacent alternate conformations, so crystallographers are

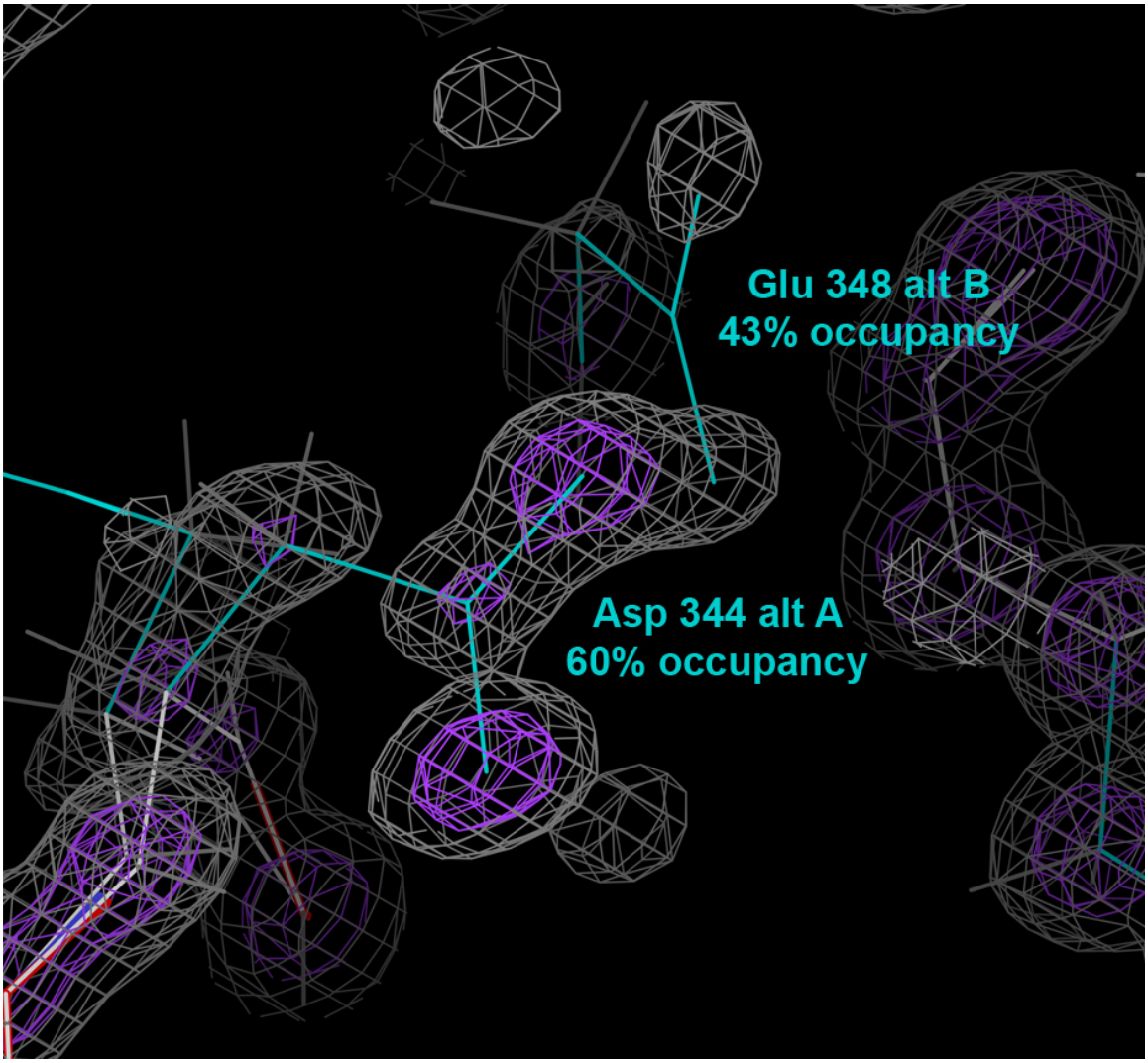

FIGURE 4.1: Asp344 and Glu348 of xylose isomerase (1muw) are clearly visible in the 0.86 Å resolution electron density map. They are assigned A and B alternate conformation labels, respectively, implying they are mutually exclusive for this region of space. However, their occupancies sum to $> 100\%$, implying they must co-exist at least 3% (or up to 43%!) of the time; such a situation would have unfathomably high energy and therefore would essentially never occur in reality. These occupancies are thus incorrect and must be adjusted, e.g. to 60%/40% or 57%/43%, to eliminate the logical fallacy in the model.

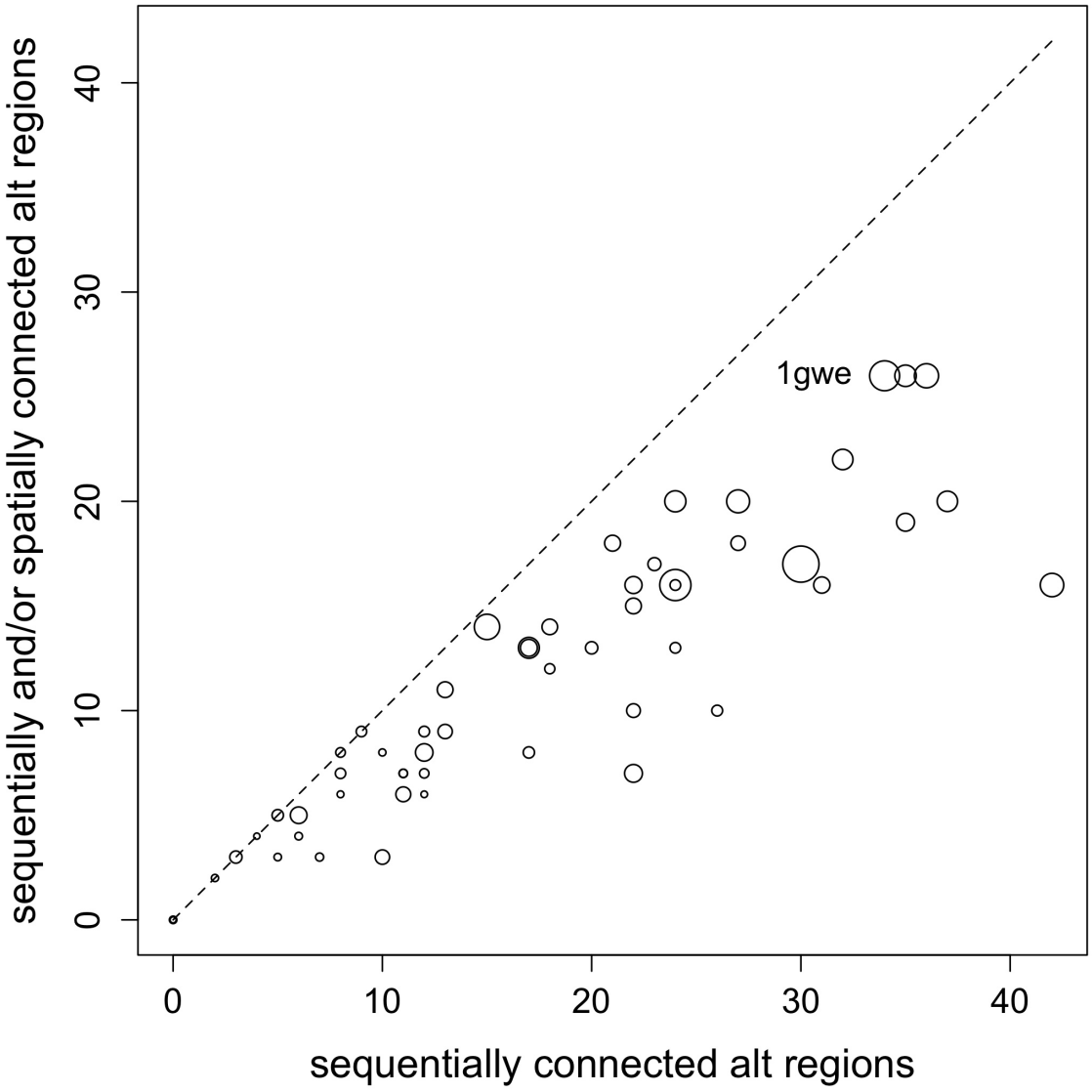

FIGURE 4.2: Large structures enable large coupled alternate networks. A simple program was used to group alternate conformations into connected regions/networks. The x-axis shows the number of such regions if residues are grouped together only if they are adjacent in sequence. The y-axis shows the number of such regions if residues are grouped together if they are adjacent in sequence *or* in three-dimensional space (i.e. some pair of atoms from the different residues are within 3 Å of each other). The data set consists of the "A" chains from the same 54 ultra-high-resolution structures used in Section 3.4. Points are scaled based on the total number of residues in the chain. Intriguingly, alternate networks are significantly *spatially* intertwined (points well below the diagonal line) only in relatively large structures such as 1gwe (labeled; see Section 4.3).

typically left to manually define those relationships as best they can. The result is that very high-resolution structures break the trend of improving model quality with improving resolution. As the number of high-resolution structures continues to rise, this deficiency will become increasingly troublesome.

Clearly new approaches are needed for dealing with the conformational multiplicity associated with high-resolution structures. To that end, this chapter describes a collaborative initiative I participated in to compile a set of "paragon" structures, free of any definable defects. My efforts to bring nearly perfect structures to full paragon status helped define a roadmap for future efforts at designing tools for automated high-resolution structure modeling and refinement.

4.2 The quixotic quest for "paragon" structures

The paragon project is an initiative undertaken by our lab and collaborators nearby and at Virginia Tech to collate crystallographic models that have attained perfection. Such sought-after models are completely devoid of any demonstrably incorrect rotamer, Ramachandran, and bond length/angle outliers; steric clashes; and Cβ deviations, as judged by MolProbity. They also have self-consistent alternate conformation labels and logically compatible occupancy values for neighboring atoms. The hope is that paragon structures will serve as gold standards for various downstream modeling tasks, from bioinformatics studies to drug design to MD simulations.

Out of the $>$ 70,000 crystal structures in the PDB, only two pre-existing paragons could be identified: 2zqe and 3iuf. This remarkably low total highlights the striking difficulty of simultaneously satisfying all of MolProbity's cadre of stringent requirements.

I also identified a "cryptic" paragon, 3kyv, by manually confirming that the single outlier in the structure, for the Lys2 rotamer, was actually a rare occurrence ($<$ 1%) of a genuine rotamer outlier (Figure 4.3). This case exemplifies the fact

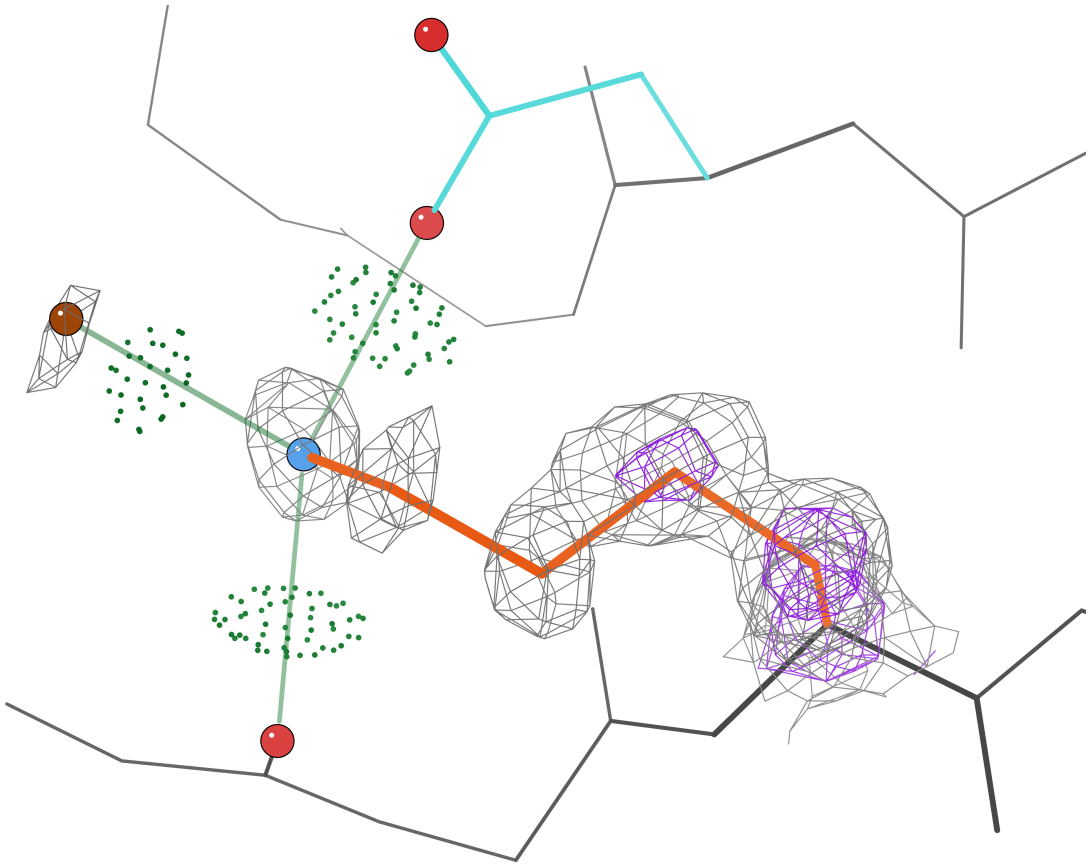

FIGURE 4.3: The only putative error flagged by MolProbity in 3kyv, a 1.10 Å neutron diffraction crystal structure of perdeuterated rubredoxin, is a rotamer outlier for Lys2 (orange). However, several lines of evidence suggest it is a valid conformer: the electron density contoured at 1.2 σ (gray mesh) and 3.0 σ (purple mesh), and hydrogen bonds (green "pillows" from Probe and translucent lines) to the backbone CO of a preceding residue, the sidechain of Asp13 (cyan), and an ordered water (brown ball, B-factor < 43). 3kyv therefore qualifies as an error-free "paragon" structure.

that outliers in our torsional distributions may occasionally be valid conformations, if compensated by local stabilizing interactions.

Another class or paragons was created by fixing the dusting of errors in a small set of nearly error-free, mostly high-resolution "near-paragon" structures. For example, Jane Richardson cleaned up the only error in 1akg with a single Pro ring flip, and Lindsay Deis repaired 1ubq with 6 manual rotamer changes and 5 Probe-

recommended amide flips.

4.3 Approaching paragon quality for a large structure

To complement these relatively small paragons, I set out to repair 1gwe (Murshudov et al., 2002), a 0.88 Å structure of catalase. This protein has 503 residues, > 800 waters, 3 sulfates, and 1 heme per monomer; moreover, it functions as a tetramer.

The structure has above-average quality for its resolution, with a MolProbity score in the 70th percentile. However, due to its large size, that still translates into a significant number of errors to address. The problem is exacerbated by the non-linear growth of potential conflicts between alternate conformations as a function of protein size (Figure 4.2).

Altogether, I ended up making over 40 changes to 1gwe. The resulting model still does not achieve strictly defined paragon status (Figure 4.4), but the process shed light on the features that next-generation high-resolution refinement tools must include.

First of all, there are no Ramachandran outliers, but all 13 residues with merely allowed (instead of favored) ϕ,ψ according to MolProbity were well supported by the electron density, and all but Pro60 were backed up by strong H-bonds (more than one in all cases but one). Interestingly, although most of these residues remain merely allowed using the new Top8000 Ramachandran distributions, Pro60 becomes favored using the new separate *cis* Pro distribution (see Chapter 5). Likewise, there are three rotamer outliers, but they are well validated by electron density, van der Waals packing, and H-bonds, and are therefore genuine outliers rather than errors.

These torsional oddities aside, many of the errors in the original deposited structure, and especially the most egregious ones, involved alternate conformations. Fortunately, some alternate-related fixes were quite simple to perform once the problem was identified. For instance, some waters were simply assigned the wrong alternate

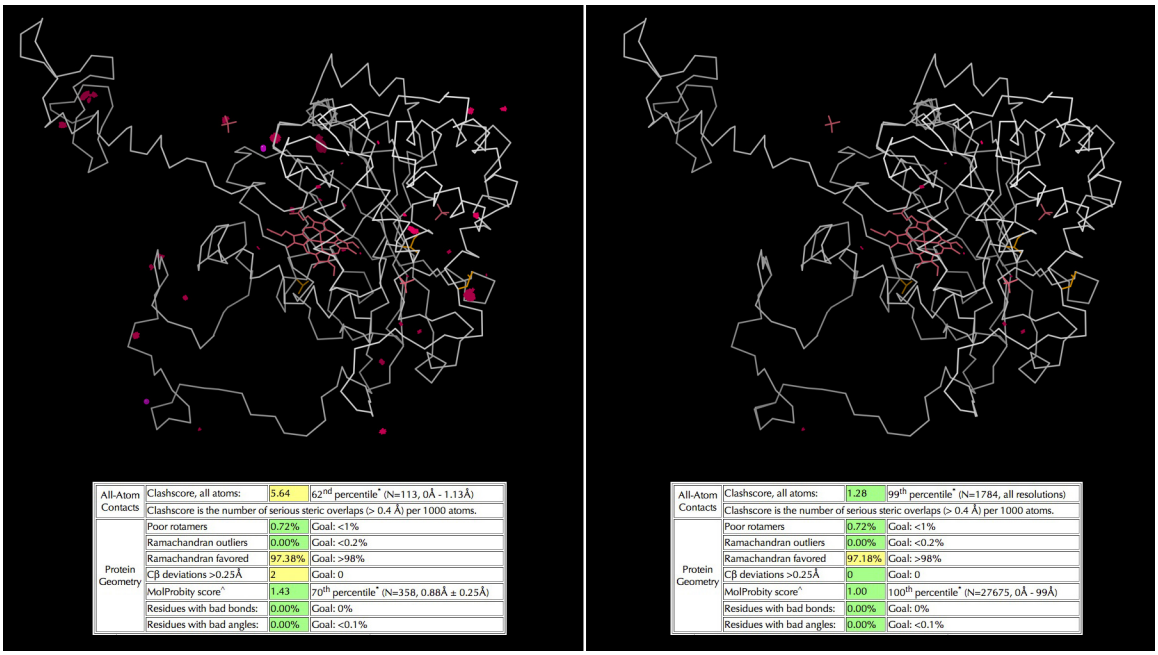

FIGURE 4.4: Near-complete paragonization of catalase. The deposited structure 1gwe (left) is above average for its resolution (0.88 Å), but nonetheless has several errors. The improved model (right) corrects many of these errors but still has a few probably unavoidable clashes.

label given their surroundings (Figure 4.5); in these cases a simple reassignment, e.g. from alternate A to alternate B, was sufficient to resolve the inconsistency. A similarly simple-minded example is Leu105, for which the sidechain and backbone alternate labels were accidentally swapped by the crystallographer(s); again, a simple reassignment solved the problem (Figure 4.6).

Apart from some such "low-hanging fruit", most errors required actual coordinate changes to repair. For example, many local dipeptides were modeled with a single backbone in spite of having alternate sidechains; as discussed previously (Davis et al., 2006) (see also Section 2.2), backrubs may better model the backbone in many such cases. Problem areas of this sort were often easily flagged by identifying concentrations of bond length and angle strain and ends of dipeptides indicative of misfit backbone geometry.

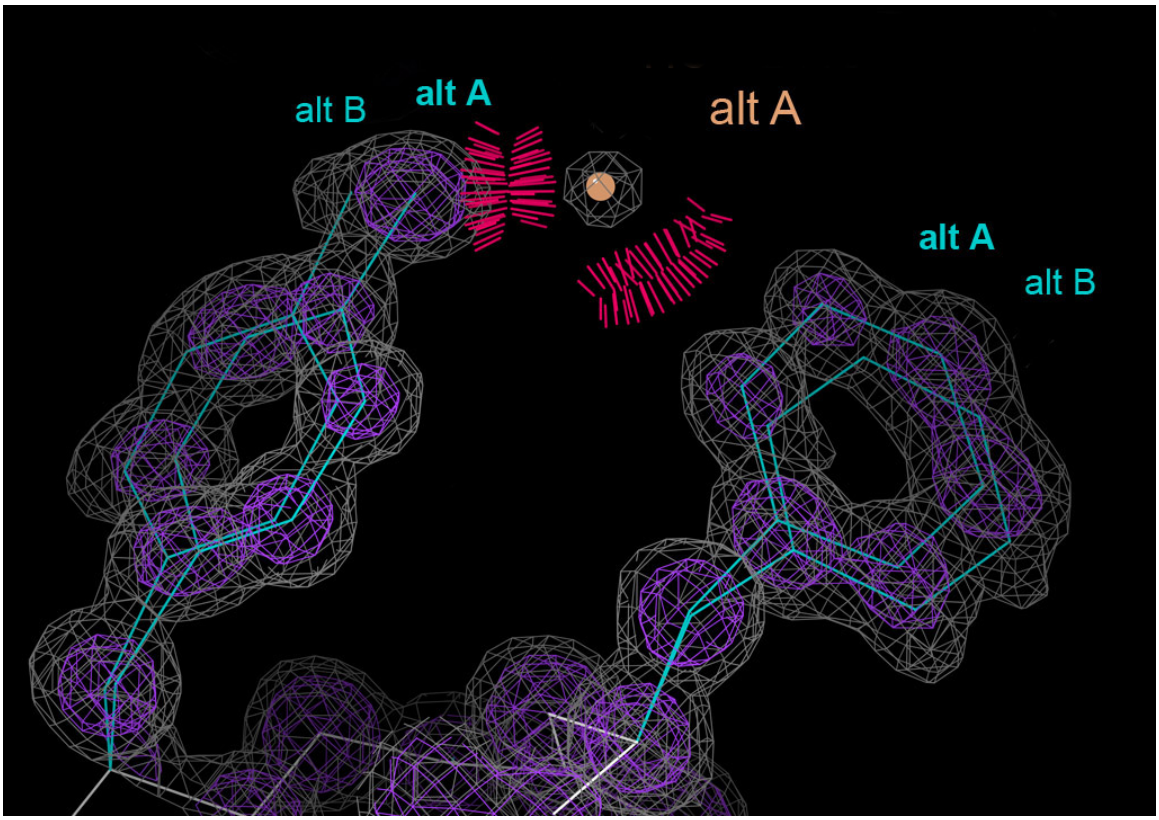

FIGURE 4.5: Incorrect alternate label for water in catalase. In the deposited structure 1gwe, HOH 2463 (peach ball) is assigned to alternate A, but that implies a clash with alternate A of both Tyr277 (left) and Phe279 (right). Those sidechains appear to have the correct alternate assignments – e.g. Tyr277's hydroxyl oxygen has stronger density for alternate A than for alternate B – so the water must be reassigned to alternate B.

Other missing alternates were more dramatic: they involved entirely unmodeled rotamers that were nonetheless visible in 2Fo-Fc and/or Fo-Fc electron density. For example, Arg464 is poorly fit and has clashes to poorly fit waters; a second alternate in place of those waters, coupled to an automated 180° Asn173 flip, not only alleviates the problems but also introduces several well-formed H-bonds (Figure 4.7). This type of convergence – with errors disappearing and favorable interactions surfacing – strongly suggests that my changes are genuine corrections, resulting in a model that more closely represents the real protein.

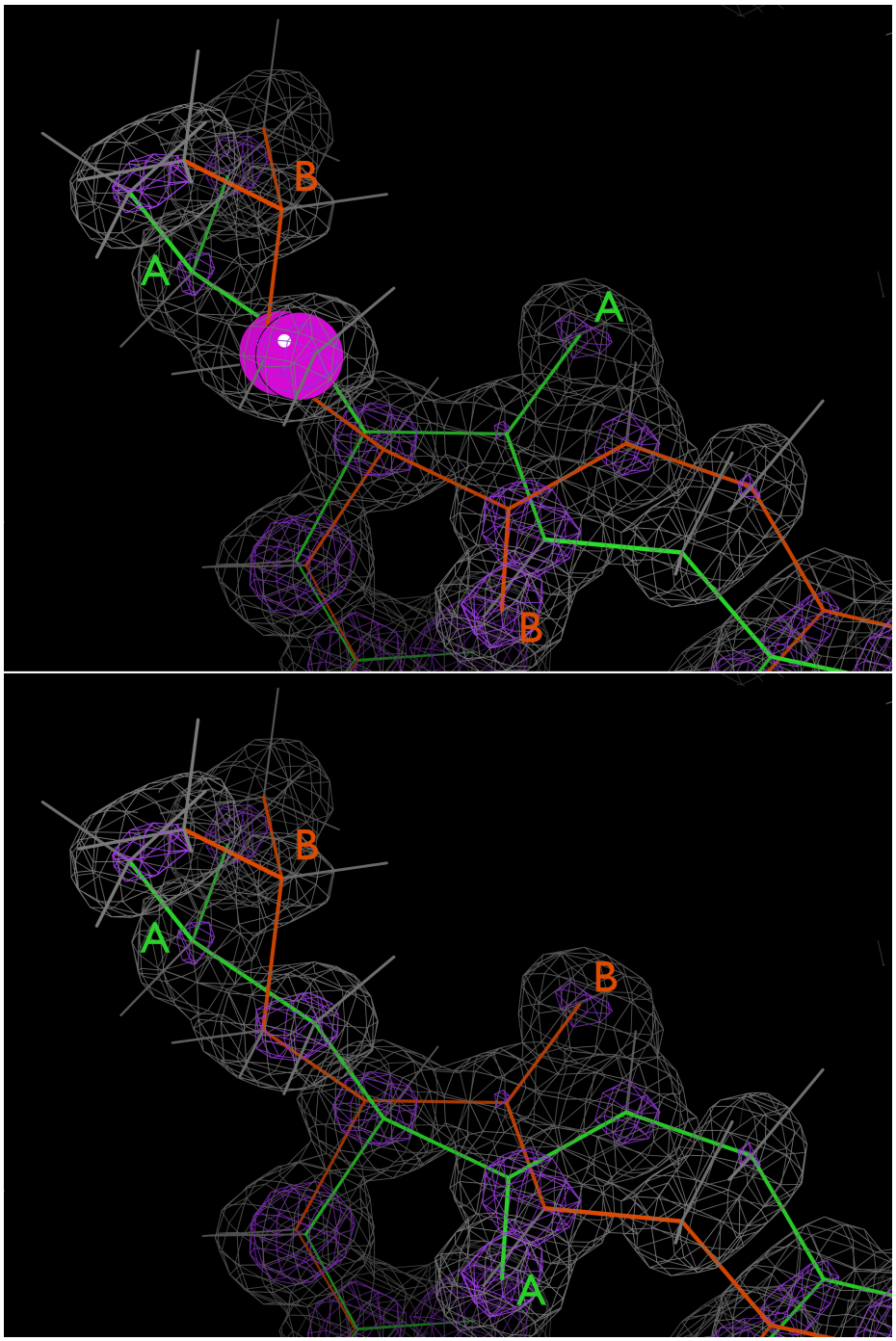

FIGURE 4.6: Swapped sidechain-mainchain alternate labels in catalase. Top: The original alternate conformations for Leu105 are mismatched – the A sidechain is matched with the B backbone and *vice versa* – resulting in Cβ deviations > 0.4 Å. Bottom: Simply swapping the backbone alternate labels solves the problem.

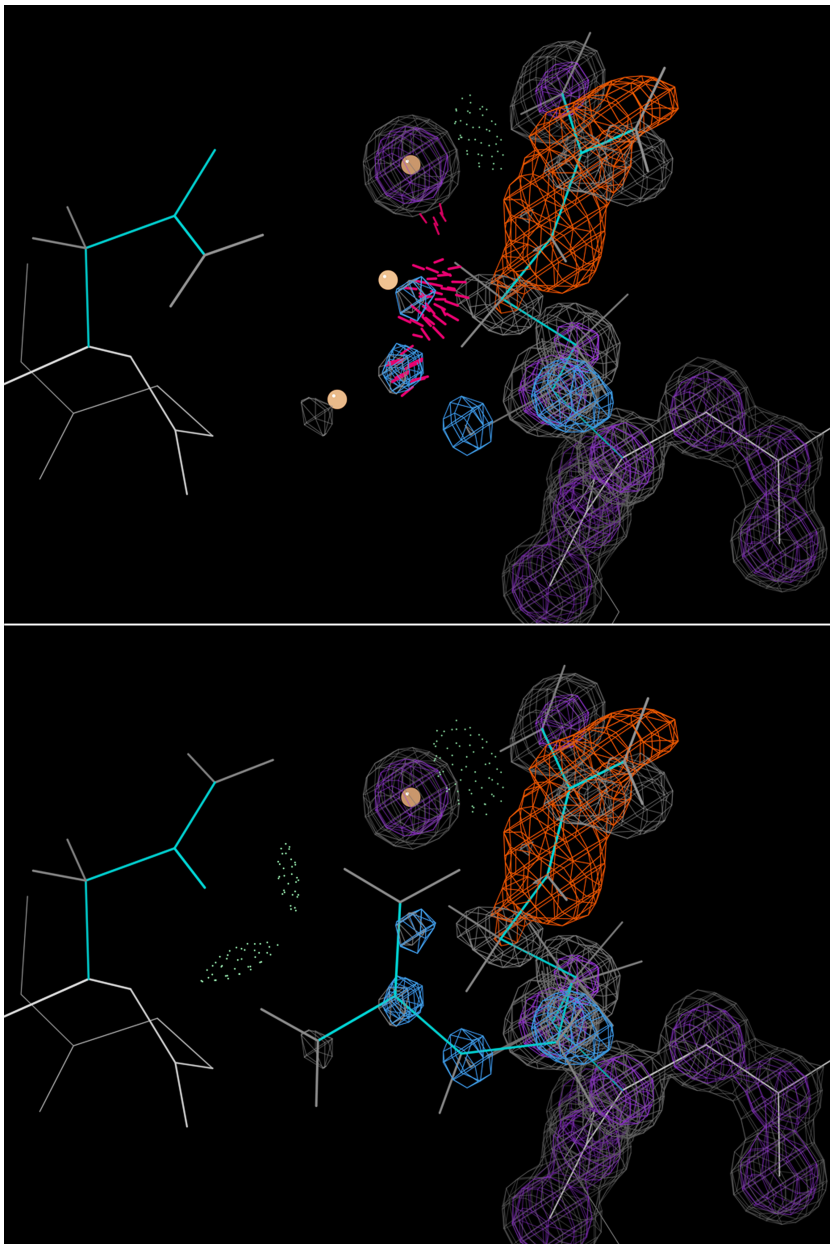

FIGURE 4.7: Hidden alternate Arg sidechain in catalase. Top: In the deposited structure, the Arg464 sidechain (cyan, right) is a poor fit to the negative Fo-Fc density (orange). Two waters (tan balls, middle) are also poor fits to the positive Fo-Fc density (blue) and to very weak 2Fo-Fc density (gray, purple) compared to a well-ordered water (top/background); as a result, they clash (pink spikes) with Arg464. Bottom: Replacing the waters with an alternate Arg464 sidechain (cyan, middle) nicely fits the density. After a concomitant amide flip of the adjacent Asn173 (cyan, left), favorable H-bonds (green pillows) are formed, completing the refit.

In a similar but more elaborate example, Arg486, Glu483, and Glu448 should be part of an alternate sidechain network (Fraser et al., 2011) but only the latter is deposited with alternates (Figure 4.8). Several clashes and poor fit to the density flagged this region as suspicious. I reconstructed the network by adding an alternate B for Glu483, swapping alternate labels for Glu448 based on its interactions with Glu483, adding an alternate A water to be mutually exclusive with Glu483 alternate B and form an H-bond with a nearby amide, adding an alternate B for Arg486, and adjusting occupancies for all sidechains and waters to be consistent across the network (Figure 4.8).

In this example, alternate A should certainly be assigned a slightly higher occupancy than alternate B based on differences in electron density; I chose a collectively exhaustive 70% and 30% based on visual inspection. Yet if additional low-level "hidden" conformers also exist (alternates C, D, ...), then the occupancies for alternates A and B should add to something less than 100%, though A should still be higher than B. This issue is directly germane to eliminating logical fallacies like steric clashes. For example, two alternate sidechains from adjacent residues with a mutually exclusive atomic overlap would clash if they both had 51% occupancy (with a second alternate at 49%) regardless of which alternate states they were assigned to, because they would by implication co-exist in the molecule at least some fraction of the time (at least 2% and at most 51%). However, they would not clash if they both had 49% occupancy (with a second alternate at 41% and a minor third at 10%), since they could then exist in the molecule at different times (i.e. belong to different alternate states). Unfortunately, optimization of occupancy assignments is far from a solved problem, in part because most crystallographers instead choose to simply refine B-factors when mobility appears to be present. Therefore, identifying or inferring the presence of additional conformers will be critical to future efforts at creating self-consistent alternate networks.

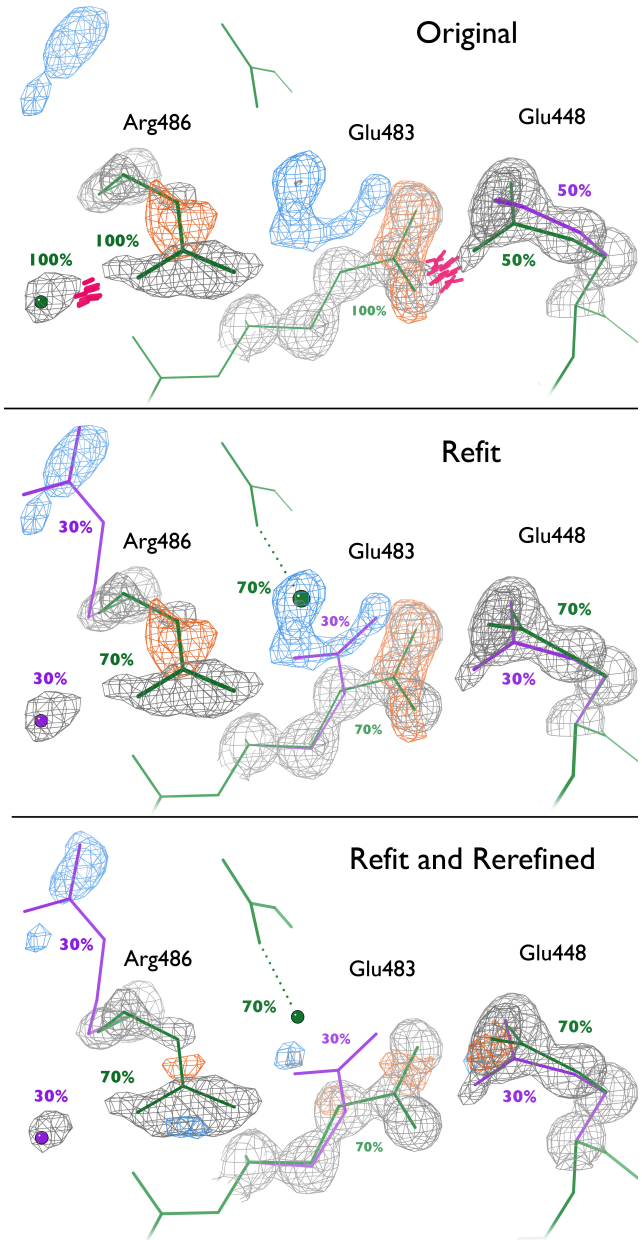

FIGURE 4.8: Extension of an alternate network in catalase. Top: The deposited structure (1gwe) has 50%/50% alternate conformations (green: A, purple: B) for Glu448, but single conformations for the adjacent Glu483 and Arg486 – this in spite of severe steric clashes (pink) and poor fit to the negative (orange) and positive (blue) Fo-Fc electron density. Middle: The refit model has alternate conformations for all three coupled sidechains with consistent 70%/30% occupancies. An alternate A water (green ball in middle) that is mutually exclusive with Glu483 alternate B and forms an H-bond with the nearby Glu488 (top) has also been added. Bottom: Re-refinement with the refit model eliminates many of the difference density peaks.

Relationships between alternate conformations are evaluated in light of contacts between adjacent sidechains, which are typically effected by hydrogens. Unfortunately, hydrogen atoms are usually invisible in crystallography. However, at ultra-high resolution, as with with 1gwe at 0.88 Å, I was able to see evidence for hydrogen positions in the form of positive Fo-Fc density peaks, and thereby evaluate the successes and failures of automated hydrogen placement by Reduce (Word et al., 1999a). I soon discovered that some apparent steric clashes were due to simple mistakes in Reduce.

For example, Reduce's Hβ hydrogens for Trp216 imply clashes with the surroundings in both directions that cannot be alleviated by χ1 tweaks (Figure 4.9, top). I explored the possibility that these hydrogens have relatively unique chemical character due to strain in the Cα-Cβ-Cγ bond angle, but Trp216's Cα-Cβ-Cγ is actually quite typical for its χ1 bin (Figure 4.9, bottom). It seemed quite likely, then, this problem derives from the original choice to use hydrogen bond-lengths based on the positions of nuclei instead of the centers of electron clouds in Reduce (Word et al., 1999a) and van der Waals radii compatible with those bond lengths in Probe (Word et al., 1999b). Other members of the Richardson lab, including Lindsay Deis and Bryan Arendall, have now worked with members of Jack Snoeyink's lab, including Vishal Verma, to address this problem by updating Reduce's hydrogen bond-lengths (and Probe's van der Waals radii accordingly) such that they at least roughly match both those in PHENIX and those implied by spherical fits to quantum-mechanics-derived electron density maps computed by Nigel Moriarty for simple systems such as isolated amino acids. To further investigate 1gwe Trp216, I used their preliminary updated versions of Reduce and Probe (informally dubbed Reducer and Prober), which for methylene hydrogens have shorter hydrogen bond-lengths by 0.13 Å and longer van der Waals radii by 0.05 Å. Indeed, the Trp216 Hβ clashes are resolved, supporting the idea that the problem in this case was with the details of our all-atom

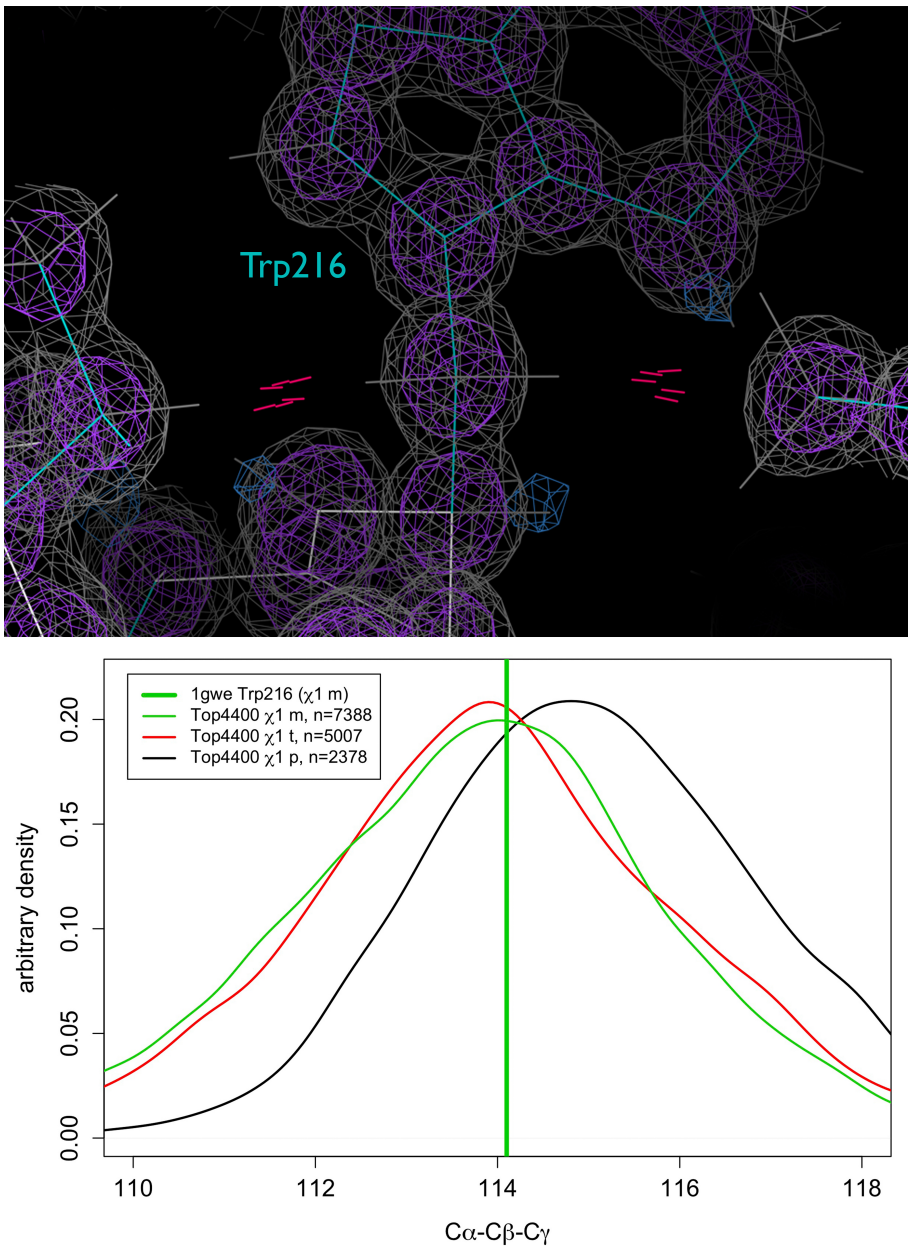

FIGURE 4.9: Wrong bond lengths for Trp Hβs in catalase. Top: The hydrogens Reduce places on the Cβ of Trp216 in 1gwe clash with the nearby Ile269 Hβ (left) and the Thr204 Cγ2. All three heavy atoms – Trp216 Cβ, Ile269 Cβ, and Thr204 Cγ2 – are well supported by the 2Fo-Fc density, so the problem is not with their positioning. Bottom: Trp216's Cα-Cβ-Cγ bond angle is normal given that its χ1 is in the m bin, so its Hβs would not be expected to flex closer to or farther away from the Cα-Cβ-Cγ plane relative to expectation. Subsequent analysis with preliminary updated C-H bond-lengths and van der Waals radii (not shown) showed that the precise details of our all-atom contact parameters were at fault.

contact parameters.

In a similar category, Reduce incorrectly added hydrogens to methyls on the central heme in an orientation in which one hydrogen was fully eclipsed instead of having one fully staggered and the other two about 30° from eclipsed (Figure 4.10). Aram Han of Jack Snoeyink's lab has now addressed this issue in the Reduce code.

Other methyls were very close to correct in perfectly staggered orientations, but required minor tweaks ($< 10°$ rotation) in response to local packing constraints. For example, Ile55 has an unacceptable steric clash to a nearby Gln methylene group; a small methyl rotation resolves the clash and fortuitously falls directly onto a positive Fo-Fc peak (Figure 4.11). Reduce needs to be quite conservative when it comes to methyls – automated methyl rotations are turned off by default since they have been observed to go awry frequently and easily if the attached heavy atom is not placed quite perfectly – so I coded a new methyl-rotation tool in KiNG to manually remodel the problematic methyls in catalase. This illustrative example happens to be at a tetrameric contact (see below), but a few other similar ones occurred in buried monomeric regions.

Similarly, some hydroxyls were misplaced because Reduce failed to completely account for their surroundings. To wit, Ser397's hydroxyl is appropriate considering just the asymmetric unit, but it becomes obvious it is impossible in the context of the biological tetramer (Figure 4.12). This failure occurs because Reduce usually operates on the asymmetric unit only. Vishal Verma of Jack Snoeyink's lab has now addressed this issue by encoding space group symmetry consideration in Reduce; the new version is being tested within PHENIX and should be available in MolProbity soon.

As alluded to in the past two examples, I performed my initial batch of fixes and Reduce runs on the 1gwe asymmetric unit, but catalase is actually a biological tetramer. Reproduction of a tetramer unit cell for my fixed-up model through space

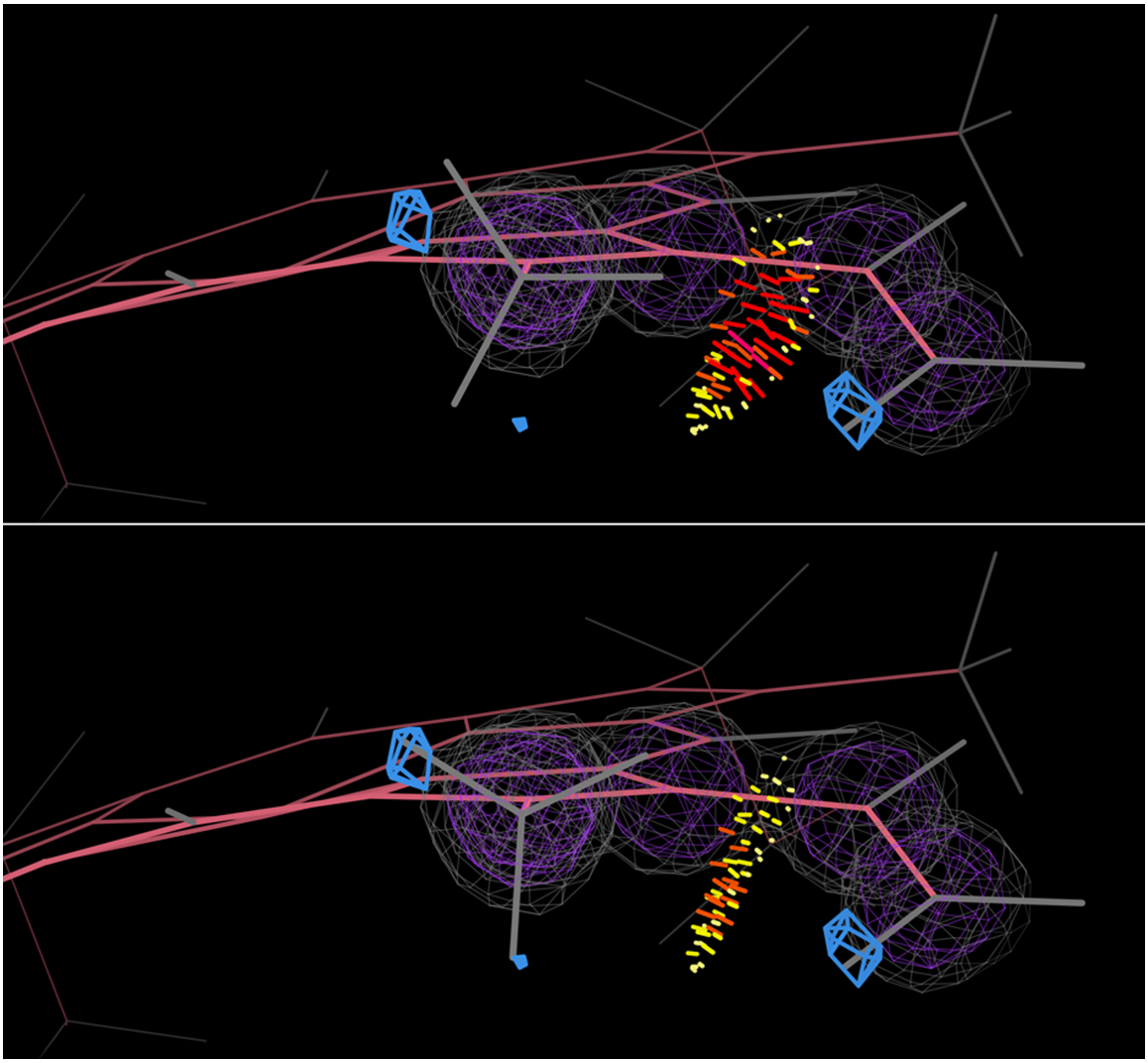

FIGURE 4.10: Heme methyl rotation to alleviate clash in catalase. Top: Reduce places the hydrogens on the CMC methyl group of the central heme such that one is fully eclipsed and thus clashes with the adjacent methylene group. The positive Fo-Fc density peaks (blue) strongly corroborate MolProbity's clash in refuting Reduce's placement. Bottom: A manual refit more correctly places the methyl hydrogens in the one of its two possible non-eclipsed positions that better fits the Fo-Fc density.

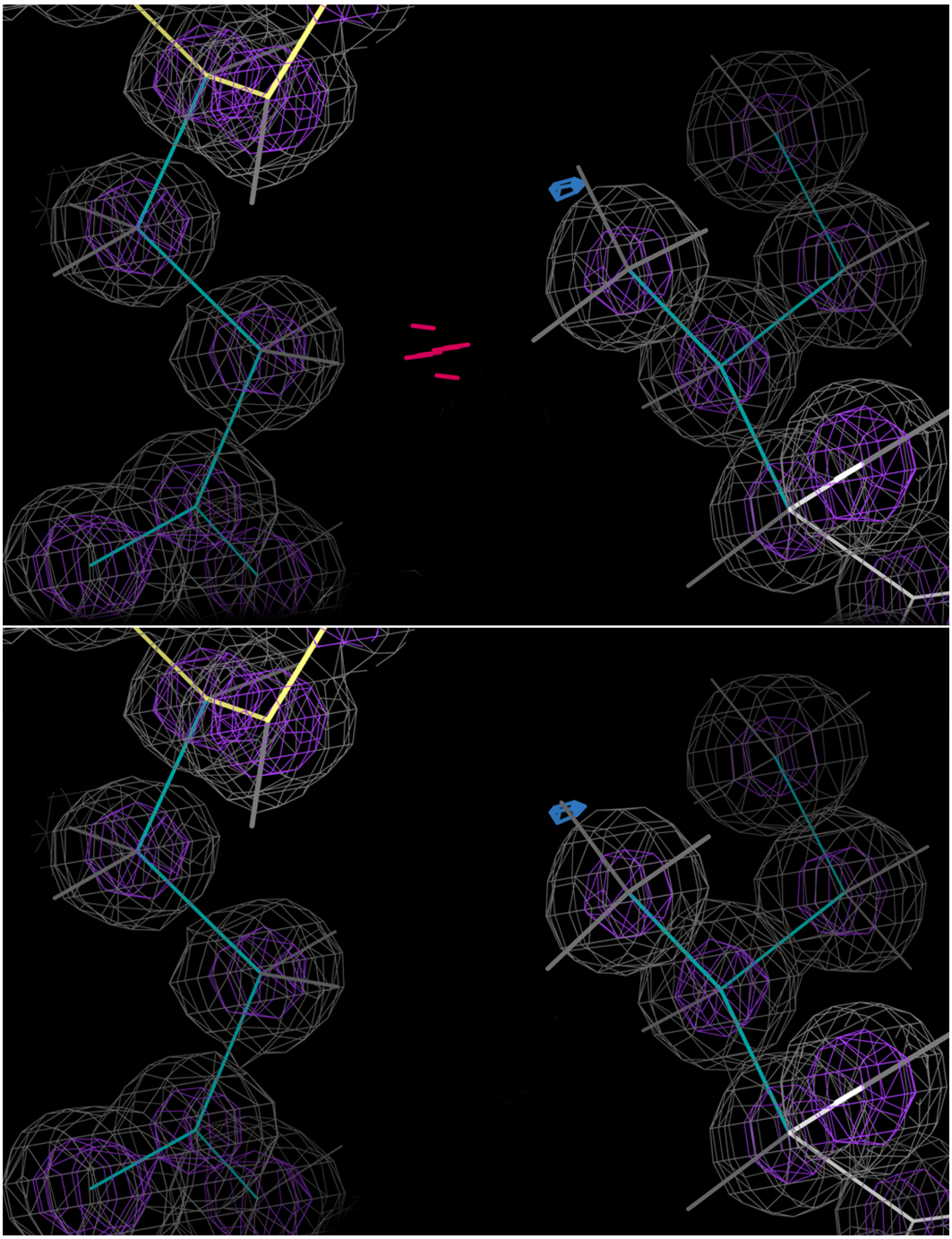

FIGURE 4.11: Ile methyl rotation to alleviate tetramer clash in catalase. Top: Ile55 (right) has a clash to Gln375's Cγ hydrogens (right) given Reduce's default hydrogen placement. Bottom: A small (9°) rotation eliminates the clash and fits a positive Fo-Fc density peak (blue).

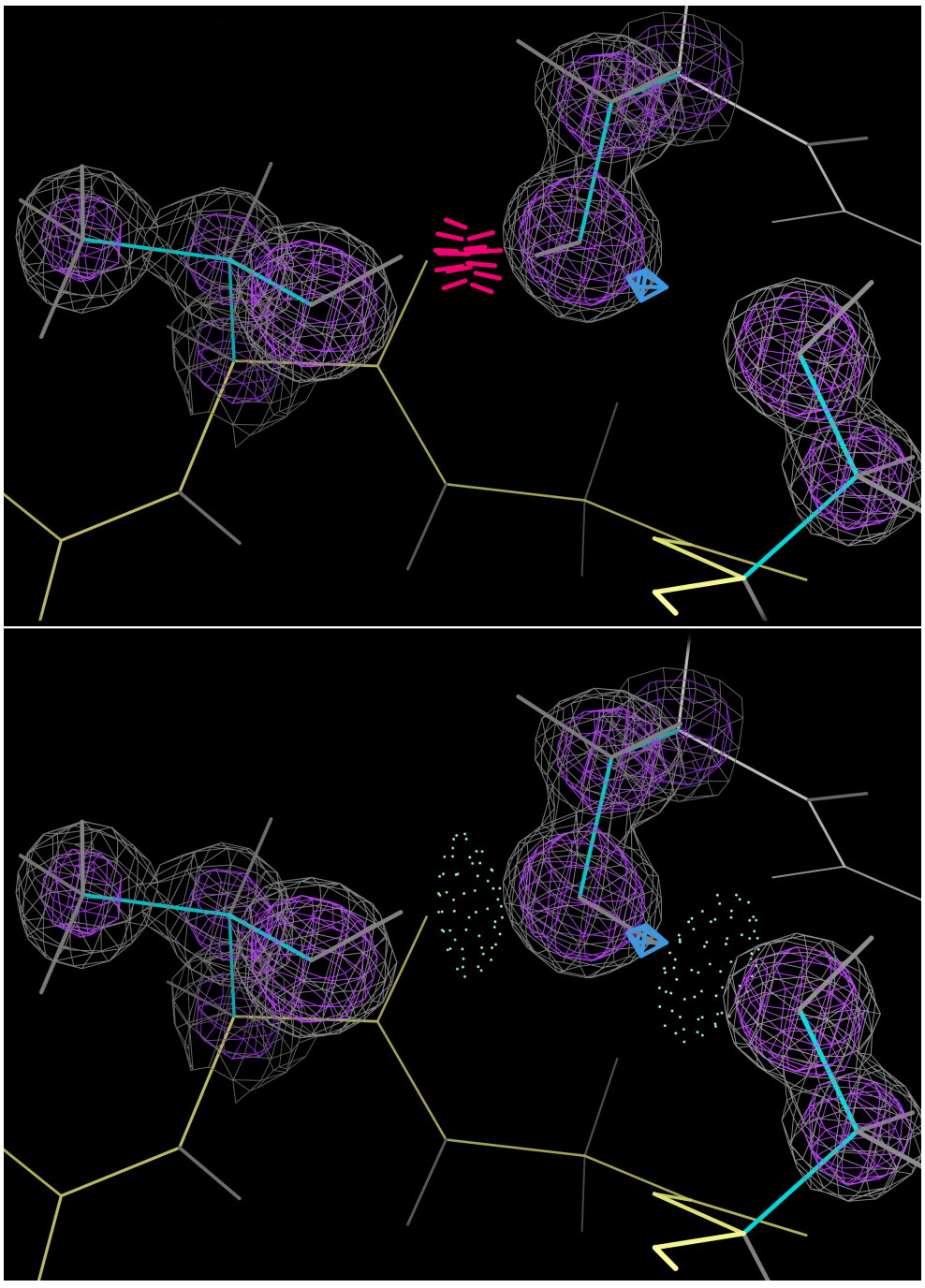

FIGURE 4.12: Ser hydroxyl rotation at tetrameric contact in catalase. Top: Given Reduce's default hydrogen placement, the hydroxyl hydrogen of Ser397 (middle) from chain A (white backbone) clashes with that of adjacent Thr11 (left) from chain B (yellow backbone). Bottom: An $\approx 80^\circ$ manual hydroxyl rotation fits a positive Fo-Fc density peak (blue), eliminates the clash to Thr11 and establishes an H-bond in its place, and also establishes an H-bond to Ser13 (right) of chain B.

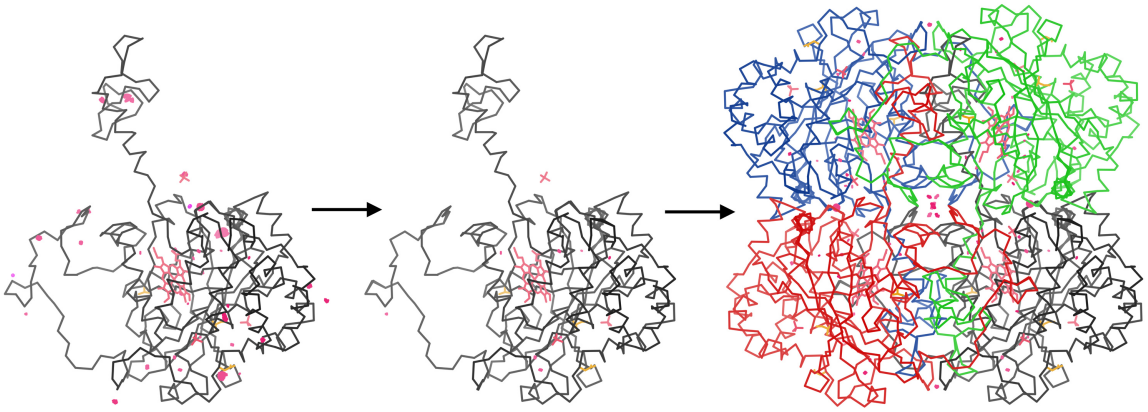

FIGURE 4.13: Tetrameric crystal contacts in near-paragon catalase. The deposited structure 1gwe (left) is improved upon by the refit model (middle) in terms of clashes, but reconstructing the biological tetramer (right) reveals several clashes at tetrameric crystal contacts, the worst of which (e.g. red clashes at very center of rightmost panel) are due simply to naming issues (see Figure 4.15).

group symmetry revealed a smattering of "new" clashes that arose at intermolecular/crystal contacts (Figure 4.13). Some of these, such as the methyl and hydroxyl rotations described above, required relatively simple atomic coordinate changes to address.

However, a pair of remaining clashes are caused by the more subtle problem of "local asymmetry", a phenomenon which has also been noted in other homooligomers (Goodsell and Olson, 2000). The first centers around the Leu105-Gly106 peptide, which is adjacent to a two-fold axis of symmetry within the unit cell. Because of this unusual location, Leu105 in chain A clashes with another symmetry-related copy of itself in the adjacent chain B (Figure 4.14, left). In the real molecule, it is clear that what is labeled alternate A in chain A must actually be paired with what is labeled alternate B (not A) in chain B. To model this in the structure, one must swap the alternate labels in either chain A or chain B (Figure 4.14, right). As noted by the depositors (Murshudov et al., 2002), however, such a change would destroy the

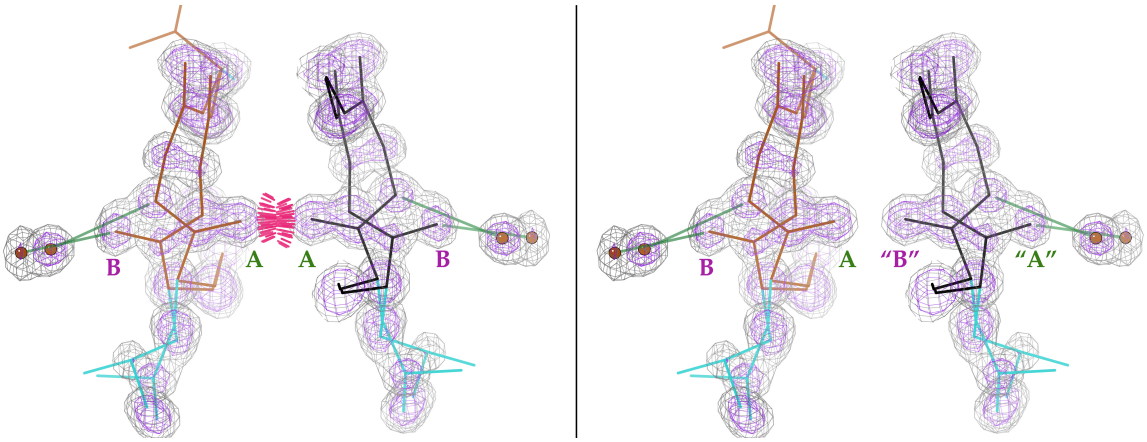

FIGURE 4.14: Broken symmetry for Leu105 at tetramer contact in catalase. Left: The near-paragon model has "self clashes" between alternate A of Leu105 in chains A and B. Right: The problem can be alleviated by effectively swapping half the alternate labels, but the original space group is no longer valid by the same logic described for Figure 4.15.

space group symmetry and necessitate a tetrameric unit cell with a different space group. Strictly speaking, then, the current $P4_22_12$ space group is inappropriate. Intriguingly, this backbone non-uniqueness is conserved among many catalase species and is adjacent to highly conserved, catalytically essential residues, so it may in fact be functionally relevant (Murshudov et al., 2002). Tyr378 is an analogous case with a sidechain-sidechain instead of backbone-backbone clash (Figure 4.15), although in this case the local non-uniqueness is not thought to be functionally relevant based on the distance from the active site.

Subsequent re-refinement of my paragonized model with PHENIX (Adams et al., 2010) led to very minor coordinate changes: most atoms moved less than 0.5 Å and the vast majority moved much less than 0.1 Å. My fixes were therefore quite compatible with the experimental data.

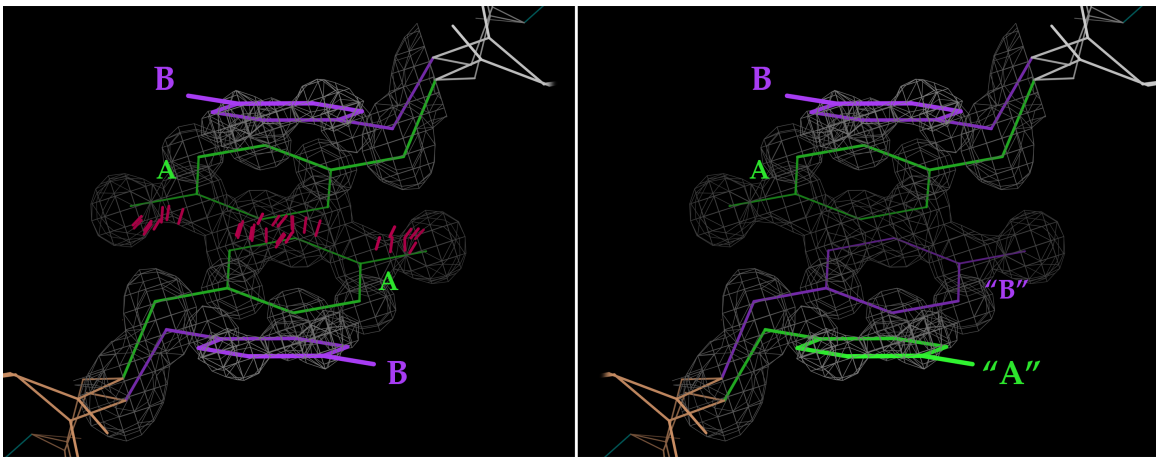

FIGURE 4.15: Broken symmetry for Tyr378 at tetramer contact in catalase. Left: The near-paragon model has "self clashes" between alternate A (green) of Tyr378 in chain A (top-right, white backbone) and alternate A of Tyr378 in chain B (bottom-left, peach backbone). This occurs because Tyr378 happens to coincide with an axis of symmetry for the tetramer, such that instantiating alternate A implies an impossible model. Right: The problem can be alleviated by effectively swapping half the alternate labels, i.e. assigning alternate B (purple) to one of the two original alternate A sidechains. Now alternate A and alternate B each imply a possible model, but the original space group – by which these symmetric chains were placed by a sequence of rotations and translations – is no longer valid because the subunits are no longer truly identical.

4.4 Discussion

High-resolution data in crystallography (typically defined as better than 1.2 Å) contains information about small inter-atomic spacings, so is is usually very desirable for a crystallographer who wishes to define his/her protein's structure as accurately as possible.

However, the increased atomic discernment possible at high resolution also reveals the presence of well-populated substates that would be obfuscated or invisible with medium- or low-resolution data. Alternate conformations are identifiable even in the best structures in sets of no more than about 3 and with no less than about 10% occupancy; because a 90%-to-10% ratio implies an energy difference of only about 2.2 kT, these observable conformations co-exist with nearly equal energies. Therefore, with greater power to discern atomic positions comes greater responsibility to coalesce multiple conformations from the native-state ensemble into a cohesive structural model.

We were forced to consider this tradeoff in our quest for paragon structures (Section 4.2), which are essentially "Platonic ideals" with no detectable errors whatsoever (given our assumptions about physics and chemistry as encoded in MolProbity, the limitations of the PDB file format, etc.). Indeed, the only pre-existing, ready-made paragons we identified were at medium-to-high resolution, where only a single conformation was visible.

On the other hand, my efforts to create a paragon starting from the 0.88 Å structure 1gwe (Section 4.3) highlighted the complexities of dealing with networks of often mutually exclusive alternate conformations. In failing to eliminate literally all MolProbity errors, I also stumbled upon several modeling issues that should probably be reconsidered, among them the hydrogen bond-lengths used by Reduce and how to assign space groups when alternate conformations occur at axes of symmetry.

The failed effort to completely "paragonize" 1gwe was thus very educational, and should inform near-future efforts to build more powerful tools for defining multi-conformer models. Ideally, such an algorithm would be able to automatically model convoluted alternate networks based on repulsive steric clashes, attractive H-bonds or charge-charge interactions, presence vs. absence of ordered solvent molecules, and other factors. Ultimately, the multi-conformer models made possible by these advances will be a rich data source for understanding the broad biological implications of near-native-state conformational ensembles in proteins.

5

Torsional Bioinformatics

5.1 Torsional validation in MolProbity

Accurate macromolecular structural modeling is impeded by many obstacles; chief among them are the vastness of conformational space and the difficulty of efficiently and precisely evaluating free energies. The magnitude of the conformational search space varies widely by application: *ab initio* protein structure prediction requires large fractions to be considered, whereas crystallographic refinement typically focuses on conformations relatively similar to the model being refined. However, essentially every protein modeling task involves a scoring component in which various competing conformations are weighed and counterweighed based on their free energies (or, when entropy is ignored (Hu and Kuhlman, 2006), simply energy, i.e. internal energy or enthalphy). For example, proposed conformational or configurational changes, such as small backbone movements, rotamer jumps, and/or mutations, may be accepted or rejected based on their computed energies, ultimately leading to the overall lowest-energy conformation and/or sequence.

Unfortunately, this step remains difficult, because accurate methods deeply rooted

in state-of-the-art physicochemical theories are too computationally expensive to be extensively used, and simplified versions using only classical physics and idealized chemistry are often too inaccurate for precise atomic-level calculations. Various approaches to scoring functions have been attempted, with different strengths and weaknesses.

Molecular mechanics force fields such as Amber (Cornell et al., 1995) (see Section 2.4) use a simple functional form comprised of additive contributions from various physicochemical forces: harmonic restraints on bond lengths, angles, and most dihedrals; van der Waals attraction and repulsion with a 6-12 Lennard-Jones potential; and Coulombic electrostatic interactions between atom-centered point charges. The coefficients on these terms, which determine the relative contributions of different energetic influences, are generally derived by choosing values resulting in simulations that better reproduce experimental thermodynamic measurements for very simple systems.

An alternative philosophy is hybrid physical-statistical scoring functions (use of the phrase "energy functions" to describe such systems is considered *passé* in some quarters). This genre is best exemplified by Rosetta (Rohl et al., 2004), which uses two separate scoring functions with different granularity, the first to generate plausible folds in the initial stages of *ab initio* structure prediction, and the second to refine proposed conformations in the neighborhood of the native conformation. In the low-resolution/reduced-representation scoring function, a sidechain centroid is used to represent each residue, and solvation, electrostatic, and H-bonding effects are modeled probabilistically on a residue level. To propose plausible conformations for local regions, Rosetta borrows "fragments" from experimentally determined structures; importantly, it is assumed that these conformations implicitly reflect the most important local energetic considerations, and therefore that intra-fragment energetic evaluation is unnecessary. In the high-resolution/all-atom scoring function,

by contrast, explicit van der Waals and H-bond terms and an implicit solvation term (Lazaridis and Karplus, 1999) play critical roles. At this stage, relying on implicit energetics within fragments is insufficient, because even small torsional changes resulting from minimization may have large energetic effects; therefore, probabilistic Ramachandran (ϕ and ψ angles) and rotamer (χ angles) terms are used instead. This two-pronged approach to "energy" evaluation, coupled with fast (albeit incomplete) conformational search, has enabled much of Rosetta's well-documented success.

Keating and colleagues have taken a quite different approach that has enabled rapid evaluation of the energy of an arbitrary sequence on a given backbone fold (Grigoryan et al., 2006). They first used the deterministic DEE and A* algorithms to find the optimal rotamer combinations for tens of thousands of sequences threaded onto the backbone of interest, using a structure-based (e.g. molecular mechanics) energy function. After this one-time computational investment, they were able to fit a simple linear model relating sequence to energy by assigning weights on terms for various singles, pairs, triple, etc. of amino acids at specific residue positions. The result was ultra-fast energy evaluation (a factor of 10^7 speed-up) for any given sequence on the backbone used for training, but there was a cost in accuracy (on the order of 1-5 kcal/mol) relative to the energy function being approximated. However, given that the types of energy functions used for structure-based protein design are themselves imperfect, the authors argued that this fitting error is acceptable in light of the massive gain in speed.

Our response to the hurdles of evaluating macromolecular energetics is to learn what conformations are realistic – and which ones aren't – directly from experimental structures themselves. MolProbity (Chen et al., 2009c) embodies this approach with a conglomeration of empirically based assessments.

Expected bond lengths and angles involving protein backbone heavy atoms are taken from standard values (Engh and Huber, 2001), which are derived from polypeptide-

like fragments of small-molecule crystal structures in the Cambridge Structural Database (CSD) (Allen, 2002). The Cβ deviation, a measurement of proper tetrahedral geometry at the Cα locus (Lovell et al., 2003), is computed relative to an ideal Cβ position that is defined by a combination of these ideal backbone parameters.

Similarly, the covalent bond lengths and angles used to place hydrogens (Word et al., 1999a), which are invisible to most crystallographic experiments, and the van der Waals radii used to subsequently evaluate all-atom non-covalent packing interactions (Word et al., 1999b; Bondi, 1964), the majority of which involve hydrogens, are based on empirical distributions in small-molecule crystal structures. However, these parameters reflect nuclear positions rather than centers of the electron clouds, which actually underlie inter-atomic steric interactions, and are therefore slightly too long at present. Therefore, work is ongoing by other members of the Richardson lab and our collaborators to define shorter, electron-cloud-centric bond lengths – and correspondingly longer van der Waals radii – that will be more appropriate for macromolecular structure validation.

MolProbity also makes strong use of four-body torsions for validation: particular local conformations are compared against empirical distributions of ϕ and ψ backbone dihedrals and χ sidechain dihedrals. Because these distributions have been derived using high-resolution, quality-filtered protein crystal structures, they are reliably useful for structure validation: we can claim with confidence that outliers falling outside their populated regions are at best unusual and at worst erroneous. Such forbidden regions can be roughly mapped based on implied steric clashes in theoretical conformers with given torsional combinations, but their precise boundaries are best defined using large amounts of good data in the form of high-quality crystal structures.

In this chapter, I describe updates to these almost decade-old torsional distributions, made possible by the continuing expansion of individual and high-throughput

structural biology efforts. The new versions do not radically alter our concept of protein backbone and sidechain energetics, but they do qualitatively improve our ability to distinguish conformations that are disfavored but possible from those that are surely disallowed.

5.2 Building a bigger and better data set

The Protein Data Bank has grown rapidly in the years since the Richardson lab's most recently published quality-filtered structural database for proteins, the Top500 (Lovell et al., 2003). In 2007, lab members attempted to take advantage of this new data by creating the Top5200. It maintained similar standards of resolution and structure quality but, due to sheer logistics, required a more automated selection protocol that made use of PDB homology clusters (updated weekly) and MolProbity score. Unfortunately, chains with MolProbity score > 2.0, up to > 2.7 in some cases, were unintentionally included – although only chains from structures with resolution < 2.0 Å were included, as intended. In 2010 I helped implement a stop-gap successor, the Top4400, by simply eliminating all chains in the Top5200 with MolProbity score > 2.0. This database was inherently suboptimal because no attempt was made to find suitable replacements.

The Top5200 was used for my investigations of mutation-coupled backrubs (Chapter 2) and the Top4400 was used for the Validation Task Force paper (Read et al., 2011), and both have been distributed informally to collaborators, but otherwise neither was fully accepted as our next-generation database (e.g. neither is available on our website).

To facilitate new structural bioinformatics studies, we have now constructed the Top8000 databases of high-quality protein structures. We ran Reduce (Word et al., 1999a) on all crystal structures in the PDB as of March 29, 2011 containing at least one protein chain with $\geqslant 38$ residues (according to the MolProbity "oneline" script),

in order to allow Asn/Gln/His flips (`reduce -flip`) throughout the structure, including at interfaces where multimer partners may participate in hydrogen-bonding networks. Single protein chains were then extracted along with any "het" atoms or waters with the same chain identifier.

Next, for each chain, a MolProbity score (Davis et al., 2007; Chen et al., 2009c; Keedy et al., 2009) (see also Section 6.2) was calculated. This score is an estimate of the resolution at which a structure's steric clashes, rotamer quality, and Ramachandran quality would be average. Thus the average of resolution and MolProbity score is a combined experimental and statistical indicator of structural quality.

In terms of filtering, chains that were marked by the PDB as obsolete as of April 13, 2011, from structures retracted in the Murthy (University of Alabama at Birmingham) falsification scandal (`http://www.wwpdb.org/UAB.html`), atomically incomplete ($< 25\%$ of residues with sidechains), or too short (< 38 residues) were eliminated. Finally, the PDB's chain-level homology clusters as of March 29, 2011 (actually released earlier that week on March 25, 2011) were downloaded.

After conversations within the lab, we required each chain to have resolution $<$ 2.0 Å, chain MolProbity score < 2.0, $\leqslant 5\%$ of residues with bond length outliers ($> 4\sigma$), $\leqslant 5\%$ of residues with bond angle outliers ($> 4\ \sigma$), and $\leqslant 5\%$ of residues with Cβ deviation outliers (> 0.25 Å). We then selected the best chain (in terms of average of resolution and chain MolProbity score) per homology cluster. There was a small number of ties within clusters (for $< 1\%$ of the final chain tallies); these were resolved, arbitrarily but reproducibly, by alphabetical order of PDB code + single-character chain ID.

The step of selecting the best chain from each homology cluster was done separately for the 50% ("most stringent" homology filtering), 70%, 90%, and 95% ("least stringent" homology filtering) clustering levels. The homology filters, to varying levels, prevent redundancy and thus over-representation of certain motifs or sub-

structures. Moreover, by comparing the results of a bioinformatics study with the different homology filters, one can consider the influence of evolutionary relatedness on a given feature's prevalence as opposed to simple energetic favorability; our previous data sets were collated using only the 70% filter and therefore did not allow such analysis.

Availability of electron density maps from the Electron Density Server (EDS) (Kleywegt et al., 2004) was tabulated but not used for selection for the "primary" versions of the Top8000. However, a set of "secondary" versions was also compiled in which the availability of a map in the EDS was required during the database creation process for each entry. Map availability could also be used as a less arbitrary tiebreaker (see above) for future data sets, since a lack thereof could suggest that the authors were reluctant to deposit their diffraction data or that the EDS was unable to successfully produce an acceptable map given the deposited data, either of which is worrisome.

Originally a filter was planned to eliminate "suspiciously good" chains with resolution better than 1 Å, chain MolProbity score = 0.5 (the optimal score), and clashscore = 0 (the optimal score). However, we determined that such a filter would eliminate several true "paragons" (truly error-free models), and thus we did not use it.

The "geometry" filters (bond lengths and angles and Cβ deviations) were not used in the Top500, Top5200, or Top4400 – they are new to the Top8000. Fortunately, although these filters eliminate a substantial number of individual chains, they eliminate a remarkably small number of homology clusters; this is because in many such clusters, some *other* chain passes the geometry filters, and thus can represent the cluster. We therefore gain in quality within each cluster with little loss in quantity of clusters.

Chain MolProbity scores, as opposed to file MolProbity scores, were used here,

Table 5.1: Residue counts in Top8000 versions vs. older data sets. *Our most-used, default set, with approximately 8000 chains, hence the name Top8000.

Data Set	EDS Requirement	Homology Filter	Chains	Residues (no filter)	Residues (mc B < 30)
Top500	no	70%	500	109,799	97,368
Top5200	no	70%	5199	1,195,418	1,176,398
Top4400	no	70%	4,403	1,010,596	974,917
Top8000	no	50%	7,232	1,686,207	1,430,119
Top8000	no	70%	*7,957	1,858,193	1,573,349
Top8000	no	90%	8,562	1,990,507	1,680,600
Top8000	no	95%	8,825	2,041,499	1,720,878
Top8000	yes	50%	6,107	1,444,973	1,228,077
Top8000	yes	70%	6,663	1,575,216	1,336,807
Top8000	yes	90%	7,138	1,678,587	1,419,978
Top8000	yes	95%	7,342	1,719,227	1,451,433

although in the future, other lab members may investigate some alternative that accounts for both or applies some average chain MolProbity score "correction factor" differentially for chains from single-chain vs. multi-chain structures.

Table 5.1 contains the final counts of protein chains and protein residues (without and with a mainchain B-factor < 30 filter, for Ramachandran analysis) in the Top8000 versions without the EDS requirement (the "standard" or "normal" version) and with the EDS requirement (a special-case use version) as compared to previous data sets.

All Top8000 single-chain PDB files plus additional chain- and residue-level data are available in the supplemental material for this thesis.

5.3 Updating Ramachandran analysis

One of the earliest observations pertaining to protein structures, coming nearly 50 years ago (with its official 50th anniversary celebration in Bangalore in January 2013), was that mainchain ϕ and ψ torsion angles are highly correlated and that ϕ,ψ

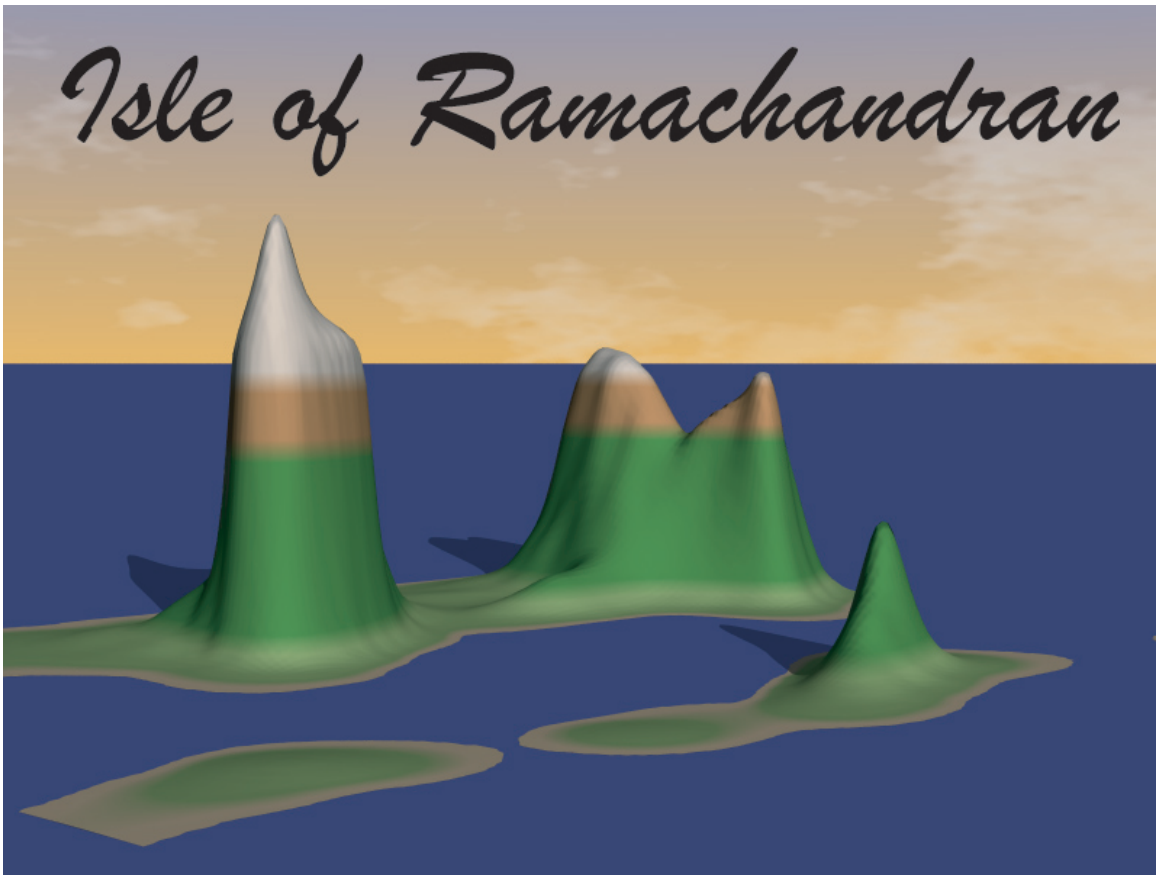

FIGURE 5.1: The Ramachandran archipelago. "Land" represents favored regions, "shores" represent allowed regions, and "water" represents sterically or otherwise disallowed regions. The highest "snow-capped peak" is α helix; the next-highest are β sheet and polyPro. Credit: Ian Davis.

pairs fall only into certain specific regions of the "Ramachandran plot" (Ramachandran et al., 1963). Topographically speaking, the two most prominent mountains correspond to α helix and β sheet, other shoals and plateaus correspond to less populated but still nonetheless genuinely observed conformations, vast oceans are almost entirely vacant due to local steric repulsions between backbone atoms, and ocean shores represent energetically unfavorable conformations that are occasionally observed because of compensating favorable interactions such as hydrogen bonds (Figure 5.1).

The Ramachandran plot has proved invaluable for a cornucopia of protein structure modeling tasks. For structure prediction and design, one would like to accurately estimate the *most likely* conformation for a given residue. To do so, it can be useful to compile a separate ϕ,ψ distribution for each amino acid type, and perhaps even a separate sub-distribution for each neighboring amino acid type (Ting et al., 2010). However, despite the size of the PDB, this approach can still be difficult due to insufficient availability of well-determined residues in high-resolution structures that thus have precisely positioned backbone atoms. One solution is a Bayesian approach that deviates from global expectation only when the data is sufficient to support such a conclusion (Ting et al., 2010).

The goal of structure validation, on the other hand, is to determine whether or not the conformation of a given residue in a known structural model is *reasonable*, by cleanly differentiating energetically disfavored but allowed values (that generally avoid clashes by opening bond angles slightly) from those that are physically possible only under very unusual local circumstances. This differs fundamentally from validation using the entire distribution, in which a structure would be scored based on its adherence to the database-wide distribution; such an approach would be inappropriate not just for unusual/iconoclastic structures, but for any structure that did not adhere quite closely to database-wide averages of secondary structure and amino acid content (e.g. an all-helical structure, or one without any helices).

For example, the γ turn (Figure 5.2), in a sparse positive-ϕ region of the Ramachandran plot below Lα, is labeled a serious outlier by ProCheck, which uses unsmoothed distributions made with older, unfiltered data (Laskowski et al., 1993). However, it is disfavored but allowed according to MolProbity, which uses smoothed distributions made with more, quality-filtered data (Lovell et al., 2003); that continues to be true in my new Top8000-based distributions (see below). This statistical analysis reflects the underlying chemical reality that the γ turn's i-1 to i+1 mainchain

H-bond at least partially countervails its steric strain (Figure 5.2).

As a result, the outer fringes of the distributions are of more pressing interest for validation than are the peaks, so it is preferable to avoid the problems associated with insufficient data for rare amino acids, and instead compile a minimal set of categories in order to better define the "shores". Most residues can be collapsed into a "general case" category, with separate categories only for amino acids with significantly different distributions. In the past, only glycine and proline were treated separately (Laskowski et al., 1993; Vriend, 1990; Kleywegt and Jones, 1996), but residues immediately preceding proline were also singled out from the Top500 (Lovell et al., 2003), as suggested previously (Karplus, 1996).

However, with an order of magnitude more data than the Top500 (Table 5.1), we can now more finely subdivide these residue-type categories and still maintain excellent outer contour definitions for validation purposes. For example, we suspected that residues with hydrophobic, branched-Cβ sidechains, namely isoleucine and valine, could be moved to a separate category, and that proline could profitably be separated into *cis* and *trans* categories; this was tested for the report of the wwPDB X-ray Validation Task Force (VTF) (Read et al., 2011). I tested this idea using the Top4400 data set (our most reliable data set available at the time) by comparing the contours for the 16 amino acids used for the general category (excluding Gly, Pro, and pre-Pro as always, and now also excluding Ile and Val) to each other. I observed only minor differences (Figure 5.3), supporting the decision to separate out Ile/Val and *cis* vs. *trans* Pro but to otherwise maintain the paradigm of a "general case" category. I also investigated including Thr with Ile/Val because it is also has a branched-Cβ sidechain, but discovered only after looking that its distribution is more like the general case than like Ile/Val; it is probably the most deviant of the 16 but not badly so. We are fairly sure the cause of this is that the hydroxyl is much smaller than a methyl or methylene, so that Thr does not have as much steric clash

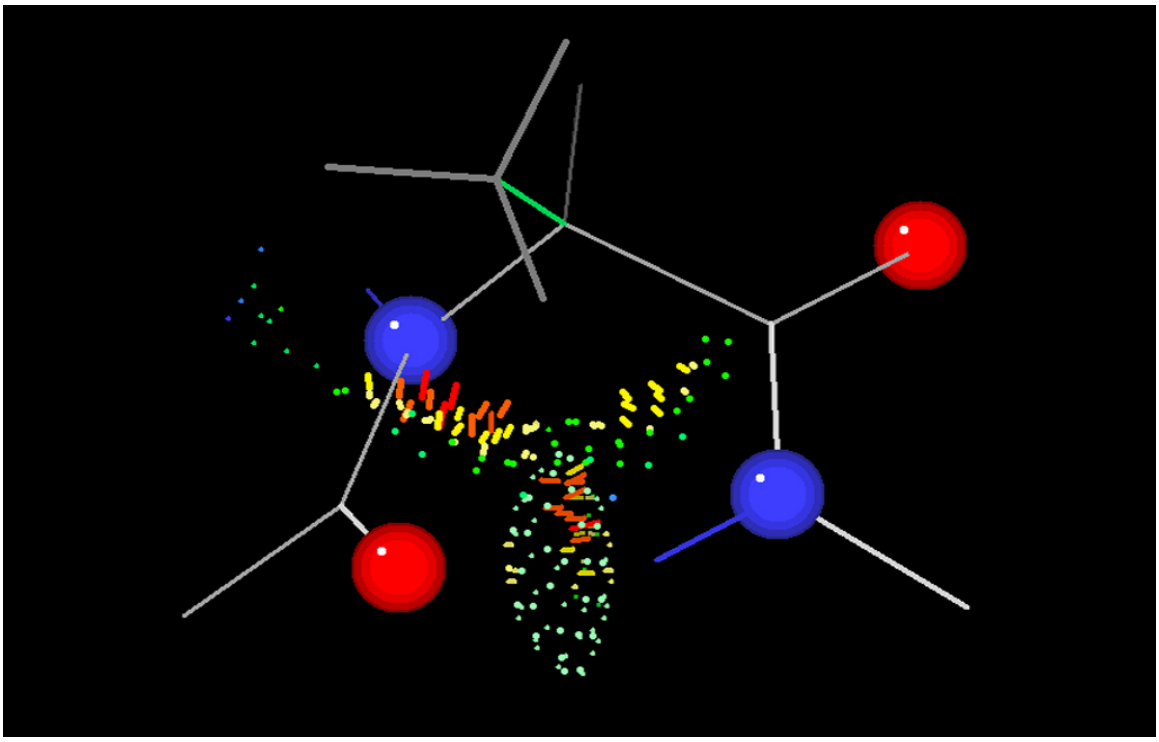

FIGURE 5.2: The γ turn, a rare but possible motif. This example is built with ideal bond lengths and angles (Engh and Huber, 2001) with ϕ,ψ near +75°,-50°. Based on our quality-filtered distributions, in many cases the stability gain from the i-1 to i+1 mainchain H-bond (green pillow) outweighs the stability loss from the overly tight sidechain-mainchain packing (red dots/spikes) – or from the bond-angle adjustments that may occur to relieve these repulsive forces in real examples without ideal geometry – resulting in a low but measurable occurrence frequency.

as Ile/Val do with local backbone. The other six categories have quite unique outer contours (Figure 5.3) and therefore indeed merit separate treatment.

I subsequently compiled versions of these final six Ramachandran plots using our currently state-of-the-art data set, the Top8000 (Section 5.2). To achieve smooth probability distributions with sharply defined boundaries between favorable and forbidden areas, I applied kernel density estimation in two passes to both achieve smooth contours in sparse regions and tightly hug the steep cliff next to helix (Lovell et al., 2003). In the first step of this method, a cosine of fixed width α_0 is placed over each data point, and the value at each desired grid point is determined by sum-

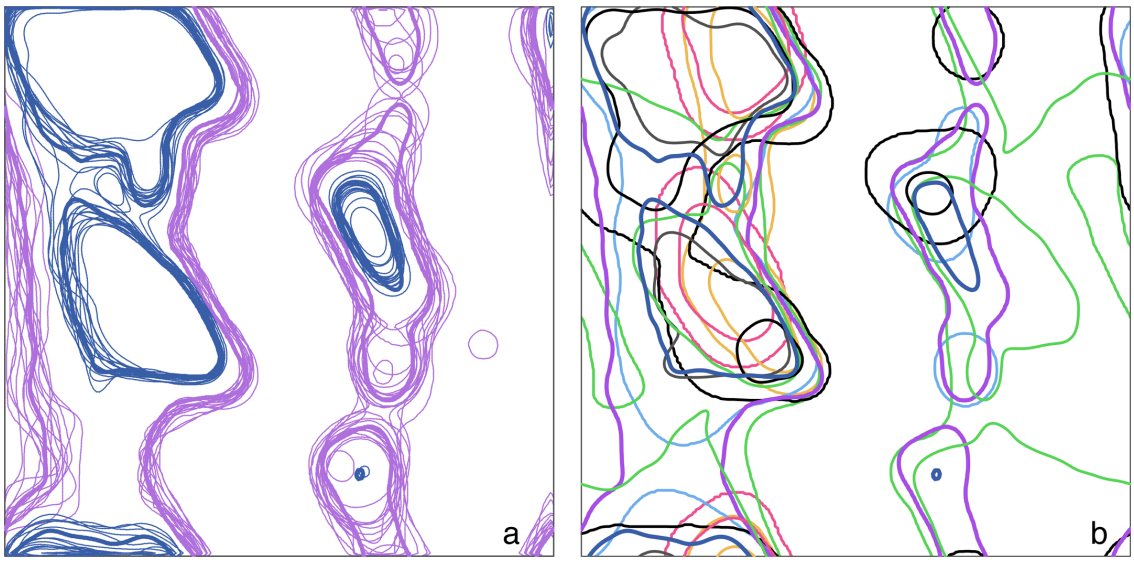

FIGURE 5.3: Outer contours of Top4400-derived Ramachandran plots for specific amino acid categories; in both panels, the general-case contours are shown as wider lines (dark blue and purple). (a) Overlapped contours for each of the 16 amino acid types that are included in the "general" distribution because they match quite well; 98% contours are in dark blue, 99.95% contours in purple. (b) Overlapped contours for the 6 categories recommended by the VTF (Gly in green, *trans* Pro in gold, *cis* Pro in red, pre-Pro in black, Ile/Val in cyan, and general in wider dark blue and purple), proposed for separate evaluation because they are each very different. Made for (Read et al., 2011).

ming all data points' cosine-based contributions. The second, "density-dependent" step is similar, but now each data point receives a cosine with a unique width α_i that is inversely proportional to its density value after the first step and scaled by a constant k. Finally, each grid point is converted from a sum-of-cosines value to a percentile by computing the percentage of data points that are at lower values. In general, there is some interplay between the widths of the cosines placed on data points for the two smoothing passes and the contour levels used to define "favored" and "allowed" regions. Therefore, we experimented with various combinations, with the aim of unifying these values across the six distributions to the extent possible. We ultimately determined it was reasonable to use the same values as before: 10° for

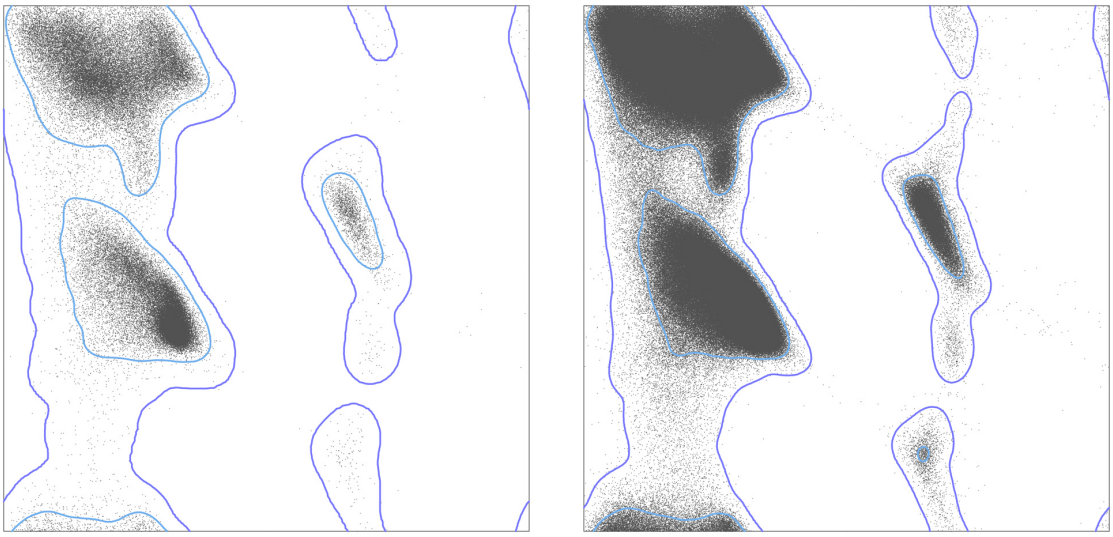

FIGURE 5.4: Top500 vs. Top8000 general-case Ramachandran plots. Left: Top500 distribution based on 81,234 residues with backbone B-factor < 30, excluding Gly, Pro, and pre-Pro. Right: Top8000 distribution based on 1,061,639 residues with backbone B-factor < 30, excluding Gly, Pro, pre-Pro, Ile, and Val (see main text). Smoothed contours encompassing 98% (light blue) and 99.95% (dark blue) of the data delineate "favored" and "allowed" regions, respectively.

the first-pass cosine width α_0, and 13° (general case) or 16° (other subsets) for the second-pass k parameter (Lovell et al., 2003). We also maintained the 98% contour (i.e. the area containing 98% of the data) definition for "favored" in all six distributions, the 99.95% contour definition for "allowed" in the updated general-case distribution, and the 99.9% contour definition for "allowed" in four of our five other distributions. However, we switched to a 99.8% contour definition for "allowed" in the *cis* Pro distribution; the resulting contour is tolerably smooth and still tight to the body of the data despite the relatively small amount of data.

The resulting plots reflect over an order of magnitude larger data set than was originally used in MolProbity (Figure 5.4). Despite this massive increase in data, for those distributions which are essentially updates to predecessor distributions (general case, pre-Pro, Gly, to some extent *trans* Pro), the outer contours are merely refined,

not radically altered (Figure 5.5). For example, an allowed shoal for $+\phi$ and $\psi \approx +110^\circ$ and a favored peak for $+\phi$ and $\psi \approx$ -130° arise for the general case, a new allowed peak near +75°, +160° appears for pre-Pro, and allowed symmetric plateaus near $\phi = 180^\circ$ are significantly extended for Gly (Figure 5.5).

On the other hand, the distributions for Ile/Val and *cis* Pro are radically different from the distributions with which those residues would have previously been evaluated: the Ile/Val outer contours are significantly less permissive than the old general case, and the *cis* Pro contours are shifted "up and to the left" and severed into two patches with a disallowed ψ region in between (Figures 5.5 and 5.6).

There is new detail and strengthened evidence for clusters of allowed but disfavored conformations. The two allowed regions in the lower right quadrant on the general-case plot are now clearly visible data point clusters. The upper one consists of γ-turn residues stabilized by a backbone H-bond, entirely disallowed by the older ProCheck (Laskowski et al., 1993) criteria, previously argued as allowed by members of our lab (Lovell et al., 2003), and now quite clearly populated in the high-quality data. The lower peak is a conformation necessary to form Type II β turns. It is most favorable for Gly (Figure 5.5); impossible for Pro, pre-Pro, Ile, and Val; and allowed but disfavored for the general case, most frequently seen for Asp.

The final set of plots can be seen in Figure 5.7. The six new distributions and scores are now implemented in both MolProbity and PHENIX, and will show up very soon on the web and in nightly builds, providing the most up-to-date Ramachandran analysis possible to crystallographers and spectroscopists.

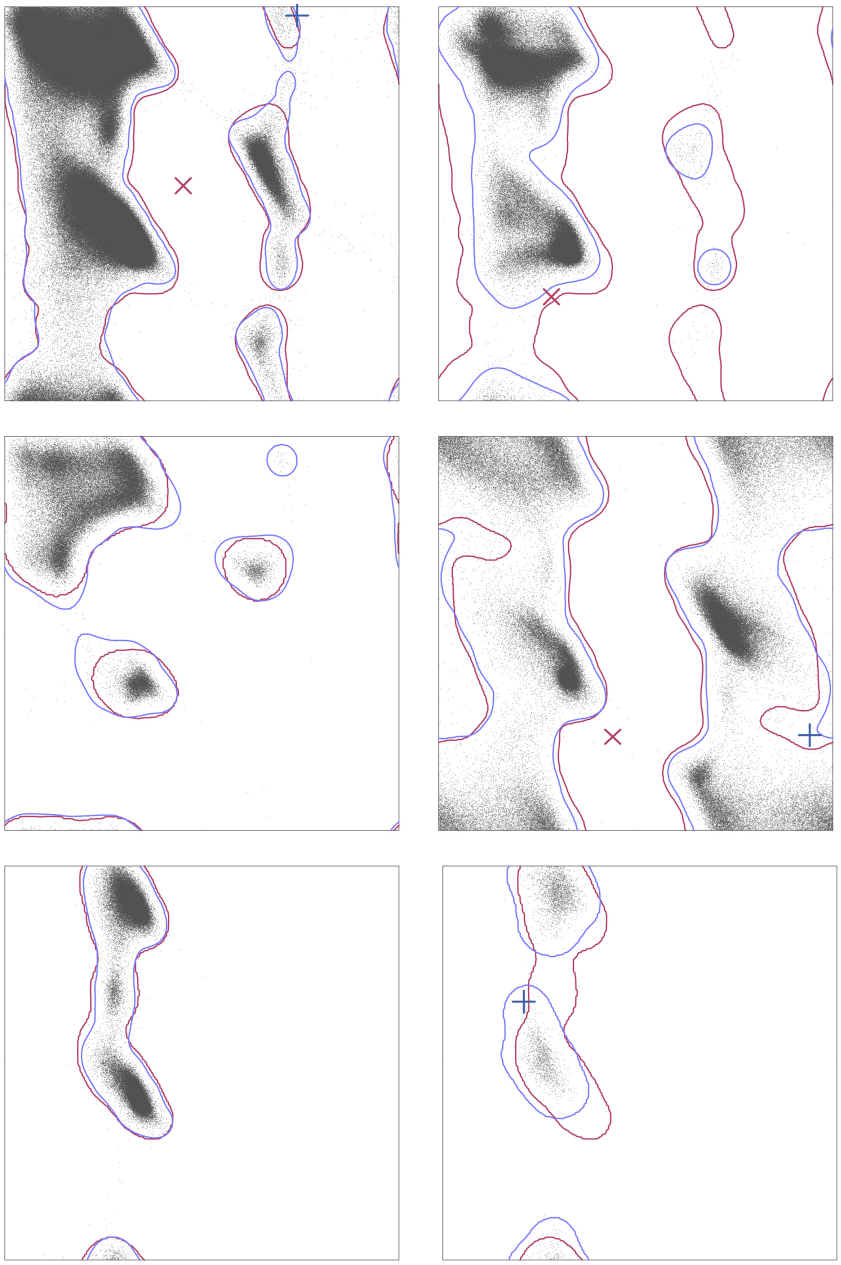

FIGURE 5.5: Top500 vs. Top8000 Ramachandran plots for all six Top8000 categories. More data from the Top8000 (small gray dots) and additional separate distributions lead to a variety of changes in the Top8000 outer (i.e. "allowed") contours (blue) relative to corresponding Top500 contours (red). Top left: general case. Top right: Ile/Val (vs. general case). Middle left: pre-Pro. Middle right: Gly. Bottom left: *trans* Pro (vs. all Pro). Bottom right: *cis* Pro (vs. all Pro). Also shown are selected residues which were previously outliers and now allowed (blue plusses) (Figures 5.8, 5.9, 5.10) and which are now (or still) outliers (red crosses) (Figures 5.11, 5.12, 5.13).

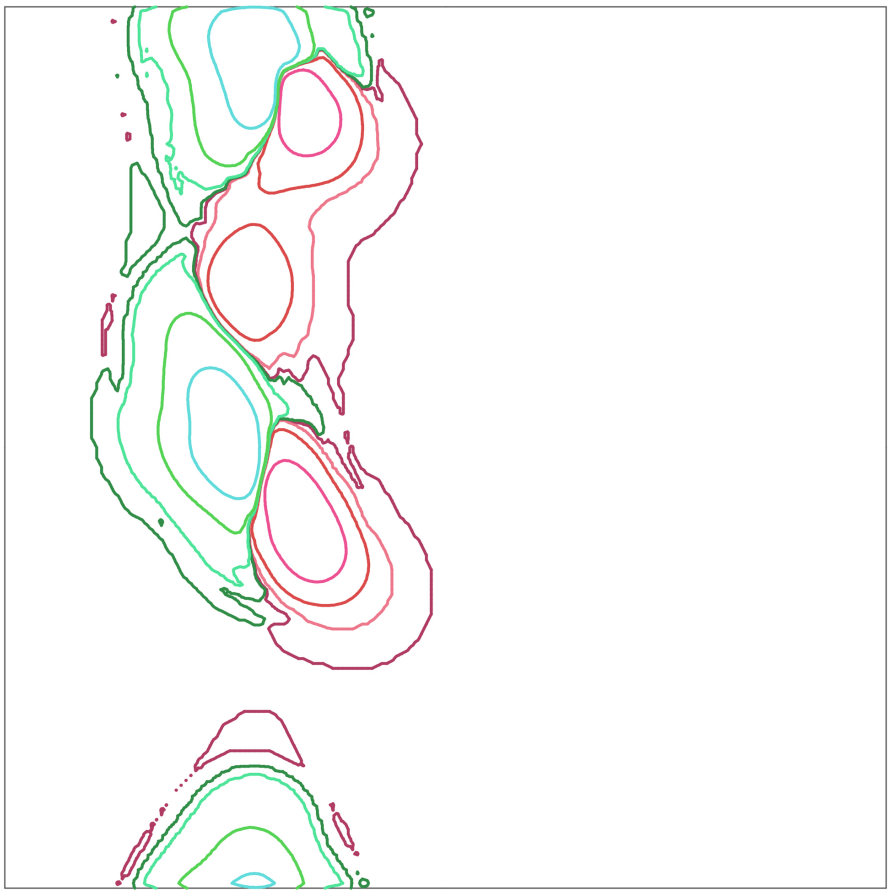

FIGURE 5.6: Ramachandran difference plot: Top8000 distribution with just *cis* Pro minus Top500 distribution with all Pro. The new separate *cis* Pro distribution is enhanced "up and to the left" (positive values, green) and depleted for significant portions of the old distribution (negative values, red). Contour levels range from 0.0001 (darkest) to 0.1 (lightest), where 1.0 is the maximum (percentile) value of the original distributions before subtraction.

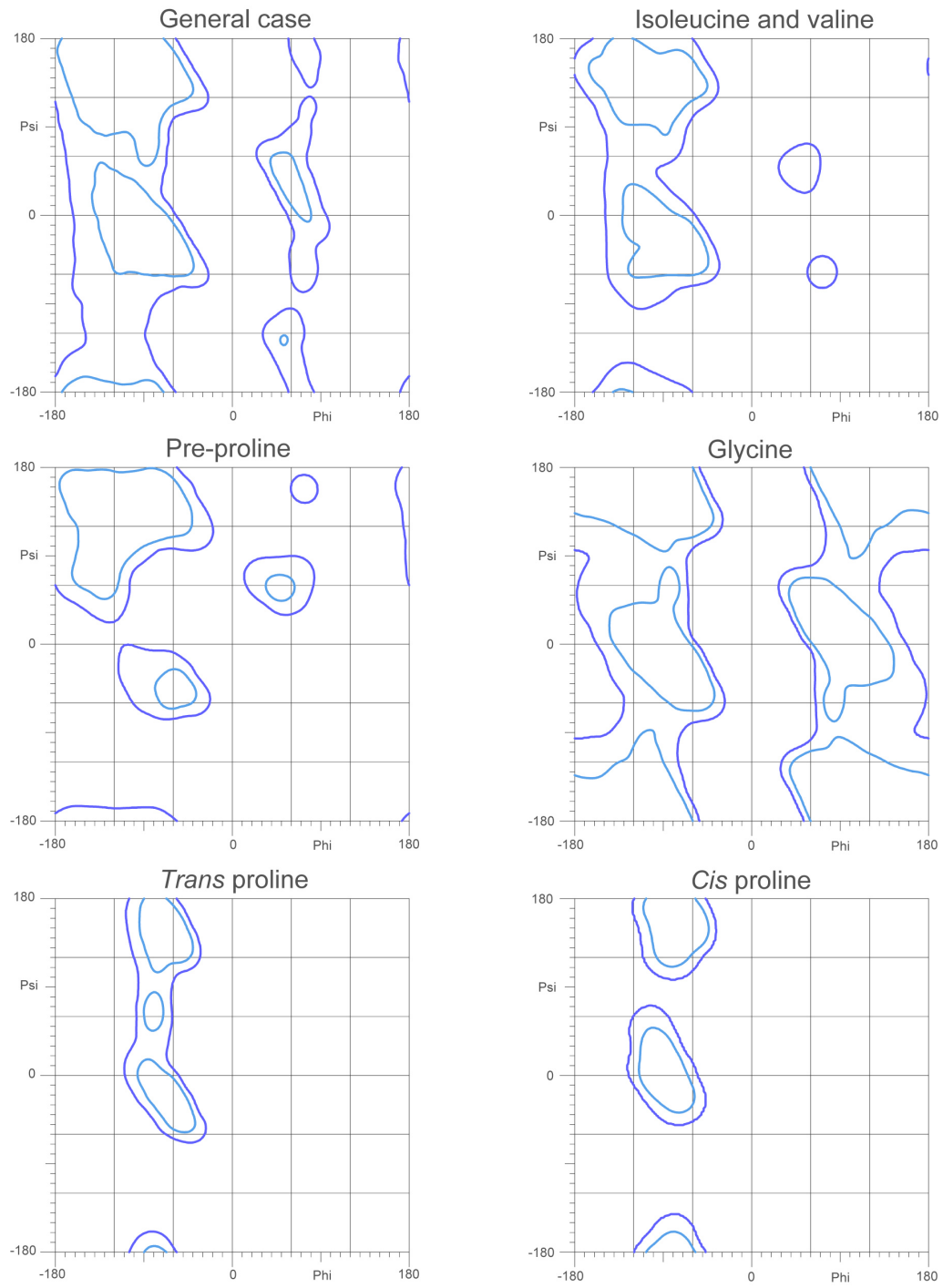

FIGURE 5.7: Final set of six Top8000 Ramachandran categories. "Favored" regions are delineated by smoothed contours encompassing 98% of the data (light blue). "Allowed" regions are delineated by smoothed contours encompassing 99.95% (general case), 99.9% (Ile/Val, pre-Pro, Gly, and *trans* Pro), or 99.8% (*cis* Pro) of the data (dark blue).

Interesting near-outliers with genuine conformations

In addition to their utility for validation purposes, these distributions invite further investigation because of their rich information content on protein backbone energetics. Outlier and barely allowed residues are of particular interest. For many such cases, unique but significantly stabilizing interactions presumably compensate for whatever repulsive forces cause the local ϕ,ψ region to be weakly populated.

For example, 1ka1 Ser264 (Figure 5.8) has ϕ,ψ +85°,+171°, which was enough to barely make it an outlier using the Top500 distributions. This conformation was previously validated as strained but real on the merits of the residue's low B-factors, good electron density, lack of clashes, and four good H-bonds, including one to an active-site phosphate (Lovell et al., 2003). Now, with over an order of magnitude more data and stricter quality filtering, a "shoal" has extended into this region of the Ramachandran plot, nearly connecting with the Lα region (Figure 5.5); with this new distribution, 1ka1 Ser264 is (more correctly) reported as allowed, albeit not favored.

Similarly, 1a88 Pro31 is a *cis* Pro with ϕ,ψ -106°,+56°, having a well-modeled conformation corroborated by good electron density and flanking hydrogen bonds (Figure 5.9). However, it is just outside the allowed region and is thus classified an outlier by the Top500 Pro distribution, which indiscriminately lumps *cis* and *trans* Pro together. With the new separate Top8000 *cis* Pro distribution, this strained but genuine conformation is labeled allowed. Interestingly, this region is highly conserved, and Pro31 in particular, by virtue of its proximity to the catalytic triad (Figure 5.9), is implicated in the chloroperoxidase enzyme's reaction mechanism (Hofmann et al., 1998). In general, a higher frequency of strained but genuine conformations are observed at active or functional sites, which emphasizes the importance of our more finely tuned torsional distributions with better delineated margins.

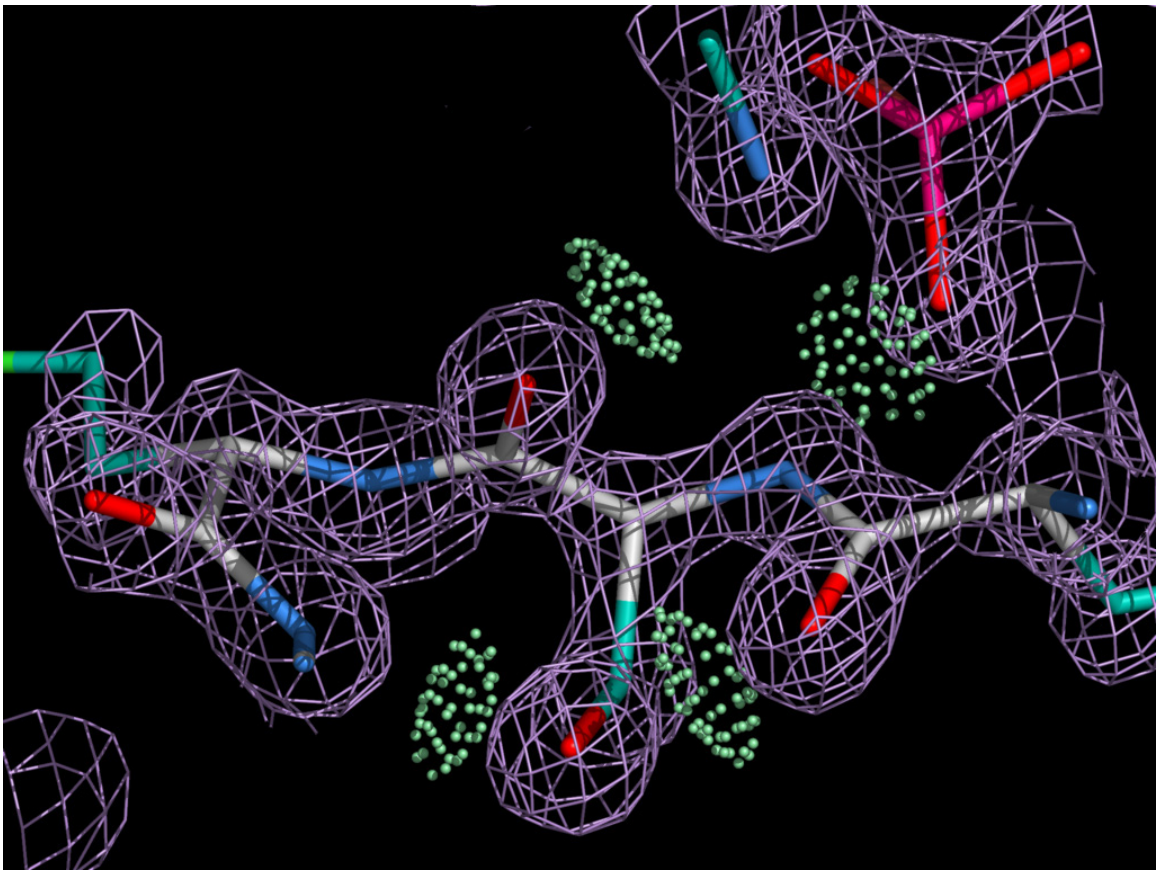

FIGURE 5.8: 1ka1 Ser264 with ϕ,ψ +85°,+171° (Figure 5.5) was considered a Ramachandran outlier using the old Top500 distribution, but was manually validated due to extenuating interactions. It is now considered allowed using the new Top8000 distribution, thereby confirming the previous manual analysis. Taken from (Lovell et al., 2003).

Another such case is 1ftr Gly150 (Figure 5.10) with ϕ,ψ +159°,-93°, which was previously labeled an outlier but is now classified as allowed based on altered contours near ϕ 180° (Figure 5.5). This conformation is genuine because its slightly unusual backbone torsions are compensated by its two mainchain-mainchain H-bonds and two more flanking mainchain H-bonds to waters.

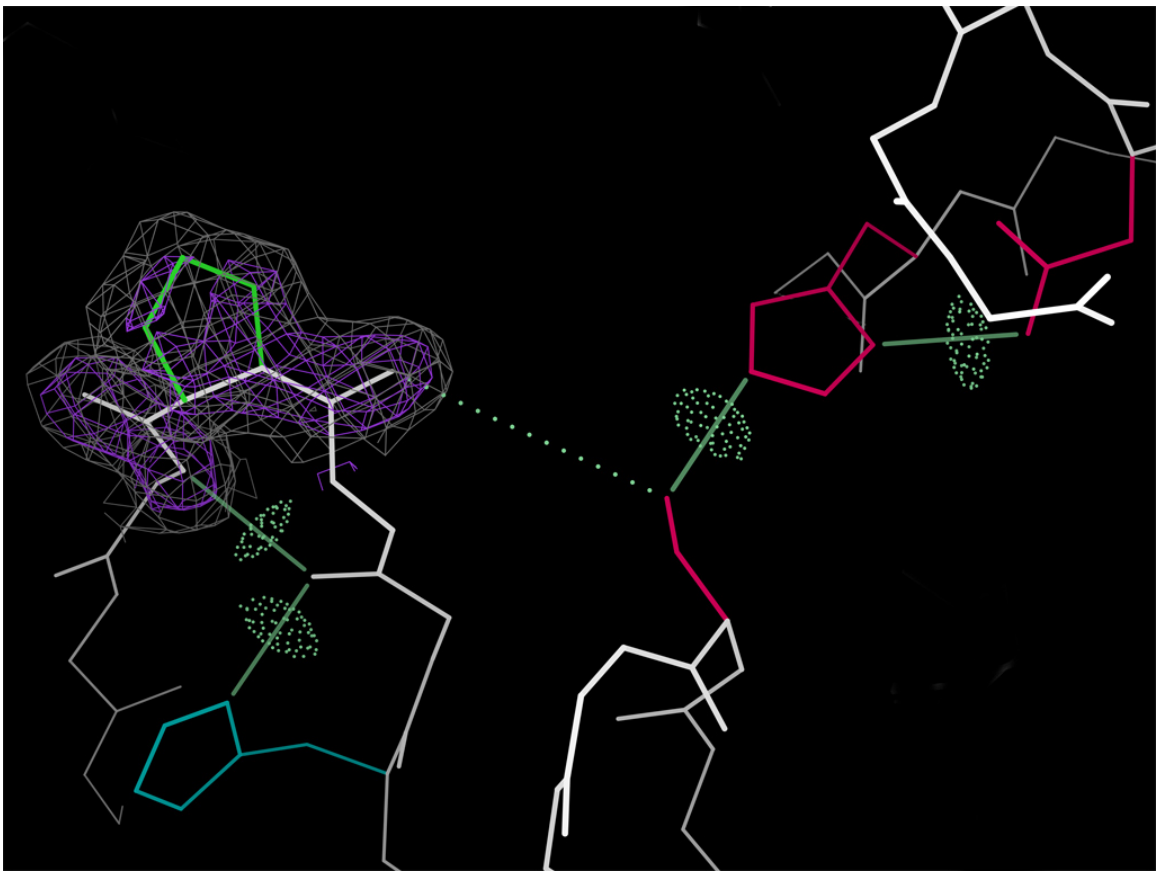

FIGURE 5.9: 1a88 *cis* Pro31 (green) with ϕ,ψ -106°,+56° (Figure 5.5) was a Ramachandran outlier based on the single Pro Top500 distribution, but is now considered allowed based on the separate *cis* Pro Top8000 distribution. Its clear electron density contoured at 1.2 σ (gray mesh) and 3.0 σ (purple mesh) and flanking main-chain H-bonds validate its conformation. Furthermore, this residue's location at the enzyme's active site adjacent to a catalytic triad (pink), with a putative H-bond thought to be important at a later step in the reaction mechanism illustrated (dotted line), helps explain its unusual but genuine conformation.

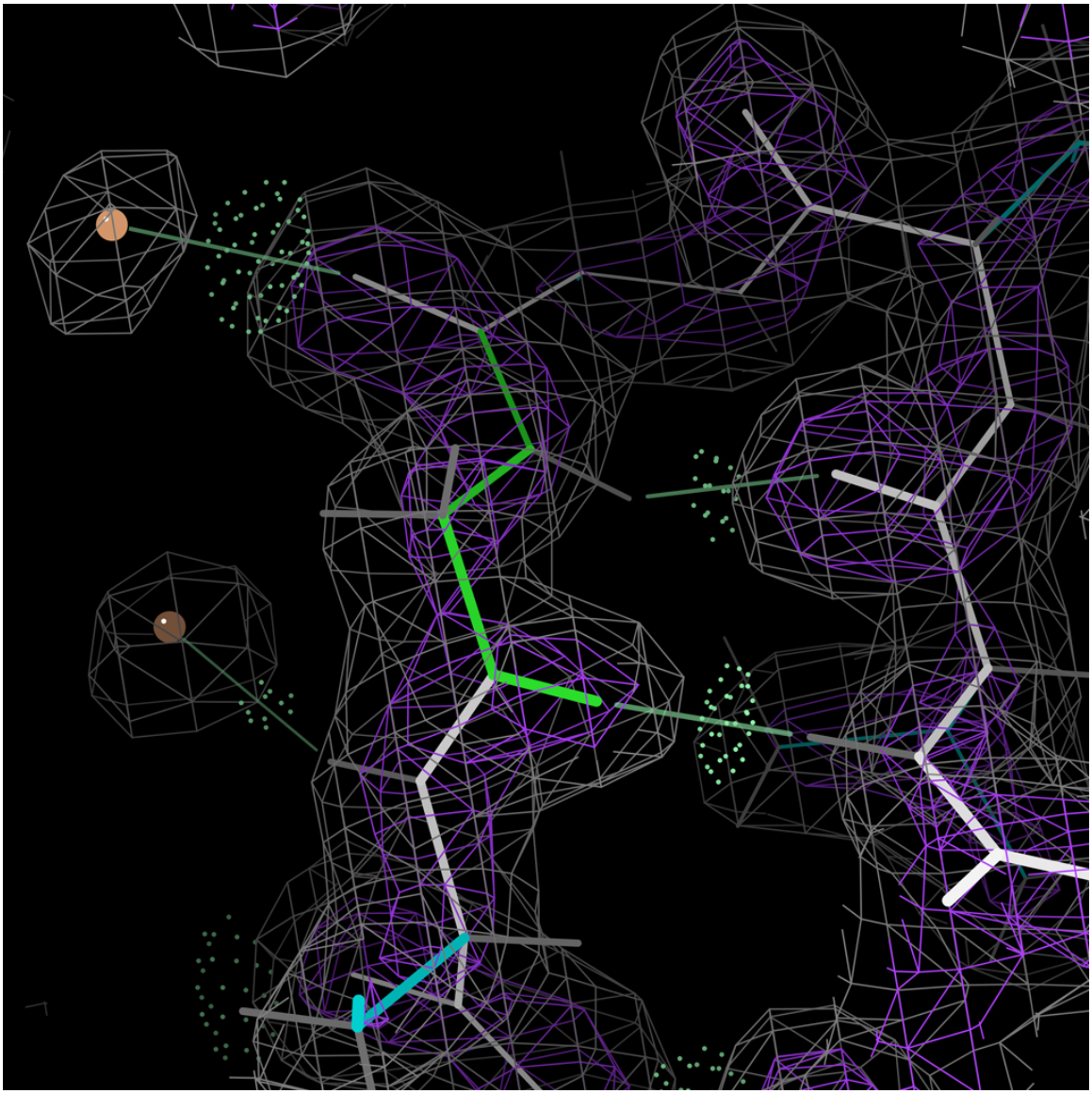

FIGURE 5.10: 1ftr Gly150 (green) with ϕ,ψ +159°,-93° (Figure 5.5) was a Ramachandran outlier based on the Gly Top500 distribution, but is now considered allowed based on the Top8000 distribution. Clear electron density contoured at 1.2 σ (gray mesh) and 3.0 σ (purple mesh) and flanking mainchain H-bonds to both mainchain and ordered waters validate its conformation.

Marginal outliers with fitting errors

The new Top8000 distributions aren't universally more permissive, however: although some residues were formerly outliers and are now allowed, even more were formerly allowed and are now outliers.

An example of the latter is 1bu8 Val246 with ϕ,ψ -77°,-85°, just below the α-helix region (Figure 5.11). It was previously barely allowed, partly because there was less data to precisely define the outlier/allowed border, but more importantly because Ile and Val were previously lumped into the general-case distribution. Now that a separate distribution for these hydrophobic branched-β sidechains is feasible, we can identify 1bu8 Val246 as an outlier rather than allowed (Figure 5.5). Indeed, all orthogonal information, from fit to density to various MolProbity markups, confirms this assertion (Figure 5.11).

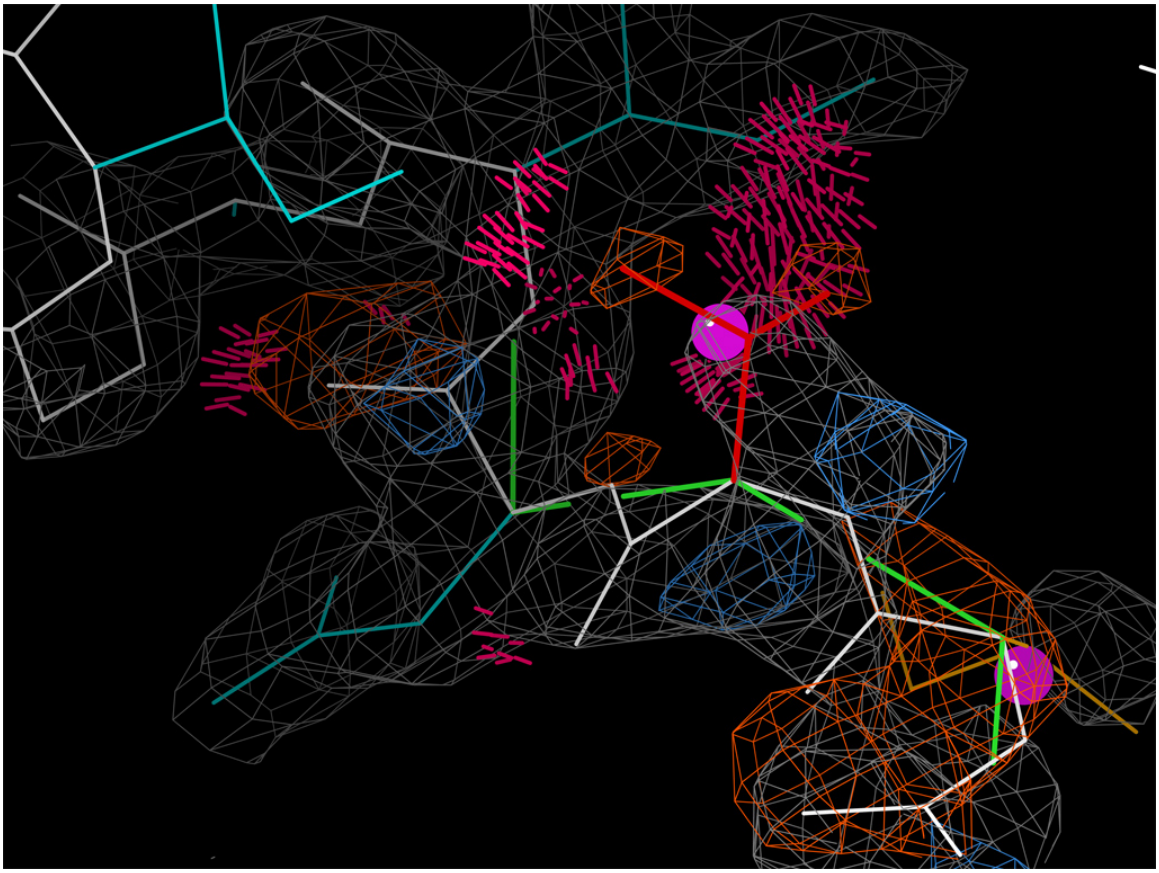

FIGURE 5.11: 1bu8 Val246 (red) with ϕ,ψ just below the α-helix region (Figure 5.5) was allowed based on the general-case Top500 Ramachandran distribution, but is now considered an outlier based on the separate Ile/Val Top8000 distribution. Numerous quality criteria corroborate the new assignment by confirming this residue is badly misfit: massive steric clashes (pink spikes), three consecutive Ramachandran outliers based on the new distributions (green lines), a rotamer outlier at the preceding residue (orange), bad Cβ deviations (pink balls), poor fit to the 2Fo-Fc electron density contoured at 1.2 σ (gray mesh) and 3.0 σ (purple mesh), and significant peaks in the Fo-Fc electron density contoured at +4.0 σ (blue mesh) and -4.0 σ (orange mesh).

Prominent outliers with fitting errors or non-standard chemistry

Other outliers are so blatant that the particular choice of allowed vs. favored contours is essentially irrelevant. One such case is 1j58 Gly308 (Figure 5.12) in the absolutely forbidden region near $\phi = 0^\circ$ (ϕ,ψ -21°,-95°). It lies amidst a hideously modeled region with steric clashes and a Cβ deviation (despite low B-factors). The electron density further contradicts the model, suggesting large (approximately 120°) peptide rotations to move the carbonyl O into density. Thus the clashes and distorted geometry are not compensated by some other force; the residue is simply misfit in the crystal structure.

Another is 1y2m Ser212, with ϕ,ψ near 0°,0°. This residue actually has a covalent modification that is unflagged in the PDB file (Figure 5.13). This rare feature alters the local chemistry, resulting in deviant geometry (from the perspective of a normal polypeptide) as detected by several MolProbity measures. However, the model fits the experimental electron density well, and is almost certainly valid (Figure 5.13). It is therefore not particularly useful for learning the subtle tradeoffs involved in protein backbone energetics.

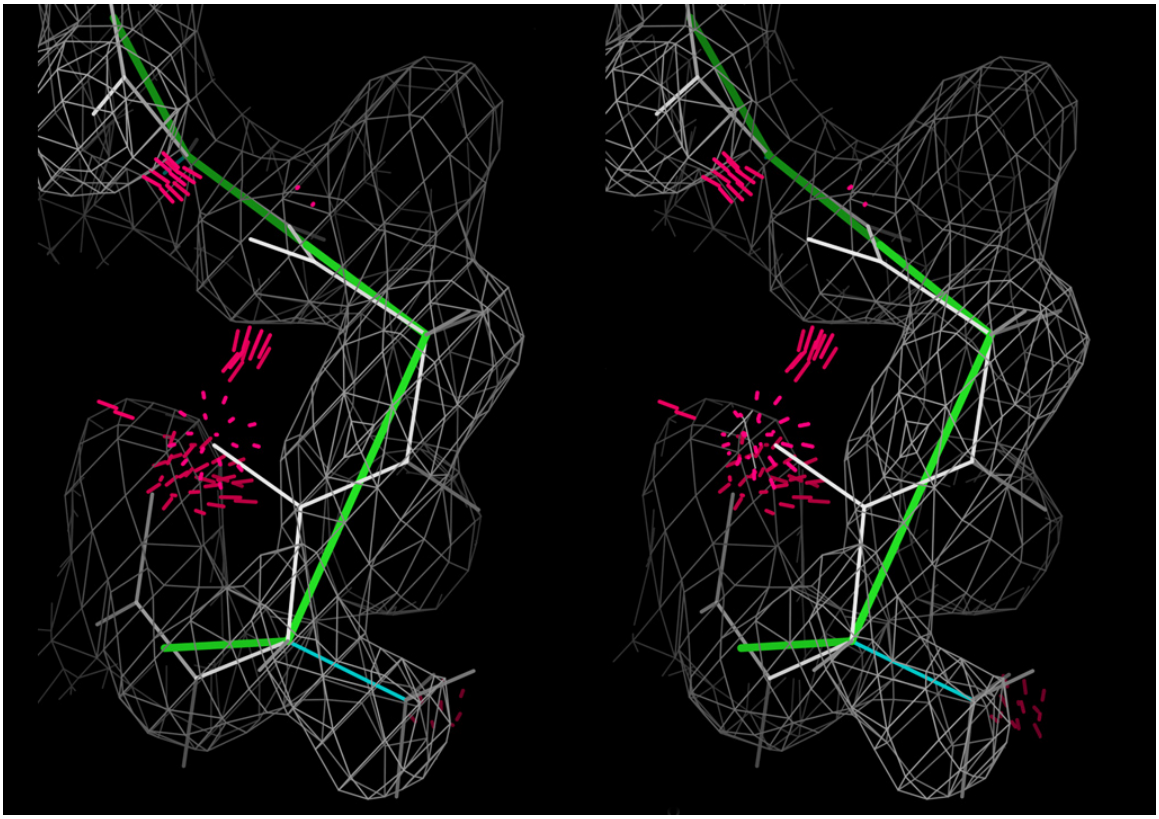

FIGURE 5.12: 1j58 Gly308 with ϕ,ψ -21°,-95° (Figure 5.5) is part of a series of Top8000 Ramachandran outliers (green lines). All orthogonal information – steric clashes (pink spikes) and poor fit to the 1.2 σ electron density (gray mesh) – corroborates its status as an error rather than a rare but genuine conformation. Stereo image.

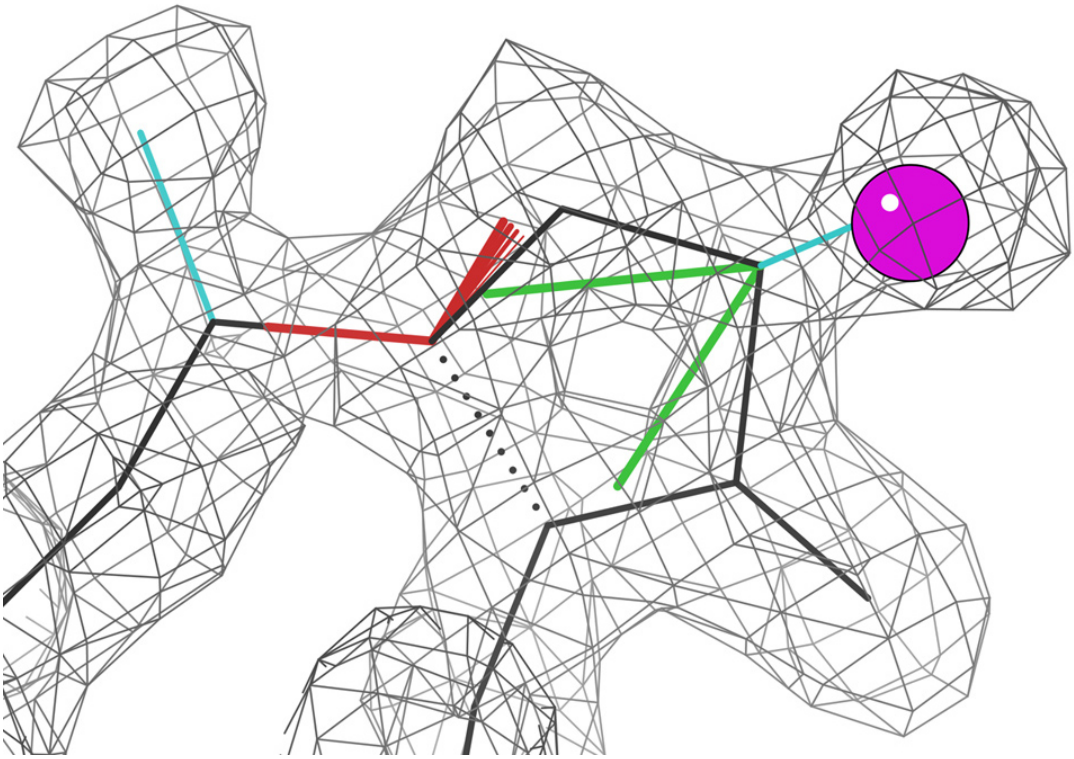

FIGURE 5.13: 1y2m Ser212 with ϕ,ψ near 0°,0° (Figure 5.5) is a Top8000 Ramachandran outlier (green lines). A widened bond angle (red fan) and a Cβ deviation (pink ball) flag it as suspicious. However, the 1.2 σ electron density (gray mesh) shows that the modeled conformation fits the experimental data quite well. In fact, this residue is so unusual not because it is fit erroneously, but because it has a covalent modification that forms a local ring-like structure (dotted line), which is why it deviates from normal protein geometry.

5.4 Future: Updating sidechain torsional distributions

The Top8000 has already enhanced our ability to discriminate realistic from dubious backbone conformations by using updated torsional distributions. In the near future, we wish to expand that analysis to sidechains.

The Richardson lab has a history of carefully describing sidechain conformational preferences. Perhaps the biggest landmark was the "penultimate" rotamer library, which reported preferred values for each discrete rotameric well and promulgated the practice of filtering out residues with high B-factors (Lovell et al., 2000). A few years later, a similar procedure to that described above for making Ramachandran distributions was used to craft smooth, multidimensional χ-dihedral-angle distributions (Lovell et al., 2003) for use in MolProbity.

Yet these studies were conducted with small data sets consisting of 240 and 500 structures, respectively. In contrast to backbone, sidechains have significantly different chemistry for essentially every amino acid type, so one must invoke more subsets – roughly 20 instead of just 6 (Section 5.3) – to achieve accurate sidechain torsional validation. The greater than order of magnitude gain in data quantity afforded by the Top8000, then, will likely be instrumental for improving our ability to discriminate favored or allowed conformations from outliers.

We have several interesting rotamer analyses planned using the Top8000. For example, we may filter out residues with poor real-space correlation to local electron density (Shapovalov and Dunbrack Jr, 2007). Using the resulting updated distributions, we plan to identify new "decoy" rotamers (Lovell et al., 2000). One productive tactic may be to find rotamer pairs that mimic the reversed Leu decoy phenomenon (Lovell et al., 2000) in that much of the sidechain overlaps closely but one atom (or rigid group of atoms) protrudes, thereby identifying decoy rotamers plus their putative correct companions. We will also investigate whether certain covalent bond

angles (e.g. Cα-Cβ-Xγ) differ significantly from expectation for specific rotamers (probably involving + χ1); even small such changes are amplified to non-negligible displacements of the sidechain end via the lever effect. As a related way to explore a sidechain end's placement possibilities, we will develop methods for sampling sidechain conformers, including not only the modal rotamers but also nearby subrotamers, that are well dispersed in Cartesian rather than simply torsional space.

Finally, in addition to the usual amino acid sidechains, the Top8000 may finally enable us to create a multi-dimensional dihedral distribution and rotamer library for disulfides. The 5 degrees of freedom and the rarity of disulfides in proteins (relative to most amino acids) has so far prevented bioinformaticians from addressing this need. In particular, the versions of the Top8000 with different degrees of homology filtering (50%, 70%, 90%, 95%) will allow us to create different distributions for different needs: one of the more loosely homology-filtered versions can be used for disulfide validation, where certain rotamers are truly overrepresented due to functional utility in certain types of folds, whereas one of the more strictly homology-filtered versions (along with imposed symmetry across the two ends) can be used for disulfide design, where structural possibility is more important than structural history.

Our work on sidechain torsions is in a preliminary stage, but by using the Top8000 and the cadre of strategies proposed above, we expect to achieve noticeable improvements in sidechain treatment in MolProbity and PHENIX in the near future.

5.5 Discussion

Strongly favored regions are labeled as such in our distributions because they are quite common across thousands of structures, presumably because they play architectural roles that ensure the protein remains stably folded. Rare but allowed conformations are therefore potentially more interesting, because they are more likely to play specific "functional" roles – indeed, they appear to be overrepresented at regions such as enzyme active sites (Lovell et al., 2003), the loci of unusual chemistry critical for biology. It is therefore important to differentiate such rare but allowed conformations from erroneous conformations that fall into (or near) outlier regions because they are modeled incorrectly.

To this end, I have created new Ramachandran distributions which are more in line with orthogonal all-atom quality criteria, as illustrated by several outlier discrimination examples. This convergence of orthogonal sources of information presumably means the new distributions are better reflective of the underlying realities of backbone-centric protein energetics. Larger amounts of more stringently filtered data have been instrumental here, as evidenced by differences in "decoy" discernment by distributions based on the older Top500 vs. the newer Top8000, thereby demonstrating the power of structural bioinformatics for improving our understanding of the fundamental determinants of protein structure. To encourage their widespread adoption, these new distributions are already implemented in MolProbity and PHENIX (Adams et al., 2010), and they have been recommended by the wwPDB VTF for official incorporation in the PDB.

6

CASP8 Assessment

6.1 CASP: the "Olympics" of structure prediction

The Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a biennial protein folding competition – the "Olympics" of structure prediction, as it were. Anyone in the world is free to form a participant group, receive sequences of proteins with temporarily secret experimentally solved structures, and submit up to five of their best structural models for each such target. CASP is made possible by crystallographers and NMR spectroscopists who graciously delay publication of their experimentally hard-earned structural information to serve the structure prediction community. On the order of 100 targets are open for competition in a typical CASP experiment. In particular, many target structures come from the high-throughput Protein Structure Initiative (PSI), a federally funded mega-project with the goal of "structural genomics", i.e. elucidating the structures of as many diverse proteins as possible.

A striking recent trend is that an increasing percentage of targets are significantly homologous to proteins of publicly known structure, and correspondingly a decreasing percentage are genuinely novel structures with previously unobserved folds. A

primary reason for this shift is that much of the "low-hanging fruit" – unique sequences encoding experimentally well-behaved, small-to-midsize proteins with novel folds – has already been harvested. PSI structures are now more finely mapping the structural space of a limited sequence space, as opposed to boldly exploring new regions of sequence space (and thus presumably also of structural space). To be sure, the CASP organizers have recently launched a "CASP Roll" experiment, with targets presented to predictors as the structures are solved instead of in biennial lumps, in order to provide more opportunities for prediction of new folds, but the usefulness of this promising approach is not yet proven.

As a result, CASP has evolved in recent years to focus more on template-based modeling (TBM), also known as homology modeling, as opposed to free modeling (FM). In these methods, an existing structure of similar sequence is used as a "template", i.e. an initial guess for the structure of the target sequence, then a refinement technique of some sort is used to predict deviations from the template based on differences between the target and template sequences. The TBM community at large employs a wide range of refinement techniques, with stochastic and deterministic search/minimization techniques, and with scoring functions ranging from statistical functions based primarily on evolutionary information to largely physics-based pseudo-energy functions. In spite of these differences in approach, one emerging commonality is that essentially all groups now use multiple templates rather than a single template – the "evolutionary" camp in particular have gotten very good at recognizing distant relationships and putting together pieces from many templates, while much physics-based prediction (e.g. Rosetta) uses many small fragments with related sequences rather than a single explicit template.

In each CASP experiment, an overall assessor or judge evaluates the success of each group, thereby providing perspective on which prediction techniques are currently state-of-the-art. Historically, success of predicted models has been based on

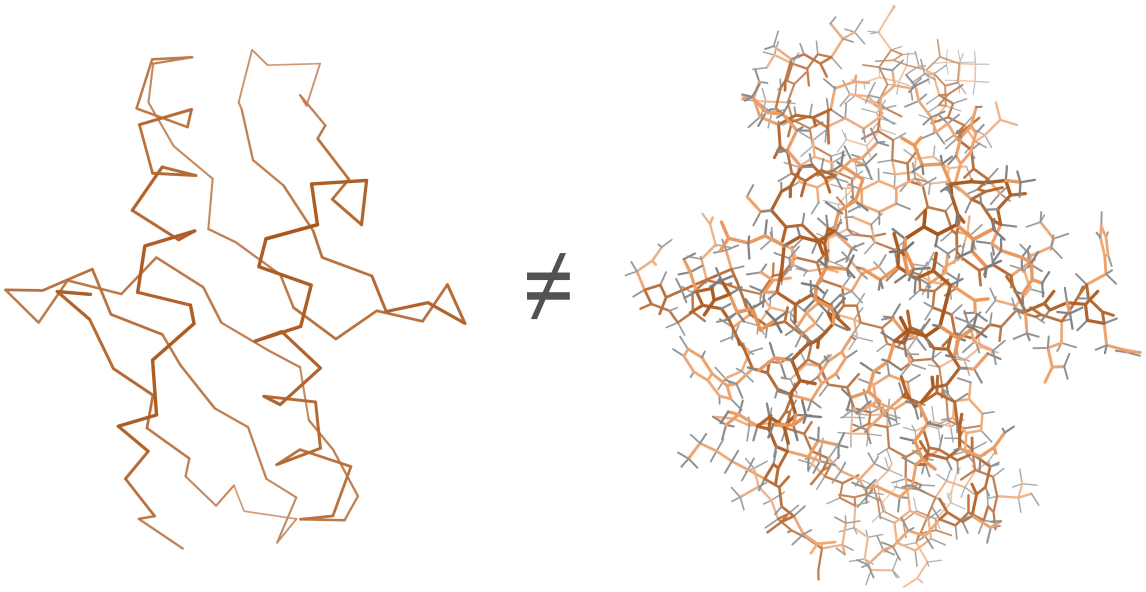

FIGURE 6.1: Cα atoms make up only about 10% of the atoms in proteins. For CASP8 assessment, we focused on the other 90% (Keedy et al., 2009). Left: Cα-only representation of T0472. Right: all-atom representation of T0472.

accuracy of Cα placement with respect to the crystal structure or properly trimmed (i.e. ignoring poorly defined regions such as floppy loops and tails (Tress et al., 2009)) NMR model. However, as it is almost trivial to note, Cαs comprise only about 10% of the atoms in a protein (Figure 6.1) – the other 90% are, for the most part, being ignored. This tradition is unfortunate, because an end user of a homology model requires a fully fleshed-out representation of a protein, not just a skeleton, in order to e.g. design a drug that binds tightly in a specific orientation.

In 2008, the Richardson lab realized that our critical perspective on structure validation, embodied in MolProbity, could be applied to predicted models as well, so that summer we served as TBM assessors for the 8th CASP experiment (CASP8) (Keedy et al., 2009). I was immediately interested in branching out from validation of experimental structures to computational models, and became intimately involved in the assessment process. As over 60,000 predicted models were gradually submitted,

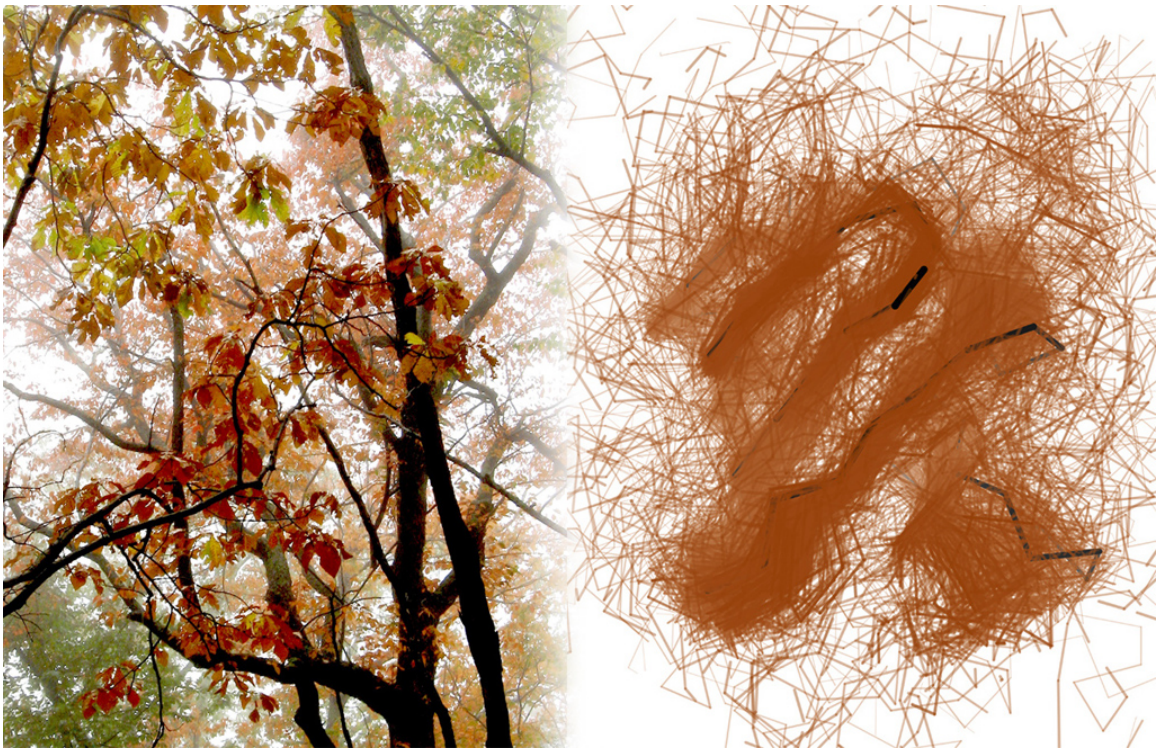

FIGURE 6.2: In attempting to sift through thousands upon thousands of predicted models submitted to CASP8, such as more than 500 (brown) for target 409 domain 1 (black) shown on the right, I began to unconsciously draw parallels to other natural phenomena, like the autumnal Great Smoky Mountains forest shown on the left.

it became quite a task to navigate the veritable forest of models (Figure 6.2) and identify the best ones, so essentially all members of our lab contributed to the process of assessing their correctness and all-atom quality. The primary goal was to incorporate all-atom quality as a component of accuracy, somewhat of a departure from previous assessments. Our lack of familiarity with the CASP status quo also gave us leeway to reevaluate several other standard practices and thereby improve assessment in fundamental ways. Finally, we packaged our model pre-processing and assessment applications into a publicly available repository, for continuing use by future predictors and (hopefully) assessors.

6.2 All-atom scores for predicted models

The traditional Cα-only score in CASP is GDT (global distance test), computed with the program LGA (local-global alignment) (Zemla, 2003). GDT is an excellent indicator of one structure's similarity to another, applicable across the entire range of difficulty for TBM targets and, to a large extent, for free modeling (FM) as well. Its power derives primarily from its use of multiple superpositions to assess both high- and low-accuracy similarity, as opposed to more quotidian metrics such as root-mean-square deviation (RMSD), which use a single superposition. Specifically, a version of GDT using relatively loose interatomic distance cutoffs of 1, 2, 4, and 8 Å called GDT-TS ("total score") has traditionally been the principal metric for correctness of predictions. However, a variant using stricter cutoffs of 0.5, 1, 2, and 4 Å called GDT-HA ("high accuracy") was used for much of the CASP7 TBM assessment because of its enhanced sensitivity to finer structural details (Kopp et al., 2007; Read and Chavali, 2007). We believe that GDT-HA probes a level of structural detail similar to that achieved by our new measures (see below), and we therefore continue to use it widely here.

To supplement GDT, we devised six new all-atom scores. The first two provide information on model-only quality, and thus uniquely can be computed by predictors before model submission. The last four provide information on model-to-target match, and thus can only be computed by the assessors (or for retrospective studies). Importantly, hydrogens were necessary for four of the six measures, and were therefore added to all models and targets with Reduce (Word et al., 1999a) prior to assessment.

MolProbity score

The first score, MPscore, is the familiar MolProbity score (Davis et al., 2007; Chen et al., 2009c; Keedy et al., 2009). It is based on a log-linear fit of three quality parameters to resolution, such that a model's MPscore is the resolution at which its individual quality parameters would be the expected values:

$$\begin{aligned} MPscore = & 0.5 + 0.42574 * log(1 + clashscore) + \\ & 0.32996 * log(1 + max(0, pctRotOut - 1)) + \\ & 0.24979 * log(1 + max(0, 100 - pctRamaFavored - 2)) \end{aligned} \tag{6.1}$$

where $clashscore$ is all-atom clashscore (Word et al., 1999a), $pctRotaOut$ is the percentage of residues that are rotamer outliers (Lovell et al., 2000), and $100 - pctRamaFavored$ is the percentage of residues that are Ramachandran outliers or allowed, i.e. not favored (Lovell et al., 2003).

Mainchain reality score

The second score, MCRS (mainchain reality score), is similar in spirit to MPscore but is skewed toward mainchain atoms:

$$MCRS = 100 - 10 * avgSpike - 5 * pctRamaOut - 2.5 * pctLengthOut - 2.5 * pctAngleOut \tag{6.2}$$

where $avgSpike$ is the per-residue average of the sum of "spike" lengths from Probe (indicating the severity of steric clashes) between pairs of mainchain atoms, $pctRamaOut$ is the percentage of residues that are Ramachandran outliers, and $pctLengthOut$ and $pctAngleOut$ are the percentages of residues with mainchain bond lengths and bond angles respectively that are outliers $> 4\ \sigma$ from ideal. MCRS initiates a model's score at 100 then subtracts points for errors, meaning that extremely poor models could end up with negative scores; we therefore placed a "floor" by truncating scores at 0.

Sidechain end positioning

The third score, GDC-sc (global distance calculation for sidechains), applies superposition-based scoring to the functional ends of protein sidechains. Instead of comparing residue positions on the basis of Cαs, GDC-sc uses a characteristic atom near the end of each sidechain type (Keedy et al., 2009) for the evaluation of residue-residue distance deviations. More concretely, the traditional GDT-TS score is a weighted sum of the fractions of residues whose Cα atoms are superimposed within limits of 1, 2, 4, and 8 Å; using the LGA backbone superposition, the GDC-sc score is instead a weighted sum of the fractions of residues whose corresponding model-target sidechain atom pairs fit under 10 distance-limit values from 0.5 Å to 5 Å (8 Å would be a displacement too large to be meaningful for a local sidechain difference).

Hydrogen bond correctness

The fourth and fifth scores, HBmc and HBsc, measure the percentage of target H-bonds recapitulated by the model. These scores are similar to the H-bond correctness score used in CASP7 (Kopp et al., 2007), but separate mainchain H-bonds (mainchain-mainchain only) from sidechain H-bonds (sidechain-sidechain or sidechain-mainchain). H-bonds were defined by Probe (Word et al., 1999b) using default parameters, although we used a slightly more lenient probe radius for model (but not target) sidechain H-bonds to reward correct atom pairings with imperfect geometry (Keedy et al., 2009).

Rotamer correctness

The sixth and final score, corRot, is analogous to HBmc and HBsc but applies to sidechain rotamers instead: it measures the percentage of target rotamers recapitulated by the model. Rotamers are assigned to high-probability bins in MolProbity's smoothed, multidimensional sidechain dihedral distributions (Lovell et al.,

2003; Chen et al., 2009c). This means we required all dihedrals to fall into the proper bins, rather than just one or a few (e.g. just $\chi 1$, or $\chi 1$ and $\chi 2$). Our method is therefore a more stringent test of functional sidechain end placement, although it makes prediction of long sidechains statistically/inherently more difficult than for short sidechains.

NMR structures are typically ensembles containing many models with different coordinates, so each residue in an NMR target does not have a single rotamer, as is the case with X-ray targets. To address this hurdle, I worked with our lab's resident NMR expert, Jeremy Block, to define reasonable NMR target rotamers. We decided to include only residues with "consensus" rotamers across the ensemble: 85, 70, 55, and 40% agreement for sidechains with one, two, three, and four χ angles, respectively. These criteria led us to use on average 46% of target sidechains, with a range from 25-65%.

We also considered requiring a minimum number of NOE distance restraints per residue, under the assumption that the presence of experimental data would correlate with more realistic sidechain conformations that could more reasonably be included in the target set. The general trend among the NMR targets was found to be that sidechains are more likely to converge to one rotamer given more restraints, but the relationship is somewhat messy/complex (Figure 6.3). The lack of a tight linear correlation may reflect the inequitable contributions of some NOEs to determining the final set of conformations in the NMR model. Differences in methodology across refinement programs can also influence the relationship between amount of experimental restraint and degree of sidechain conformer convergence – e.g. some programs may reach the same lowest-energy sidechain conformer in many independent trajectories, ignoring evidence of other slightly higher-energy yet still (lowly) populated states – but it is unlikely that effect played a role here because 15 of the 16 NMR targets were contributed by the same experimental group, the Northeast Structural

Genomics Consortium. Ultimately we elected to avoid becoming mired in these intricacies, and instead candidly took the ensembles at face value by using simple rotamer consensus to define NMR target rotamers.

I also compared prediction of full rotamers to prediction of just $\chi 1$ (Figure 6.4). The correlation was tight, suggesting that $\chi 1$ is the most important dihedral for anchoring a correct full-rotamer prediction, but the slope > 1 and finite scatter showed that some additional "assessable information" remains in the rest of the sidechain. Given this result, we ignored $\chi 1$ prediction and focused on full-rotamer prediction for official assessment.

Ian Davis originally created MPscore (before CASP8), Rob Gillespie created MCRS and I subsequently polished it into its final form, Christopher Williams created GDC-sc, Gary Kapral created HBmc and HBsc, and I created corRot.

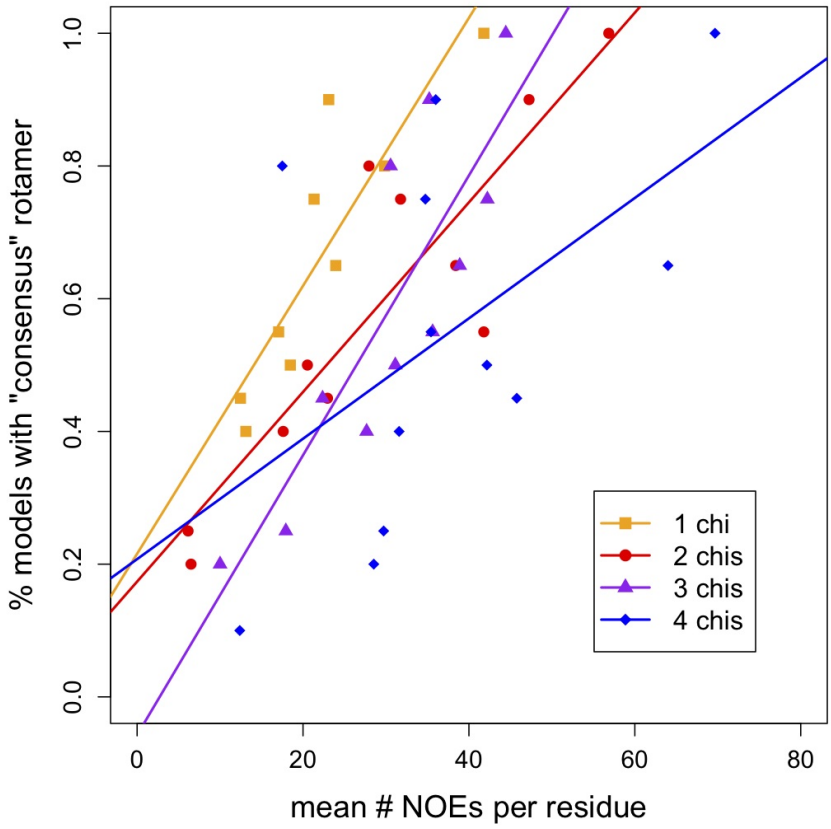

FIGURE 6.3: The number of NOEs per residue is roughly correlated with but fails to neatly predict convergence to the consensus NMR rotamer. On the y-axis is the percent of NMR-style "models" in the 16 CASP8 NMR targets that match the consensus rotamer, as defined for the corRot score. On the x-axis is the mean number of NOE restraints per residue for a given y-axis value. Data and linear fits are shown for different sidechain types with 1, 2, 3, or 4 χ angles. (Note that different numbers of residues contributed to different points on the plot because all residues with the same number of χ angles and equal convergence were conglomerated.) In all cases, although especially for longer sidechains, there is a general positive trend but significant scatter.

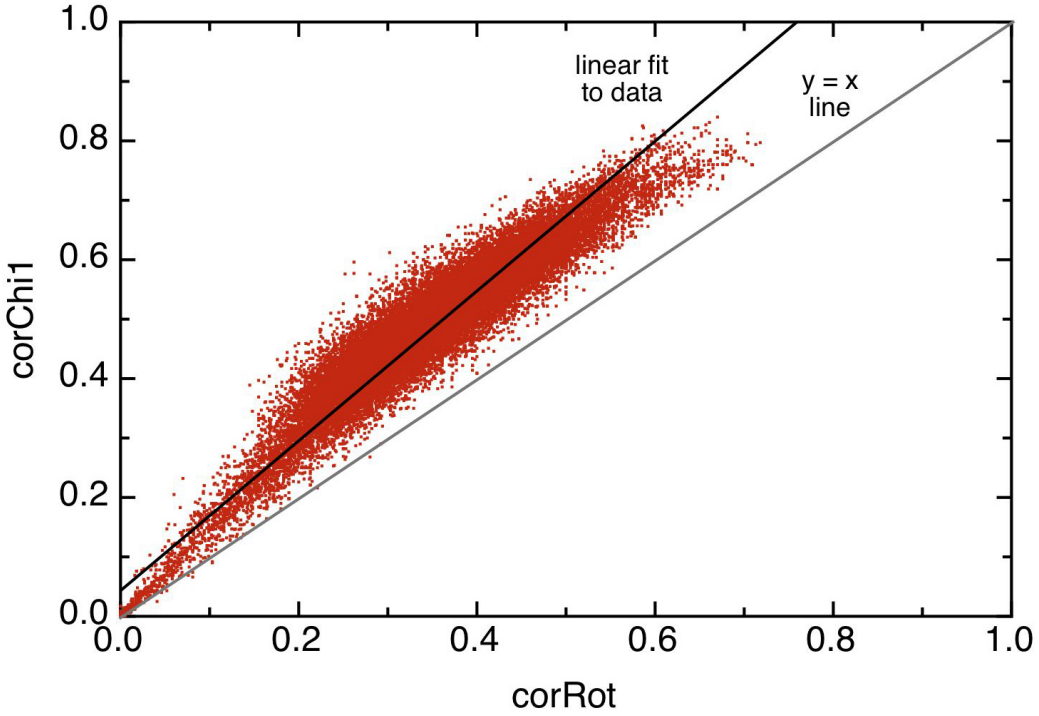

FIGURE 6.4: Prediction of just $\chi 1$ vs. full rotamers in CASP8. All models for X-ray targets are shown in red. A linear fit is shown with a black line; it lies just above the diagonal, shown with a gray line. Results were very similar for only those models with GDT-TS $\geqslant$ 50 (not shown). Single-χ sidechains were not excluded from this plot.

To learn more about the information content of these scores, we plotted them against the traditional Cα-based GDT scores on a per-model basis (Figure 6.5). As a general rule all aspects improve together, but different detailed parameters couple in different ways to get the backbone Cα atoms into roughly the right place, as evidenced by the varying levels of saturation and scatter.

MolProbity score has high scatter and relatively low slope but is linear over the entire range, and poor MCRS models exist at low GDT-HA but not at high GDT-HA. These observations suggest that modeling physically realistic mainchain may facilitate and perhaps even be essential for achieving really accurate predictions; however, this relationship needs further study.

GDC-sc has the tightest correlation to GDT-HA, presumably because it measures match of sidechain end positions between model and target, for which match of Cα positions is a prerequisite. However, it shows the most pronounced upturn at high GDT-HA, an effect detectable for most of the six plots; further investigation is needed to decipher whether this observation reflects a threshold of backbone accuracy beyond which it becomes much more feasible to achieve full-model accuracy.

The corRot score appears to capture different aspects of sidechain placement than GDC-sc, and thus seems to successfully complement GDC-sc by providing a more "local" perspective on sidechain accuracy.

The upper half of both H-bond measures shows the desirable behavior of a very strong correlation and high slope relative to GDT, but with a large spread indicative of a significant contribution from independent information.

Figure 6.6 shows a a pair of models for the same target with similar GDT-HA scores but vastly different all-atom scores. Given the two models' similarly correct Cα placement, the model with excellent all-atom quality is irrefutably superior to the one with unrealistic steric clashes and geometry outliers.

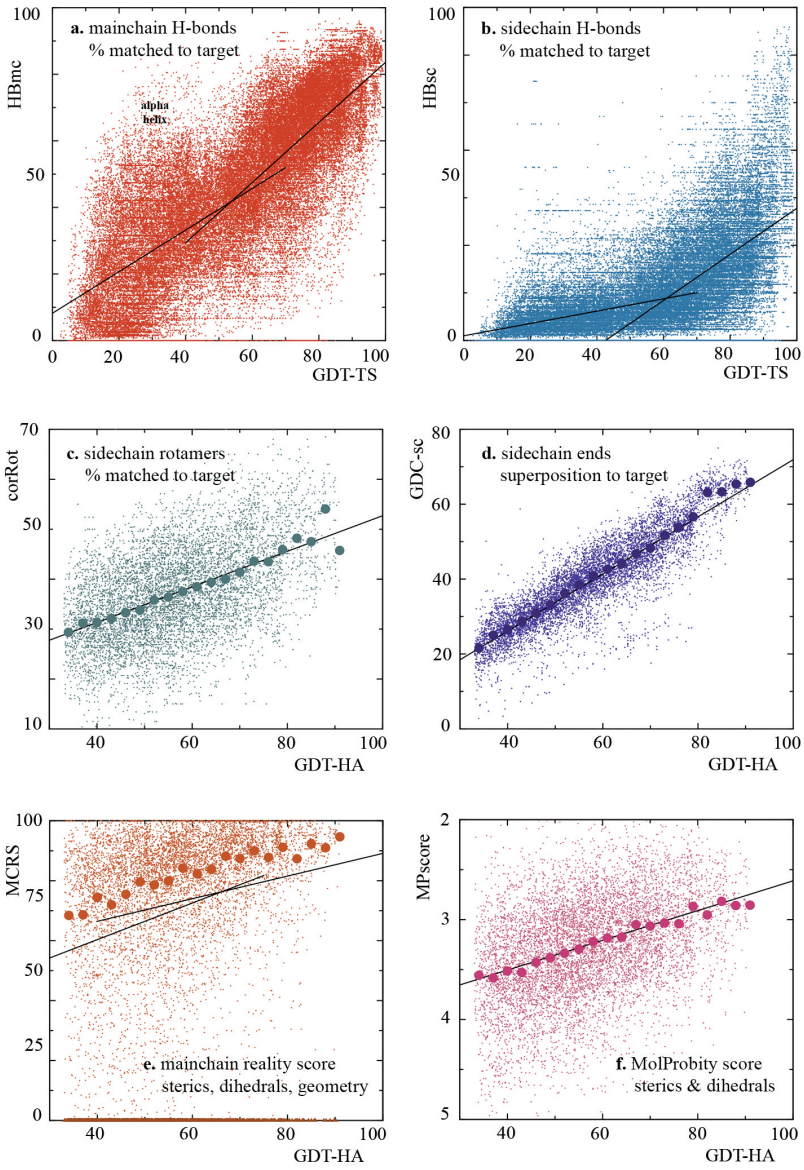

FIGURE 6.5: Distributions of new full-model scores for individual models. For all x-axes farther right is better, and for all y-axes higher is better. (a, b) All models, regardless of GDT; (c-f) only the best models with GDT-HA. Dual linear fits are on models with GDT-TS < 55 vs. $\geqslant 55$ in (a) and (b) and on models with GDT-HA < 60 vs. $\geqslant 60$ in (e); these divisions were chosen manually to highlight visible inflection points. Larger dots in (c-f) are median values for bins of 3 GDT-HA units; bins at high GDT-HA include many fewer models, producing high variability for some measures (e.g., corRot). The fit lines are well below the median points in (e), because many points lie at zero MCRS. The y-axis for MPscore in (f) has been reversed relative to other panels, because lower MPscores are better. Made for (Keedy et al., 2009).

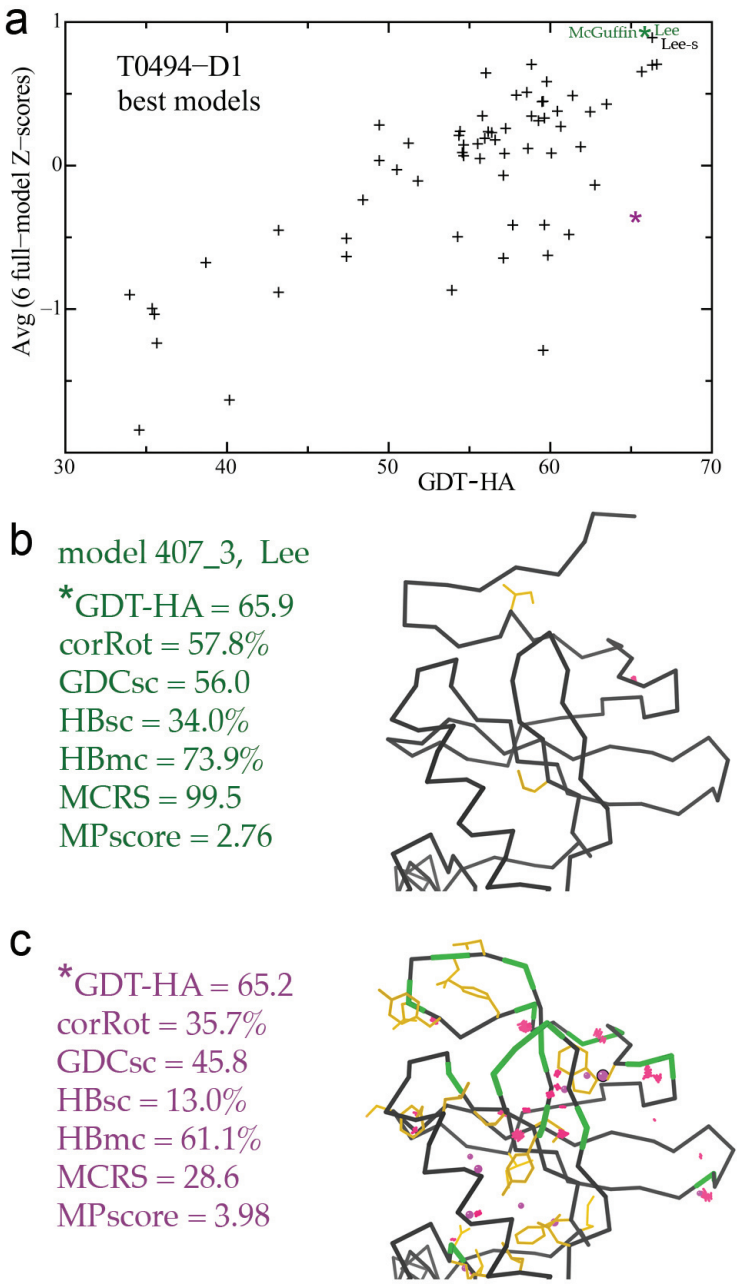

FIGURE 6.6: Differentiating models with equally good GDT scores, based on full-model performance for both physical realism and match to target. (a) Average full-model Z-score, plotted against raw GDT-HA, on individual best models for target T0494-D1 (PDB code: 2vx3, SGC, unpublished). (b) Model 407_3 (Lee) has a GDT-HA of 65.9 and the best average full-model Z-score on this target. (c) Another model with essentially the same GDT-HA (65.2) has a much lower full-model Z-score, including poorer match to target sidechains and H-bonds; the six individual scores are listed. Made for (Keedy et al., 2009).

6.3 Using all atoms to rank predictor groups

Using these versatile new metrics, we wished to cast a wide net and find not just predictor groups that excelled overall, but also those that excelled at particular aspects of homology modeling. First, though, we needed to narrow our focus to the subset of models with sufficiently accurate Cα placement for our "added value" metrics to be meaningful.

To that end, we studied the distributions of GDT scores, and found them to be strongly bimodal (Figure 6.7). This basic bimodal division also holds within most individual target domains (though there is much variability between targets in the positions and shapes of the peaks), implying that the TBM-wide bimodality is not caused by bimodality of target difficulty. Accordingly, we only considered models in the second GDT-HA peak (with GDT-HA $\geqslant$33), which have an approximately correct fold and are therefore appropriate for the more detailed, local quality assessment our new metrics provide. Consistency of "right fold" identification was assessed separately (Section 6.4).

Along similar lines, we bucked the CASP trend of assessing the model designated "model 1" by the predictors, and instead used only the best model (as judged by GDT-TS) per target. Both this choice and the GDT-HA $\geqslant$ 33 filter served to winnow the model pool to those most deserving of all-atom evaluation. Self-scoring of a group's model 1 as its actual best model was assessed separately – and found to be disappointing (Section 6.4).

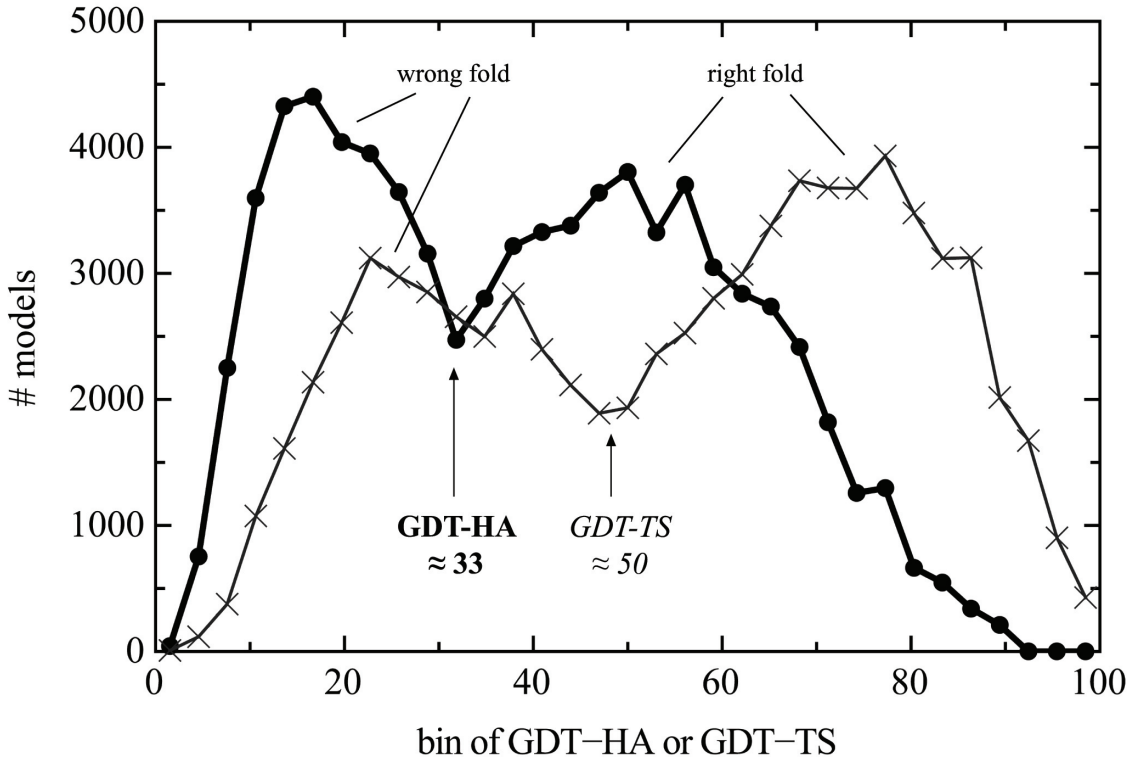

FIGURE 6.7: Bimodal distributions of GDT-HA and GDT-TS scores. All CASP8 TBM models were placed into 33 equally spaced bins, separately for GDT-HA and for GDT-TS. The division between "right fold" and "wrong fold" occurs at approximately GDT-HA of 33 (which we used for our later analysis) and GDT-TS of 50. Note that bimodal distributions were also observed within most individual targets (data not shown). Made for (Keedy et al., 2009).

At this point, all raw scores were converted to Z-scores, reflecting the number of standard deviations from the mean for each target. Predictor groups could then be ranked overall based on their performance relative to their peers, regardless of target choice. The final group scores were made available on the CASP website (`http://www.predictioncenter.org/casp8/supp_ranking.cgi`). Figure 6.8 shows the resulting "2-D ranking" plot, with average all-atom Z-score plotted against average GDT-HA Z-score. David Baker's group is the overall "winner", to the extent such a title is possible given the multifaceted nature of our assessment. However, other groups excel at certain aspects but not necessarily others. For example, Lee and Multicom earned recognition on the strength of both their sidechain modeling and their Cα placement. On the other hand, Yasara won plaudits for excellent all-atom quality, despite roughly average Cα placement (although they predicted mostly easier targets).

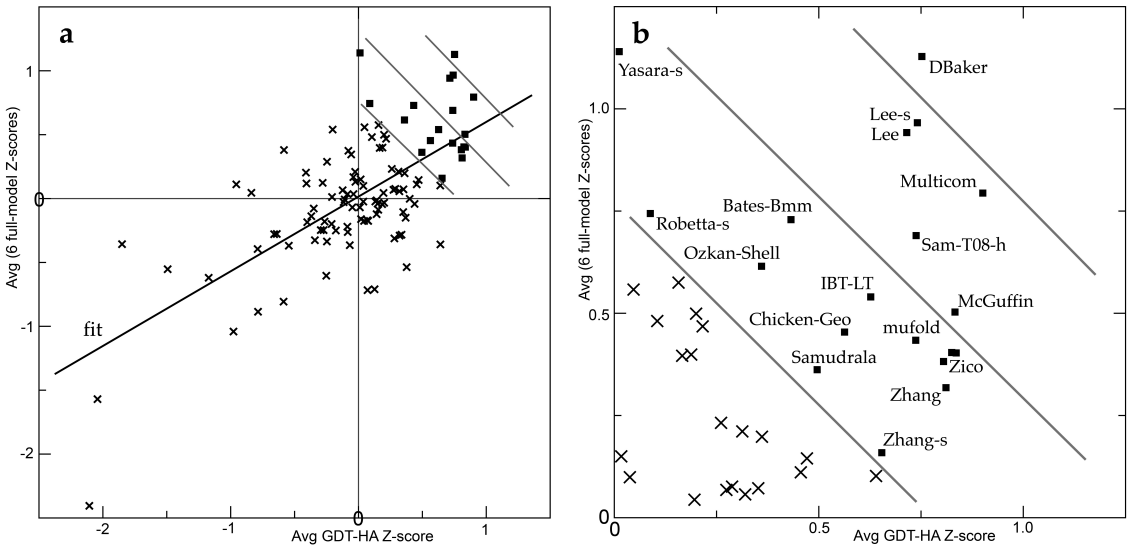

FIGURE 6.8: 2-D scoring of CASP8 predictor groups. (a) Group-average Z-score for the 6 full-model scores, plotted vs. group-average Z-score for GDT-HA. (b) Close-up of the upper-right quadrant from panel a, with the groups highlighted that did well on the combined score from both axes (emphasized by the diagonal lines). Group Z-scores are averaged over best models with GDT-HA $\geqslant 33$; groups with a qualifying model for < 20 targets are excluded. Made for (Keedy et al., 2009).

Although overall group rankings are certainly relevant for calibrating the prediction community's strengths and weaknesses, exemplary successes on particular targets are also worthy of praise. For example, Figure 6.9 shows the model with the single best individual Z-score for percent correct rotamers relative to the target where that percentage is over 60%, thereby excluding models with exceptional rotamer predictions on difficult targets where the average percent correct is actually rather low. Notice the outstanding sidechain predictions in this region of hydrophobic core, including Leu, Ile, and Val residues. Other models with high GDT but lower rotamer correctness are shown for comparison. Note that this group (DBaker) did not have the best overall rotamer correctness Z-score, but did have some outstanding models like this one.

As another example, Figure 6.10 illustrates the most dramatic cumulative GDT-TS plot, for T0460, with two individual models very much better than all others: 489_3 (DBaker; green backbone in Figure 6.10(a)) and 387_1 (Jones-UCL). The target is an NMR ensemble (2k4n), shown (black in Figure 6.10(a)) trimmed of the disordered section of a long β-hairpin loop. This is a TBM/FM target, because although there are quite a few reasonably close templates, they each differ substantially from the target for one or more of the secondary-structure elements. Only the two best models achieved a fairly close match throughout the target (GDT-TS of 63 and 54, vs. the next group at 40-44); each presumably either made an especially insightful combination among the templates or else did successful free modeling of parts not included in one or more of the better templates.

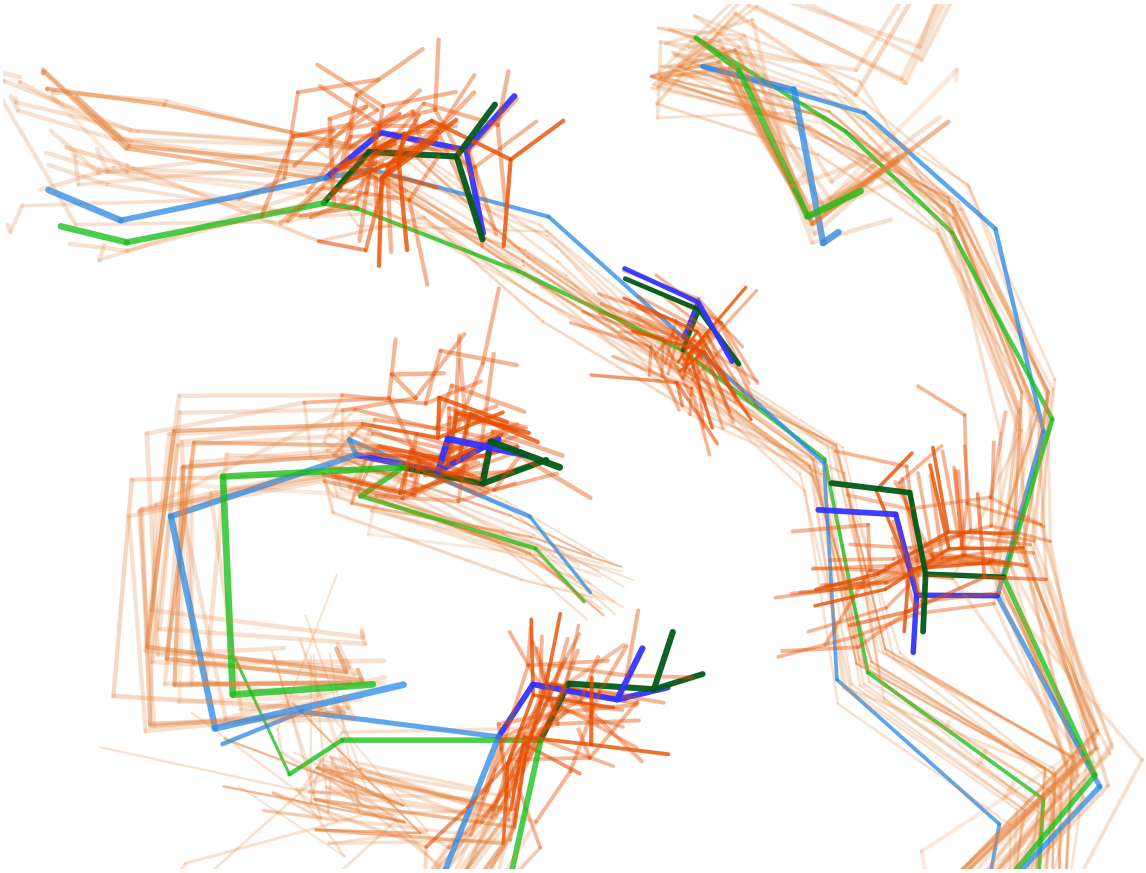

FIGURE 6.9: Individual model with outstanding rotamer correctness. Shown are Cα traces and selected core sidechains for target 492 (blue) and several models with good Cα placement (orange). Among them is model 489_1, which has the single best individual Z-score for percent correct rotamers relative to the target where that percentage is over 60% (green). Note the excellent match to the illustrated core sidechains, especially relative the competing models.

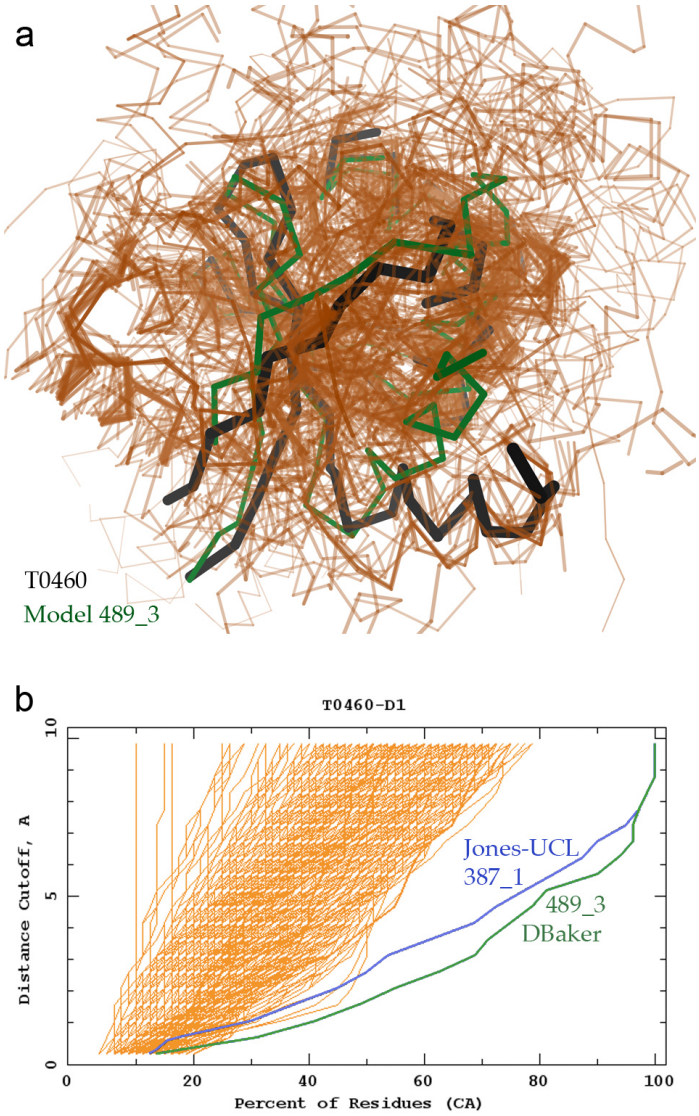

FIGURE 6.10: Two outstanding predictions for the TBM/FM target T0460-D1. (a) Cα traces are shown for the target in black, for the 134/521 best predicted models in terms of Cα placement in peach, and for the particularly exceptional model 489_3 (DBaker) in green. PDB code: 2k4n (NESG, unpublished). (b) Cumulative superposition correctness plot from the Prediction Center website. The percentage of model Cα atoms positioned within a distance cutoff of the corresponding target Cα atom after optimal Cα superposition is shown (x-axis) for a range of such distance cutoffs (y-axis); all models for T0460-D1 are shown in peach. Thus lines lower and further to the right indicate predictions that better coincide with the target. The rightmost lines are models 489_3 (DBaker, green) and 387_1 (Jones-UCL, blue).

6.4 Correct fold identification, self-scoring, and other analyses

In addition to the all-atom scoring of individual models and of predictors across targets, I contributed to several other analyses that broke the mold for CASP assessment.

6.4.1 Consistency of "right fold" identification

As described above (Section 6.3), we only tallied the "added-value" all-atom scores for models with reasonably accurate Cα superposition, i.e. those with the "right fold". However, we also sought to assess which groups excelled at consistently identifying the right fold in the first place, to help delineate the state of the art for that stage of homology modeling.

I initially explored the idea of using good Cα superposition on a set of so-called "core" residues as a measure of having correctly identified the basic protein fold. In this paradigm, the models that "saturate" in terms of core Cα-based superposition (regardless of the results of superposition using all Cαs) are the ones that got the fold correct, presumably by identifying a good template. I attempted to define such cores based on multiple sequence alignments to reveal evolutionarily conserved residues, which one might presume would be more structurally conserved as well, but my analysis showed that some conserved sequence positions are actually on significantly mobile loops. Jane Richardson and I also tried defining cores based on structural variability in the templates, both purely computationally and based on visual inspection. Unfortunately, we found this approach produced artificially large core definitions for small proteins and artificially small core definitions for large proteins, apparently because good template structures for small proteins are statistically more common.

Given these difficulties, we turned to the striking bimodal GDT distributions

(Figure 6.7), and simply defined models with GDT-HA $\geqslant$ 33 as having the right fold. Right fold percentage is also dependent on average difficulty of attempted targets, so Figure 6.11 plots it as a function of average target difficulty. Groups along the linear "outstanding edge" at the top of Figure 6.11 can be considered exemplary given their target choice. For example, Yasara (lauded above for excellent all-atom quality) succeeded on this metric by focusing on easier targets (Figure 6.8, top right), whereas Baker and IBT-LT were equally successful but on more difficult targets (Figure 6.8, top left). The central set of groups attempting essentially all targets can act as a suitable accompaniment to the full-model, high-accuracy score shown in Figure 6.8.

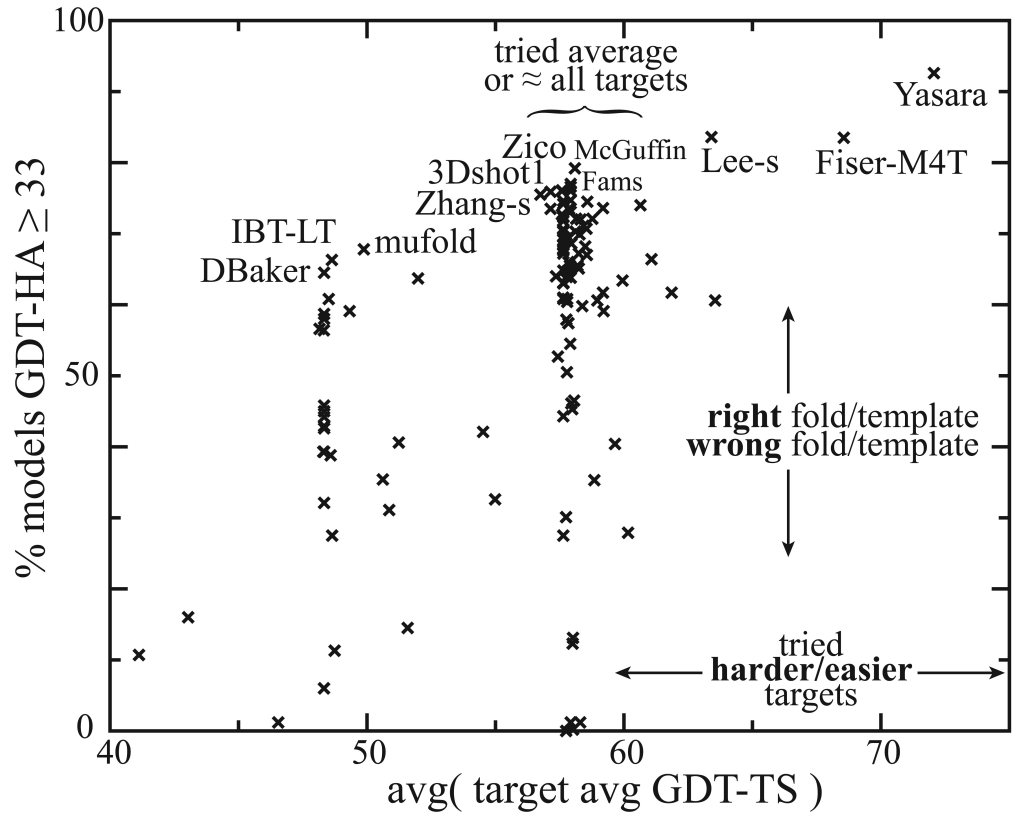

FIGURE 6.11: Percentage of models with roughly the "right fold", plotted vs. difficulty of targets attempted. The percentage of all of a group's models with GDT-HA $\geqslant 33$ ("right fold") is on the y-axis. The average across a group's attempted targets of all-model, all-group average GDT-TS (a measure of target difficulty) is on the x-axis. All groups attempting at least 20 targets are included. Names of several groups along the "outstanding edge" are labeled. Made for (Keedy et al., 2009).

6.4.2 Self-scoring

CASP predictors are allowed to submit up to five models per target, labeled "model 1" through "model 5", but traditionally only model 1 has been assessed. In CASP8 we decided to instead assess the best model, as mentioned above (Section 6.3), and separately assess self-selection, i.e. the ability to recognize the actual best model as model 1. That ability is very important to end users of predictions who want a single definitive answer, especially from publicly available automated servers.

To address this issue, we assessed self-scoring relative to that expected from random chance based on the number of models submitted for each target. Figure 6.12 plots this score against the range of GDT-TS across the up to five models for each target, which serves as a measure of willingness to explore conformational space and/or alternative prediction methods. Most groups are at least 3 σ better than random at picking their best model as model 1, but few are right more than 50% of the time. Strikingly, servers turn out overwhelmingly to dominate the top tier of this metric, making up all of the eight top-scoring groups and all but one of the top 20.

Although successful prediction and successful self-scoring are both very important to further development of the field, these observations suggest they currently remain quite unrelated, and we believe that they should therefore be assessed and encouraged separately.

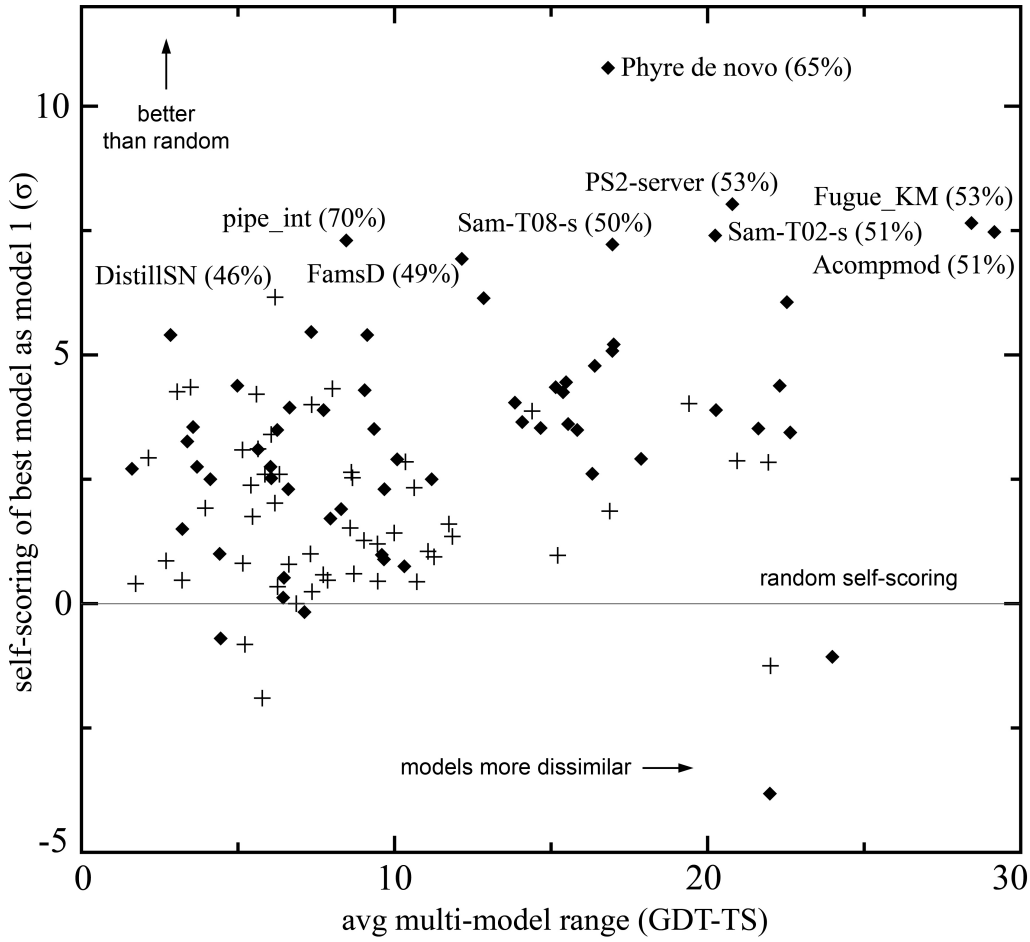

FIGURE 6.12: Ability of groups to self-select their best model as model 1. The difference from the percentage expected based on random chance (correcting for different average numbers of models) is plotted vertically (in units of standard deviations); range of scores within a group's model sets is plotted horizontally. For the best self-scorers, the group name and the percentage of "model 1s" that were actually "best models" are shown. Diamonds indicate server groups, which dominate the top self-scorers; pluses indicate human groups. Made for (Keedy et al., 2009).

6.4.3 Model compaction or stretching

During our CASP8 assessment period, assessors of previous CASPs warned us that some groups may have "over-trained" their methods to maximize their global GDT-TS score at the expense of protein realism. Specifically, they suspected that a known small subset of groups systematically scrunched up or stretched out regions of their models. To test this idea, I worked with Jeffrey Headd to examine the average standard deviation for both bond lengths and bond angles for the six suspicious groups suggested by previous assessors, relative to a control group with good geometry and GDT scores. We found that none of the suspicious groups had systematically longer bond lengths or wider bond angles by more than 0.5 σ, or systematically shorter bond lengths or tighter bond angles by more than 1 σ (Figure 6.13). This represents less than a 1% compaction in the models, which seems unlikely to produce any significant effect on overall GDT scores.

However, local compaction or stretching was much more common, especially in regions requiring a sequence insertion relative to a template structure. For example, the model in Figure 6.14 attempted to span what should be seven residues with only six, resulting in a string of bond-length outliers at 10 σ or more. This strategy places all Cαs within 4 Å of their target positions for this local window, but gets the alternation of sidechain direction wrong for half the residues on average, and therefore fails to produce a biophysically realistic model that would be useful for downstream applications. Clearly, there is still significant room for improvement on modeling insertions relative to templates. In the near future, local compaction or stretching may prove useful for diagnosing regions of homology models for which the insertion or deletion strategy should be reexamined.

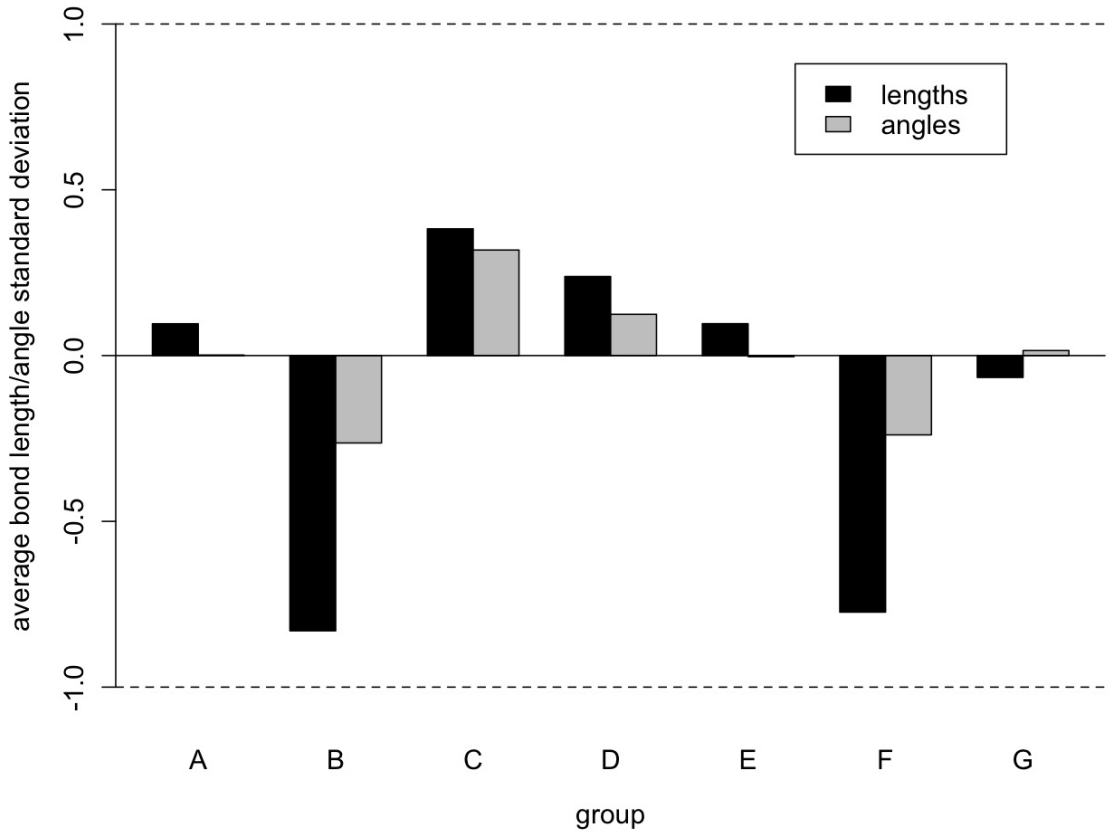

FIGURE 6.13: Lack of systematic model compaction or stretching for "suspicious" CASP8 groups. The average standard deviation is less than one in all cases. Group G is a control known to have good geometry and GDT scores. The groups are assigned arbitrary labels here to protect their identities.

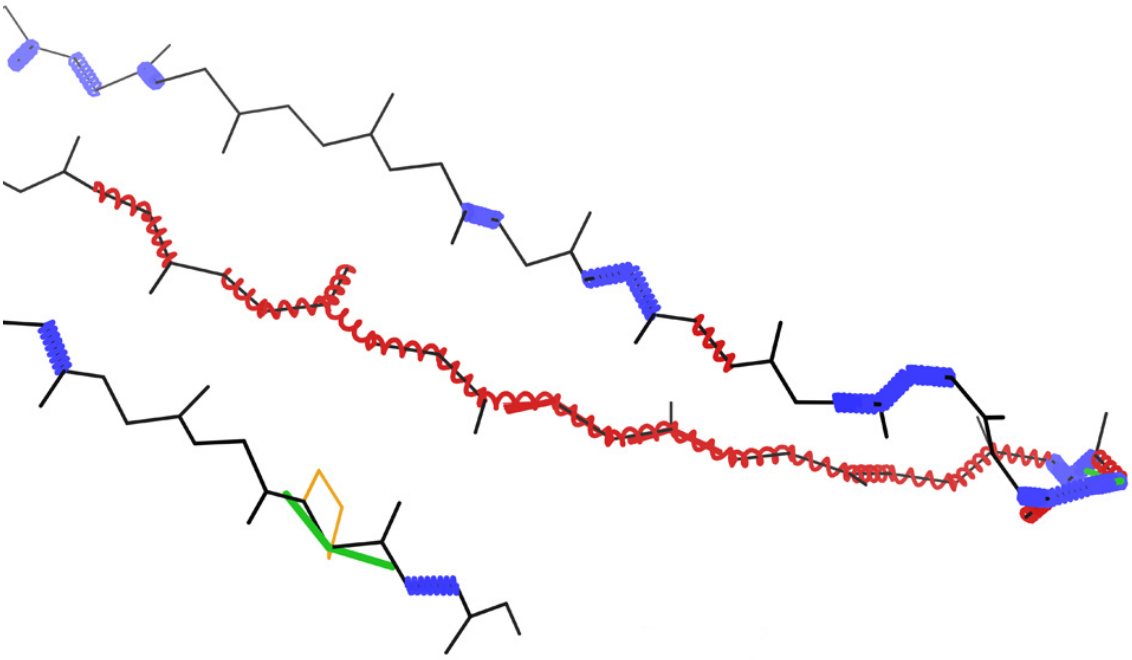

FIGURE 6.14: An over-extended β strand, with main-chain bond-length outliers up to 40 σ, marked as stretched-out red springs. T0487-D1, PDB code: 3dlb, argonaute complex. Taken from (Keedy et al., 2009).

6.5 Discussion

Some of our most striking early observations of CASP models were that (1) many models had quite poor Cα placement, and (2) even models with good Cα placement often had poor placement of all the other atoms. The overwhelming impression was that assessment practices from prior experiments (CASP1 through CASP7) "selected for" techniques that approximated the correct overall Cα trace, but failed to reward techniques that produced genuinely realistic models in a real-world, all-atom sense (with some exceptions (Kopp et al., 2007)). These practices may have been reasonable at the time, given the novelty of quasi-successful computational structure prediction, but they unwittingly reinforced the (subconscious?) paradigm that predicting peptide orientations and sidechain rotamers is a lost cause. To emphasize this point, Figure 6.15 shows a comically unrealistic "structure" from CASP8 that is clearly the result of simple neglect rather than some fundamental limit on prediction ability; until now, this predictor had no official incentive to *try* to predict these sidechains' proper conformations.

Admittedly, we will never know whether more stringent assessment practices, if they had been enforced in the past, *would have* helped advance the field in terms of all-atom quality and correctness. However, our analysis for CASP8 TBM assessment (Keedy et al., 2009) and our contributions to assessment of the "refinement" category (MacCallum et al., 2009) mark a start toward such a grand experiment. By packaging our all-atom-centric tools and format test-and-correction utility into a suite called BACPAC (Beyond Alpha Carbons Prediction Assessment for CASP), which is freely available through the official CASP website as well as our lab website (`http://kinemage.biochem.duke.edu/software/bacpac.php`), we set the stage for continuing study of the structure prediction community's ability to improve the realism of their template-based models. Indeed, many scores contained

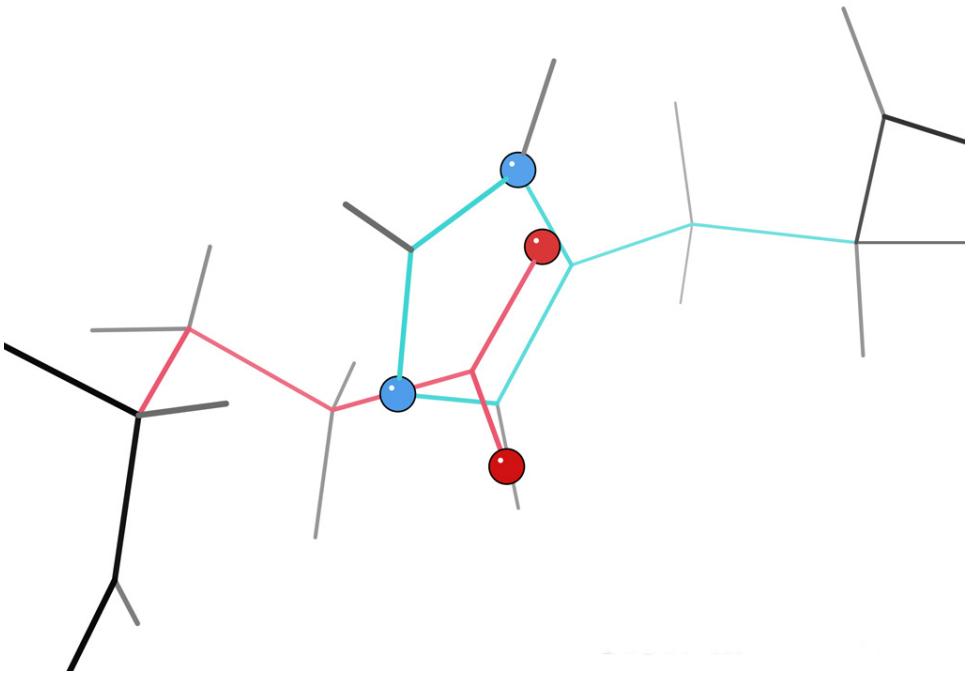

FIGURE 6.15: Modeled "atomic fusion" of Glu and His sidechains in CASP8 target T0389 model 481_1. Steric clashes from Probe are not drawn here – because they don't have to be!

in BACPAC or similar in spirit were used in CASP9: the assessors for the "refinement" category explicitly used GDC-sc and MPscore (MacCallum et al., 2011), and the TBM assessors enforced "physical plausibility" by strongly penalizing models they deemed "unrealistic" on account of too many steric clashes or stereochemical errors (although they reverted many of our other changes...) (Mariani et al., 2011). Furthermore, the Montelione lab are using scores from BACPAC for template-based assessment in CASP10; the runs will be done for them by Andriy Kryshtafovych at the CASP Prediction Center as part of their services to assessors.

More generally, continuing use of our tools for official CASP assessment has the unique potential to provide insight into the fascinating question of interplay between incentive and performance in organized scientific competitions. Such insight may

also be extensible to other, more recently formed CASP-like tournaments, such as CAPRI for prediction of protein-protein interfaces (Janin, 2002) and RNA-Puzzles for prediction of RNA tertiary structure (Cruz et al., 2012).

Our assessment results and several recent studies raise the somewhat related and equally interesting point that human intuition can play a useful role in protein structure modeling. In CASP8 TBM we observed that server groups dominate for easier targets, but human groups comprise the top groups for average and more difficult targets, which require larger excursions from the best templates. Furthermore, many of the "outstanding" models in CASP8 TBM (see Section 6.3) were produced by human groups. The CASP9 TBM assessors also found that human groups performed better than server groups, although the improvement was mild, perhaps because many human groups actually relied heavily on meta-servers (Mariani et al., 2011). Finally, non-expert players of the protein-folding video game Foldit have recently helped successfully solve several difficult prediction and design problems (Cooper et al., 2010; Khatib et al., 2011; Eiben et al., 2012).

By way of contrast, server groups outperformed human groups at self-scoring in CASP8 TBM (Section 6.4), but this may be because server groups used simpler prediction methods with correspondingly simpler interpretations of the results. Furthermore, identifying the best model from a small pool of possibilities, while desirable, is not as important a skill as producing at least one especially good model.

Thus it appears that automated computation of the form implemented in CASP servers is sufficient for certain tasks such as precise energy comparisons or relatively simple homology modeling problems, but human intervention is useful for making big jumps. With human-directed changes, there is more risk of making a starting structure worse, but there is concomitantly more potential reward of finding a significantly better structure. Clearly we will have to wait for the ideal scenario of completely reliable automated prediction regardless of target difficulty, but in the meantime it

appears that tightly coupled man-machine interfaces may have significant benefits for protein modeling.

Overall, I found CASP assessment to be overwhelming and sometimes bewildering, but ultimately rewarding. (I suspect many other contributing members of the Richardson lab feel similarly – now that assessment is over, at least.) Most importantly for my thesis work, CASP assessment provided another opportunity for comparative validation of competing alternatives for which all-atom evaluation proved critical. In this case, the vast data set of submitted models contains a very large number of *invalid* alternatives; fortunately, our all-atom scoring criteria are trained using quality-filtered experimental structures, and therefore have "an eye for" realistic conformations, allowing us to recognize them amongst the decoys. My hope is that this perspective will prove to be a welcome and influential contribution to the homology modeling field.

7

Validation of Rosetta

7.1 An introduction to Rosetta

In the past decade, the software package Rosetta (Rohl et al., 2004; Leaver-Fay et al., 2011) has emerged as the state of the art for macromolecular structural modeling. Originally intended for *ab initio* structure prediction (i.e. without starting from a template structure), it had its first breakthrough success in CASP3, yielding models within 4-6 Å Cα RMSD of the crystal structure (Simons et al., 1999). In the years since, enhanced methodologies and increased computing power have increased that accuracy to $<$ 1.5 Å Cα RMSD for relatively small proteins (Bradley et al., 2005). Rosetta was also co-opted for protein design, resulting early on in the creation of the first novel globular fold, Top7 (Kuhlman et al., 2003) – a landmark accomplishment. Subsequent successes have included the *de novo* design of enzymes that accelerate previously reactions previously uncatalyzed by any enzyme (Röthlisberger et al., 2008; Jiang et al., 2008; Siegel et al., 2010), although they are still orders of magnitude slower than many natural enzymes.

More recently, experimental data has been integrated with Rosetta in various

ways to focus the conformational search process, resulting in many spectacular successes. CS-Rosetta uses NMR chemical shifts to rapidly and accurately determine the structures of small to medium proteins (Shen et al., 2008); it was also applied to yield the structure of an "invisible" excited state of a T4 lysozyme mutant (Bouvignies et al., 2011). When sparse backbone chemical shifts, RDCs, and NOEs are combined, accurate structures of even larger proteins can be obtained (Raman et al., 2010). In addition to NMR, Rosetta has also been integrated with X-ray crystallography: a hybrid method using electron density to guide energy minimization was used to solve several challenging crystallographic data sets that had stymied expert crystallographers and existing molecular replacement techniques (DiMaio et al., 2011).

Despite the glamour of these recent reports, their success is largely predicated on the information content of the experimental data component. By contrast, "pure" *ab initio* structure prediction and *de novo* design remain challenging goals: Rosetta still reaches the native conformation only rarely in unbiased prediction simulations, and most of its proposed designed sequences fail to fold and/or function as desired. This is because experimental guidance can mask deficiencies in the stochastic conformational search process and oversimplified energy function.

Yet Rosetta's computational methodology has a relatively strong foundation: a sampling approach called fragment insertion (Simons et al., 1997), in which candidate conformations for a local region are derived from a set of conformations adopted by similar sequences in observed structures ("fragments"). Fragment insertion aims to mimic a presumed kinetic process in which a protein adopts its final structure thusfold: local regions fluctuate in and out of possible conformations dictated by local sequence, and the unanimous global conformation is dictated by the union of compatible local conformations. To facilitate this approach, Rosetta employs a physically motivated energy function augmented with empirically based terms, with

which to evaluate and subsequently accept or reject the aforementioned alternative local structures on the basis of their compatibility with each other. Notwithstanding the relative effectiveness and intuitiveness of Rosetta's fundamental computational aspects, they need improvement before the more difficult tasks of *ab initio* prediction and *de novo* design can be mastered.

This chapter details several collaborative studies with David Baker's group, the original authors and ongoing purveyors of Rosetta. I used our lab's tools and expertise in experimental structure validation to assess the validity of low-energy models generated by Rosetta, and thereby more accurately define Rosetta's current strengths and (more importantly) weaknesses. The results taught us that some of Rosetta's "failures" are due to artifacts from protein crystallization, but many more can be attributed to blind spots or neglected contributions in its energy function (especially ordered waters) and to fundamentally problematic conflicts between globally oriented statistical terms and locally oriented physico-chemical terms.

7.2 Mapping energy landscapes to find alternate states

Before our CASP8 assessment experience (Chapter 6) had had time to fade from memory, David Baker and his postdoc Mike Tyka came to us with a fascinating data set (Figure 7.1) (Tyka et al., 2010) just begging for the type of detailed examination our lab is famous for. They had used native-enhanced sampling to generate detailed maps of the energy landscapes of 111 protein domains. Hundreds of thousands of independent Monte Carlo trajectories were carried out for each protein using the Rosetta@home distributed computing project (`http://boinc.bakerlab.org/rosetta/`). Each trajectory consists of an initial low-resolution search followed by detailed all-atom refinement. To enhance sampling near the native structure, which is generally sampled quite rarely, in a subset of the trajectories bias toward the native structure was included in the move set used in the initial search and in the selection of

coarse-grained models for all-atom refinement; therefore we refer to this procedure as energy landscape mapping rather than *de novo* structure prediction. Each trajectory ends up in a local energy minimum, and the hundreds of thousands of local minima together provide a detailed map of the energy landscape.

The most striking initial observation is that the native structure almost always lies in a deep energy minimum: protein conformations with Cα RMSD of greater than 4 Å to the deposited structure almost always have higher energies (see, for example, Figure 7.1 inset). For 41% of the proteins examined the lowest-energy model is within 1.2 Å Cα RMSD from the deposited structure, and for 72% it is within 2.5 Å Cα RMSD. Of all the residues simulated, 50% show Cα-Cα deviations of less than 0.3 Å, and 90% show deviations of less than 0.8 Å from the corresponding native residue after global superposition of the lowest-energy model onto the target structure.

However, while the computed global minimum is almost always close to the native structure, it is rarely identical. My primary role in this collaboration was to investigate these quite unanticipated differences, using a multi-pronged approach entailing structure validation with MolProbity (Davis et al., 2007; Chen et al., 2009c), bioinformatics with homology tools from the PDB (Berman et al., 2000), reexamination of experimental data with electron density maps from the EDS (Kleywegt et al., 2004), and molecular visualization with KiNG (Chen et al., 2009b). All relevant data and tabulated results were made available online at
`http://kinemage.biochem.duke.edu/suppinfo/landscapes/`.

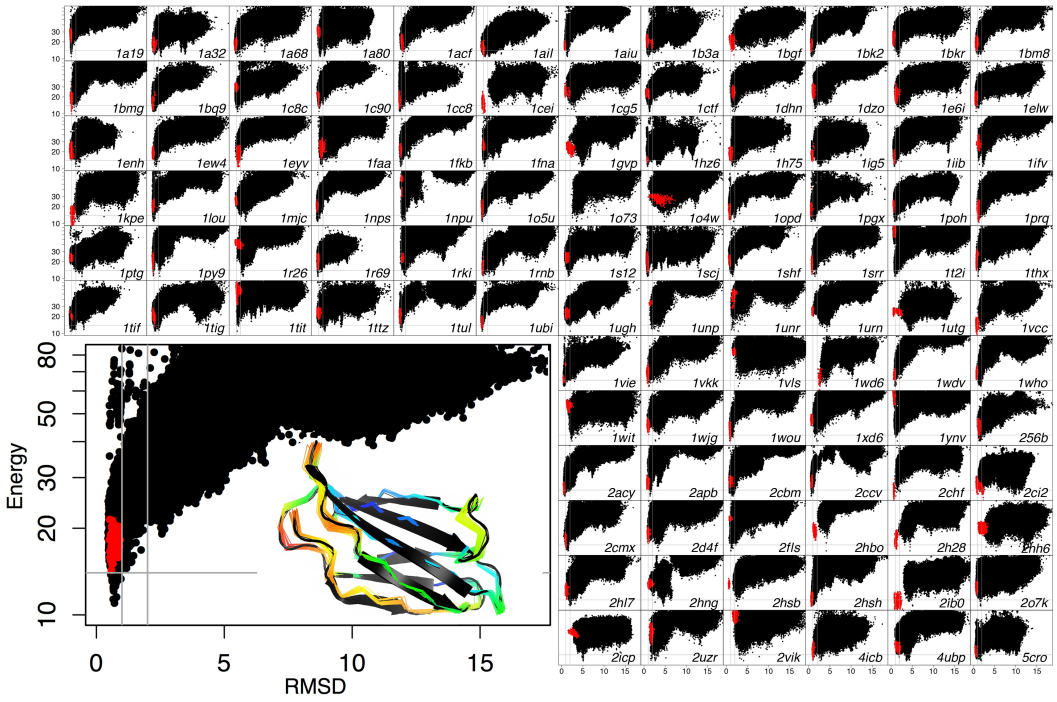

FIGURE 7.1: Computed energy landscapes. Each panel represents a different protein. The y-axis is the Rosetta all-atom energy and the x-axis is the Cα RMSD from the crystal structure; red dots are models relaxed from the crystal structure. The inset shows the energy landscape for 1ten (a fibronectin type III domain) in more detail and a superposition of the models within four energy units of the lowest-energy model (indicated by the horizontal gray line in the plot) on the crystal structure (black). Colors indicate amount of variation in the Rosetta ensemble (blue, low; red, high); variation is concentrated toward the loops. The vertical gray bars indicate the 1 and 2 Å points. Note that the y-axis has been compressed at higher values to fit in the high-energy states without losing detail at the lower (more interesting) energies. For 41% of the proteins examined, the lowest-energy structure is within 1.2 Å Cα RMSD from the deposited crystal structure (as for 1ten), and for 70%, it is within 2.5 Å Cα RMSD.

Errors in native structures

One can imagine many causative factors for the deviations, but an enticing possibility that immediately sprang to mind was that the crystal structures contained errors of the sort MolProbity is designed to detect, and thus that the computed models may be better representatives of the true energy minimum. This possibility was not inconceivable given that Rosetta's energy function (in conjunction with extensive conformational sampling) was sufficient to identify globally quite accurate models for most the 111 proteins examined, and that essentially every crystal and NMR structure contains at least a few genuine errors (Davis et al., 2007) (with only a few exceptions – see the "paragon" investigations in Chapter 4). Furthermore, I was in the perfect position to test such a hypothesis, fresh off CASP assessment and with world-renowned expertise in structure validation at my fingertips.

A few examples did provide support for this idea. For example, Thr77 and Thr101 in 1bkr are in bad rotamers that are flipped 180° relative to their correct counterparts (Headd et al., 2009), introducing several MolProbity errors. A manually corrected and re-refined version of 1bkr repairs these defects; strikingly, Rosetta's low-energy models do the same without the aid of experimental electron density (Figure 7.2). However, in this case the computational models and original deposited crystal structure differ in terms of their sidechain rotamers rather than their Cα placement, so the deviations do not appear in score vs. Cα RMSD plots like those in Figure 7.1. Furthermore, only about 4% of the local deviations in this data set were demonstrably related to errors in the experimental structure. This is likely due to the fact that errors in experimental structures involving displacements of Cα atoms that are large enough to noticeably affect global Cα RMSD seldom occur. Usually the experimental data is sufficient to overwhelmingly pinpoint the proper mainchain conformation, especially for crystal structures at moderate to high resolution. At

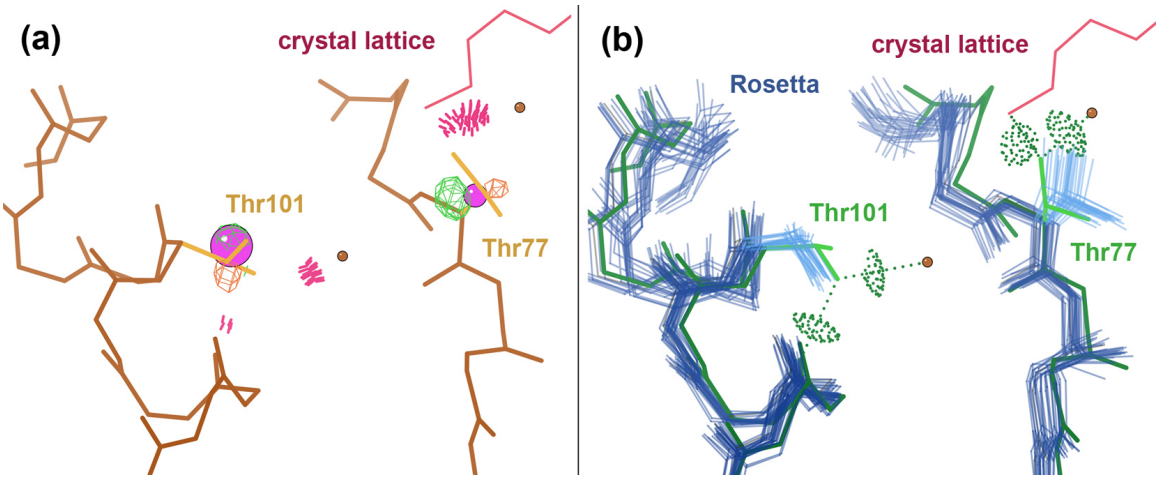

FIGURE 7.2: Correction of local errors in a deposited crystal structure. (a) MolProbity detects errors by several criteria for Thr77 and Thr101 in a crystal structure of calponin homology domain (1bkr): rotamer outliers, Cβ deviations (pink balls), and steric clashes (pink spikes) to surrounding water molecules (brown balls) and protein atoms (to a Lys side chain of another molecule in the crystal in the case of Thr77). Furthermore, the Cβ atoms for both Thr side chains fall nearer to negative 5 σ Fo-Fc difference density peaks (orange mesh) than to positive peaks (green mesh), indicating a mismatch to the experimental data. (b) The majority of Rosetta's low-energy models (blue) flip both side chains by 180° (Headd et al., 2009) to eliminate clashes, establish hydrogen bonds with surrounding atoms, and fortuitously better fit the difference density. A structure independently re-refined against the original diffraction data by the Richardson lab (green) corroborates this flip. Note that Rosetta's backbone is somewhat mobile, especially for Thr77, perhaps because stabilizing effects from the explicit water molecules and the crystal contact are not modeled. Nevertheless, in this case at least, Rosetta's energy function is sufficient to detect the proper side-chain conformations.

low resolution, gross mainchain errors sometimes occur: for example, others in our lab have identified erroneous sequence register shifts, some as long as 10 residues, in the protein portions of a 3.011 Å structure of the *E. coli* 70S ribosome. However, the worst resolution in the data set considered here was 2.9 Å, and 85% of targets had resolution no worse than 2.0 Å. In general, the more common classes of errors – and the ones we have more experience diagnosing (Davis et al., 2007; Chen et al., 2009c) – are subtle local sidechain or mainchain errors, or gross sidechain errors such as

entirely misfit rotamers, but not gross mainchain errors such as entirely misfit loops.

Errors in Rosetta models

Conversely, in many cases the experimental structure was correct and the Rosetta models appeared to have some deficiency that explained the deviation. This scenario was significantly more common, accounting for almost 20% of discrepancies. A deviation could be relatively safely assigned to this category if the native structure was error-free, no homolog (60-99% sequence identity) or "isolog" (100% sequence identity) structures were available to corroborate the Rosetta models, and the region was free of extenuating multimer or lattice interactions (see below). However, it was often more difficult to identify the precise deficiency in Rosetta leading to the misprediction, since the low-energy computational models reflect a precise balance between numerous competing energetic terms – although later collaborative work (Section 7.3) was more suggestive as to the particular energetic mis-weightings leading to poor models. That said, one easily detectable deficiency was the neglect of ordered water molecules in favor of a computationally less expensive – but correspondingly less accurate – implicit solvation model. For example, a well-ordered water in 1wd6 "peels apart" two β strands on one end of a sheet; by contrast, Rosetta adheres to ideal sheet conformation through this region due to the absence of the water (Figure 7.3).

Biological quaternary interactions

In many of the remaining cases, neither the experimental structure nor the Rosetta models was "wrong"; rather, the crystal or NMR structure represented a true complex, whereas the Rosetta models represented estimates of the isolated monomer structure. Indeed, I discovered that over 31% of deviations could be attributed to missing macromolecular binding partners or ligands.

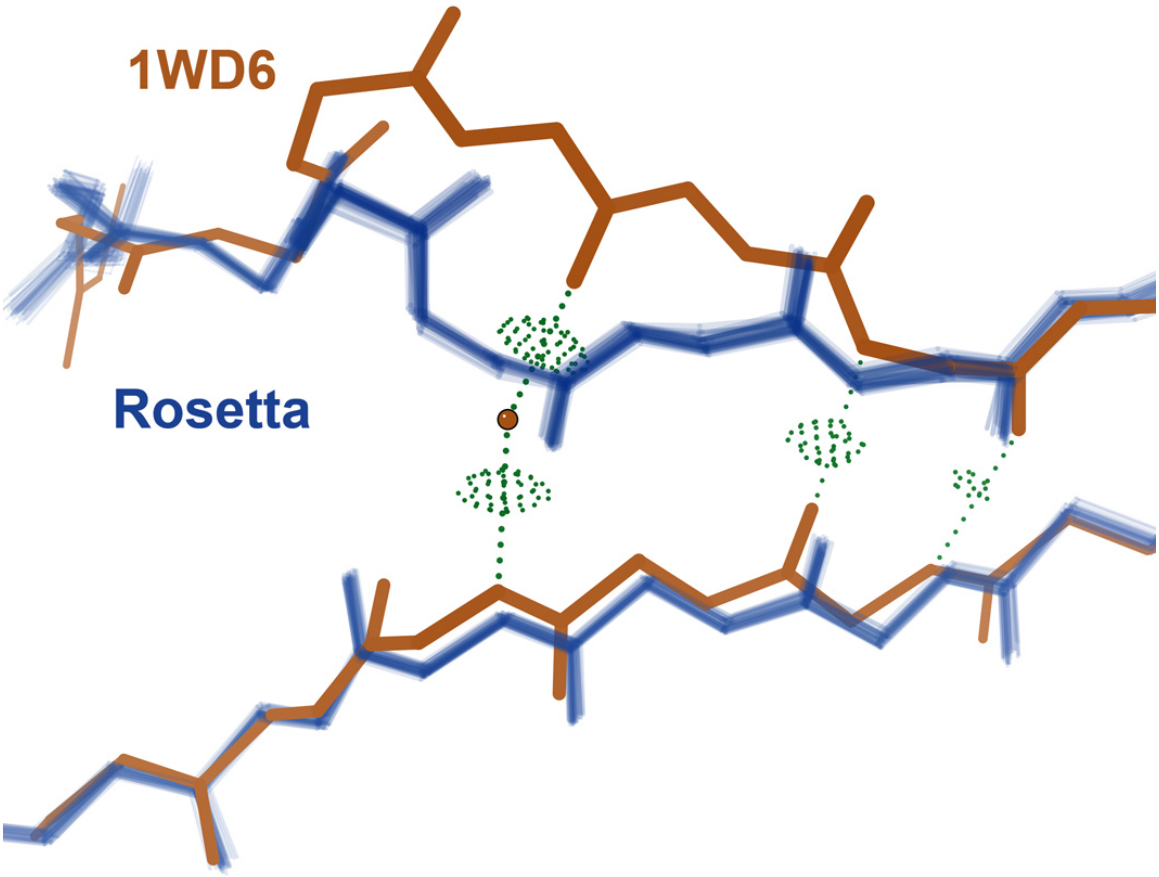

FIGURE 7.3: Example of an erroneous computed alternate conformation. For the protein JW1657 from *Escherichia coli* (1wd6, brown), an explicit water molecule (brown ball) peels apart the two strands of a parallel β sheet while maintaining excellent hydrogen bonds (green dots) to maintain the protein's structural integrity. Rosetta cannot consider the possibility of an explicit water molecule because it employs an implicit solvent model; therefore, the computed low-energy models revert to overly idealized (and in this case incorrect) β structure. The low B-factor (13.8) of the water suggests it is well ordered and precisely placed, and chain B of 1wd6 as well as other homologs confirm its position.

For many of those cases, it was impossible to ascertain the veracity of the Rosetta models as estimates of the unbound state due to a lack of corroborating evidence. For some other cases, however, experimental structures of the same protein in an apo state were available. For instance, the Rosetta models computed for 1urn, the RNA-binding domain of the U1A spliceosomal protein, in the absence of its RNA binding partner fall into two low-energy clusters: one that matches the RNA-bound conformation, and another that matches the unbound conformation as seen in the apo crystal structure 1nu4 (Figure 7.4). These observations are consistent with a significant importance for conformational selection in the RNA-binding mechanism.

Such successes increased our confidence that the computed monomers for homo-oligomeric structures may correspond to the the real conformations adopted by these proteins after synthesis but before oligomerization, and thereby fill a gap that is hard to address experimentally.

Non-biological crystal contacts

After the above mentioned rationales had been fully explored, many significant deviations between Rosetta models and experimental structures remained unexplained. One major culprit we noticed was the presence of the crystal lattice in the experimental structures (only three targets were NMR structures) but not in the simulations.

For example, the N-terminus of thioredoxin forms mainchain-mainchain β-sheet contacts with a neighboring unit cell in 1faa, but collapses toward the body of the protein in isolated-monomer Rosetta simulations (Figure 7.5).

To confirm that crystal contacts indeed caused the deviations, rather than being fortuitously co-localized with them, the Baker lab performed further computations in which all-atom energy minimization was performed in simulated lattices generated using crystallographic symmetry. Indeed, in many cases the calculations carried out in the presence of crystal contacts converge on minima considerably closer to

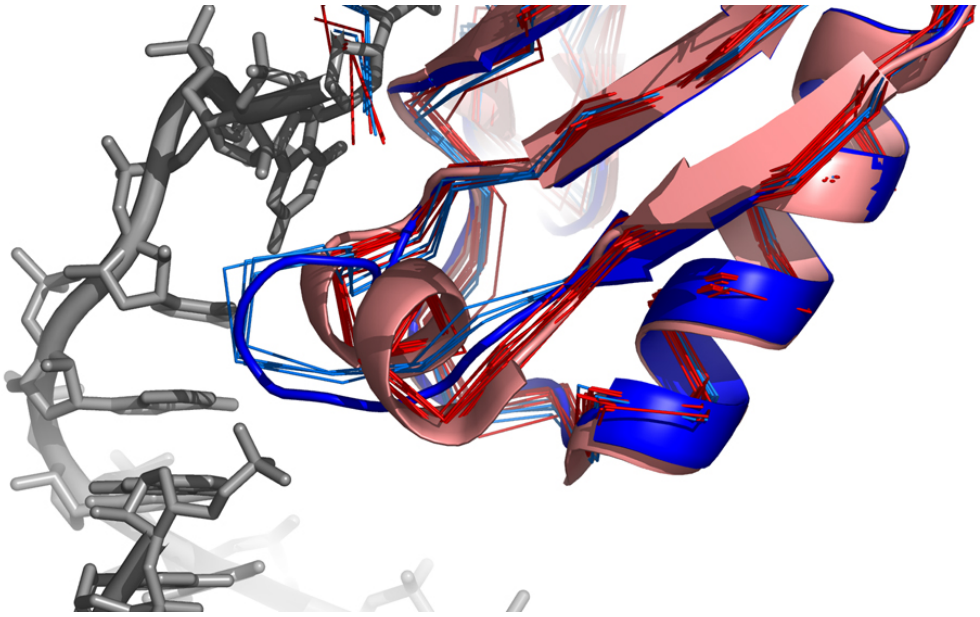

FIGURE 7.4: Influence of binding partner. The simulation of 1urn identifies two pronounced minima in the main RNA binding loop 46-52. One (thin blue backbones) matches the conformation found in 1urn (thick blue backbone) contacting the RNA, while the other (thin red backbones) forms a short helix matching the unbound conformation found in 1nu4 chain A (thick red backbone), a crystal structure of the apo form of this protein. Rosetta ranks these two minima (in the absence of RNA) equal in energy, suggesting that both the bound and apo conformations could be sampled in solution. This is further supported by the fact that chain B in 1nu4 is in a conformation close to that of 1urn.

the experimentally determined structures than did the original isolated monomer calculations.

Concomitantly, I validated deviations at crystal contacts by manually comparing the conformations in Rosetta's models vs. those in structures of the same (or occasionally a very similar) protein in a different crystal lattice. In many cases, such as the one illustrated in Figure 7.6, the alternate structure lacking the crystal contact agreed virtually perfectly with Rosetta's models. Such correspondences lend credence to Rosetta's ability to correctly model both solution states of proteins given their (mostly (Davis et al., 2006; Fraser et al., 2011)) static crystal structures and,

by extension, monomeric states prior to oligomerization (see previous section).

More broadly, these results hammer home the fact that crystal lattices impose significant restrictions, and suggest that the precise conformations of surface loops and even secondary structure elements may be artifacts of the experimental method. Functionally, they support a plastic view of protein structure in which certain regions are able to access a multitude of nearly isoenergetic minima and thus are very sensitive to binding interactions.

Along similar lines, recent reports indicate that cryogenic practices in crystallography not only reduce a protein's inherent conformational dynamics but also reshape the energy landscapes of its dynamically shifting sidechains (Fraser et al., 2011). Those studies, in conjunction with our observations described here, motivate a recalibration of how we think of crystal structures. To be precise, a crystal structure contains the subset of conformations that the crystal lattice selects out of the in-cell or in-solution population distribution, subsequently modified by cryogenic freezing.

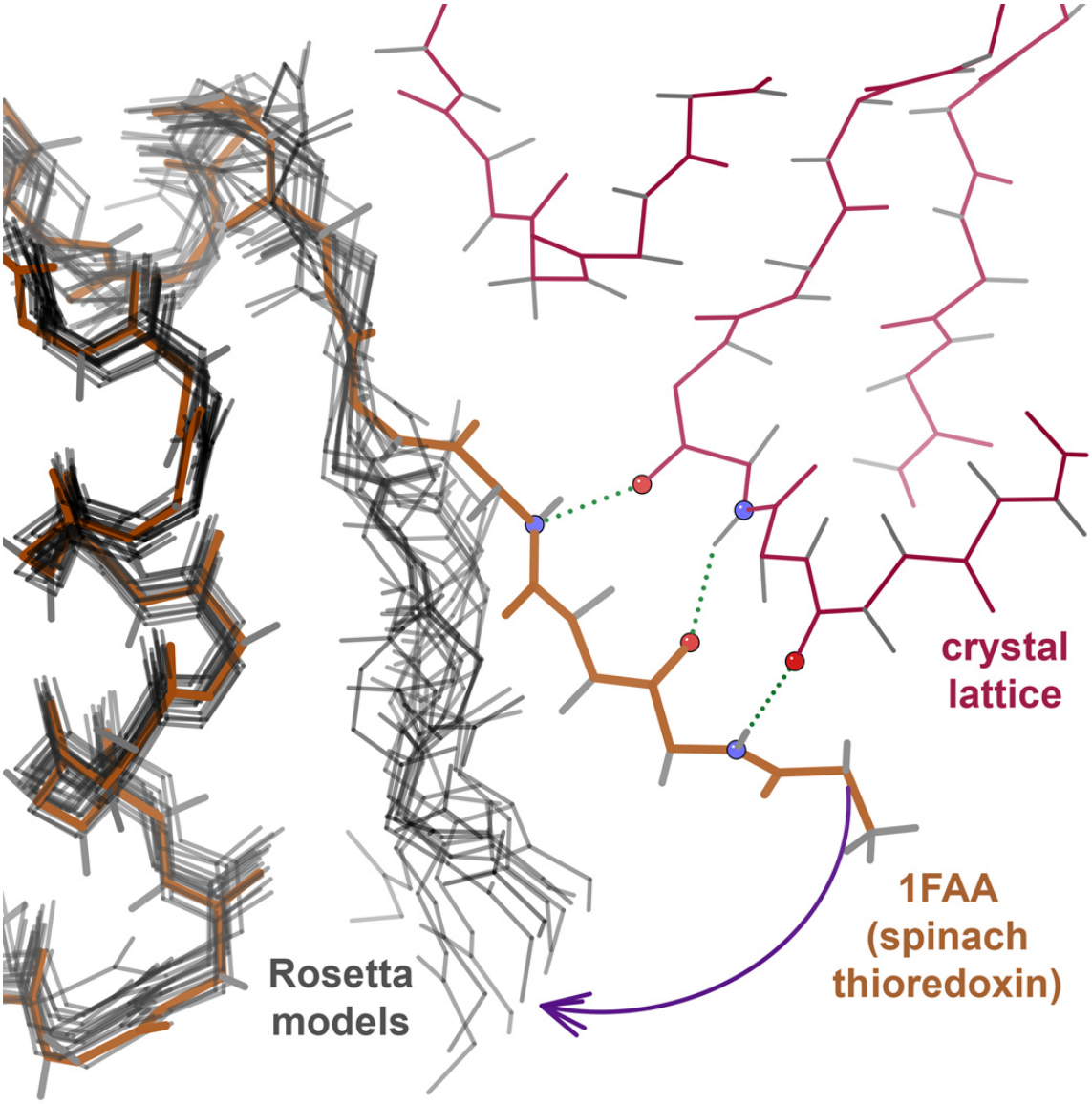

FIGURE 7.5: Effect of crystal-packing interactions. In a crystal structure of a monomeric spinach thioredoxin (1faa) (brown), the N-terminus engages in significant β-sheet-like contacts to a crystal lattice neighbor (pink). In the isolated monomer simulation, the "pull" from the crystal contact is absent, and Rosetta's low-energy models (gray) adopt a wide range of conformations that all collapse toward the body of the protein.

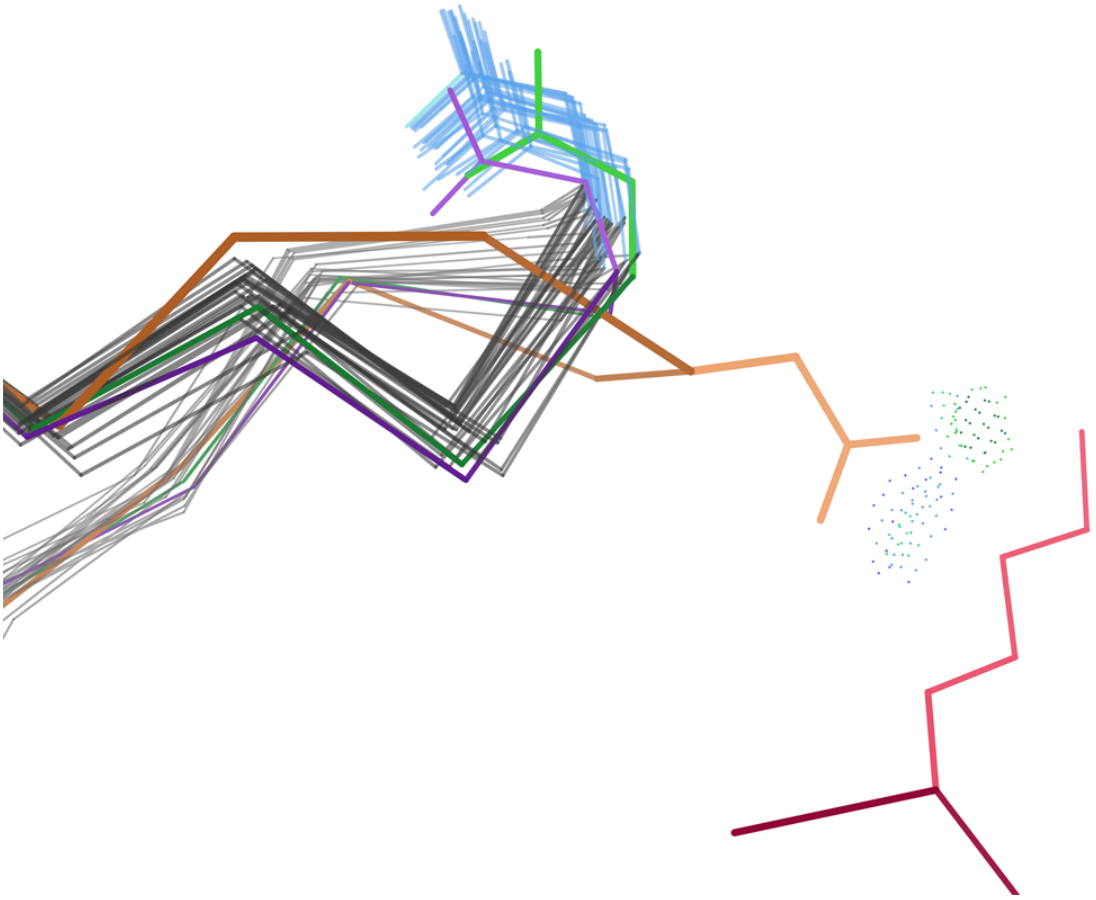

FIGURE 7.6: Effect of crystal-packing interactions. In a crystal structure of human immunophilin FKBP12 (1fkb) (brown), Asp32 contacts a lysine sidechain in a crystal lattice neighbor (pink). In the isolated monomer simulation, this interaction is absent, and Rosetta's low-energy models (gray) "pull" the loop housing the Asp inward. Other crystal structures of the same protein solved in different space groups and lacking this crystal contact, 1d6o (green) and 1d7i (purple), match Rosetta's loop conformation.

Overall, investigating the alternate states identified in this project provided an interesting contrast to CASP8 assessment (Chapter 6). For CASP, the target crystal structures could be taken as ground truth, since few predicted models were close enough for us to consider the type of localized deviations at the heart of this project. Here, by contrast, aggressive minimization seeded by native and homologous fragments in a top-notch energy function produced a sufficiently detailed "energy landscape mapping" that native minima were often quite identified quite accurately. Given this high level of success, we could turn our attention to the well-defined but smaller discrepancies and actually consider the possibility that the computational models were in some respects "better" than the experimental structures. This paradigm shift presented an intriguing twist on this thesis's theme of validating alternatives.

7.3 Understanding false Rosetta energy minima

Two more data sets provided a contrast to the plausible models discussed above by focusing on Rosetta's failures instead of its pleasantly surprising successes. The first contained significantly deviant global folds; the second focused on mispredicted single arginine sidechains. Rather than inducing despair, the natural "glass half empty" perspective induced by these failures led me to usefully identify several notable deficiencies in Rosetta's energy function.

7.3.1 False global energy minima

The first data set, provided by Mike Tyka, contained 14 target structures and anywhere from 2 to 38 (typically 4 to 10) low-energy models per target. The models spanned a wide range of similarity to the target, with Cα RMSD ranging from 0.8 Å (near-native) to $>$ 14 Å (false minima) (Figure 7.7), suggesting an undesirable degeneracy in the energy function. In contrast to the data set from Section 7.2, most

of the deviations between models and targets appeared to be unrelated to crystal contacts, and instead likely reflected malfunctions in the underlying Rosetta energy function.

To investigate the fundamental causes for these unwarranted deviations, I examined the faulty models at length using a cadre of assessment techniques. One feature which immediately jumped to my attention was the high rotamer score or "rotamericity" of the models relative to native protein structures (Figure 7.8). (The effect was similar but less pronounced for Ramachandran score.) Notably, the computed models strongly preferred the absolute most probable rotamers. Native structures, by contrast, permit rotamers of all percentiles equally – this is of course true by construction, since our rotamericity score is the percent of sidechains whose χ angle combinations fall into less populated regions.

Despite this glaring artificiality, rotamericity fails to correlate whatsoever with distance from model Cα to target Cα after global superposition (Figure 7.9). Thus the globally averaged local evidence does not suggest that overly rotameric sidechains induce backbone inaccuracies. It is likely that many sidechain errors are simply problems in and of themselves, independent of global Cα trace. However, it remains possible that some individual mispredicted sidechains cause localized backbone errors, or even that in some instances a few spatially adjacent mispredicted sidechains conspire to push a global fold over the edge to a similar but non-native-like fold; this topic deserves more concentrated study.

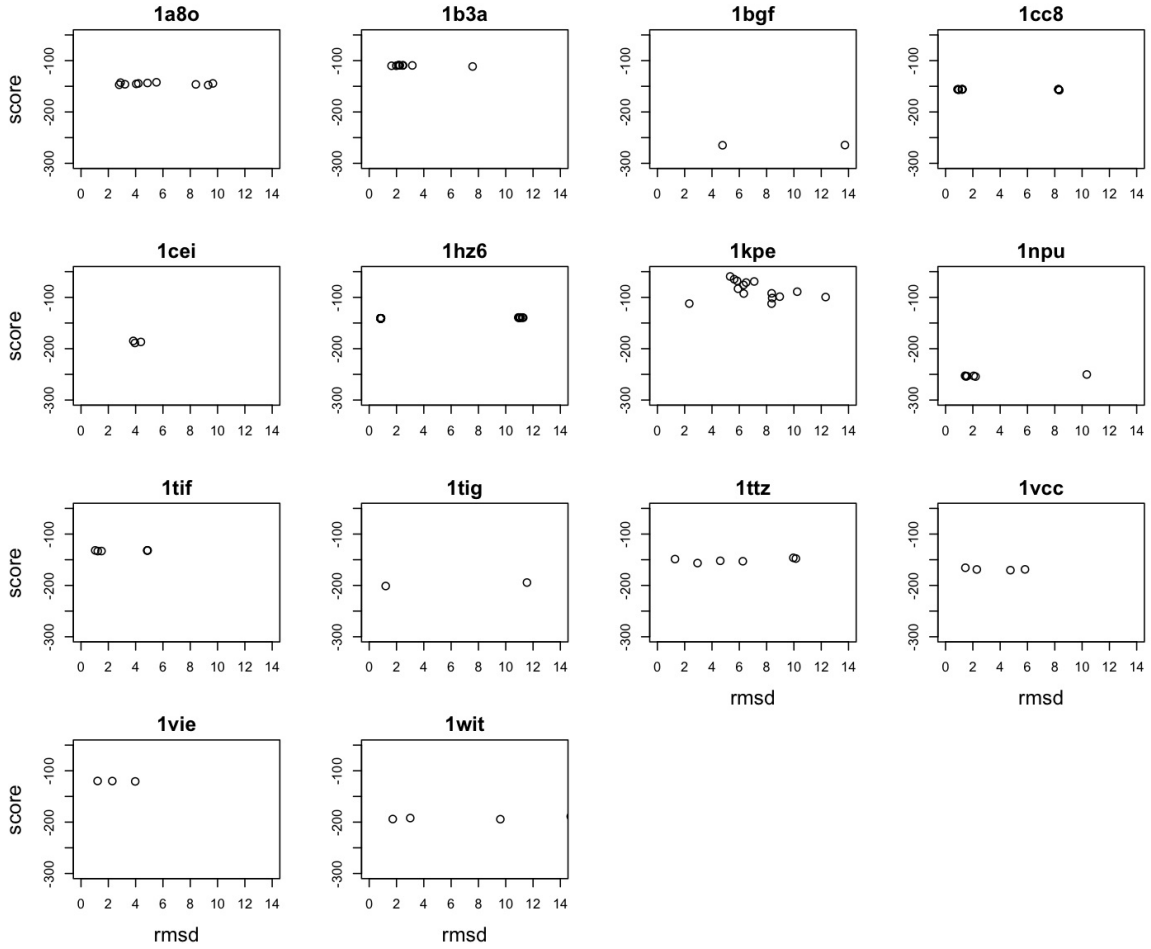

FIGURE 7.7: Rosetta models with false global energy minima. Rosetta score is plotted against Cα RMSD for computed models for 14 targets. Note that many models with very high RMSD values have similar energy as more native-like models. These models are analogous to the bottoms of "funnels" in Figure 7.1, but here they betray errors in the Rosetta energy function rather than potentially biologically interesting alternate conformations.

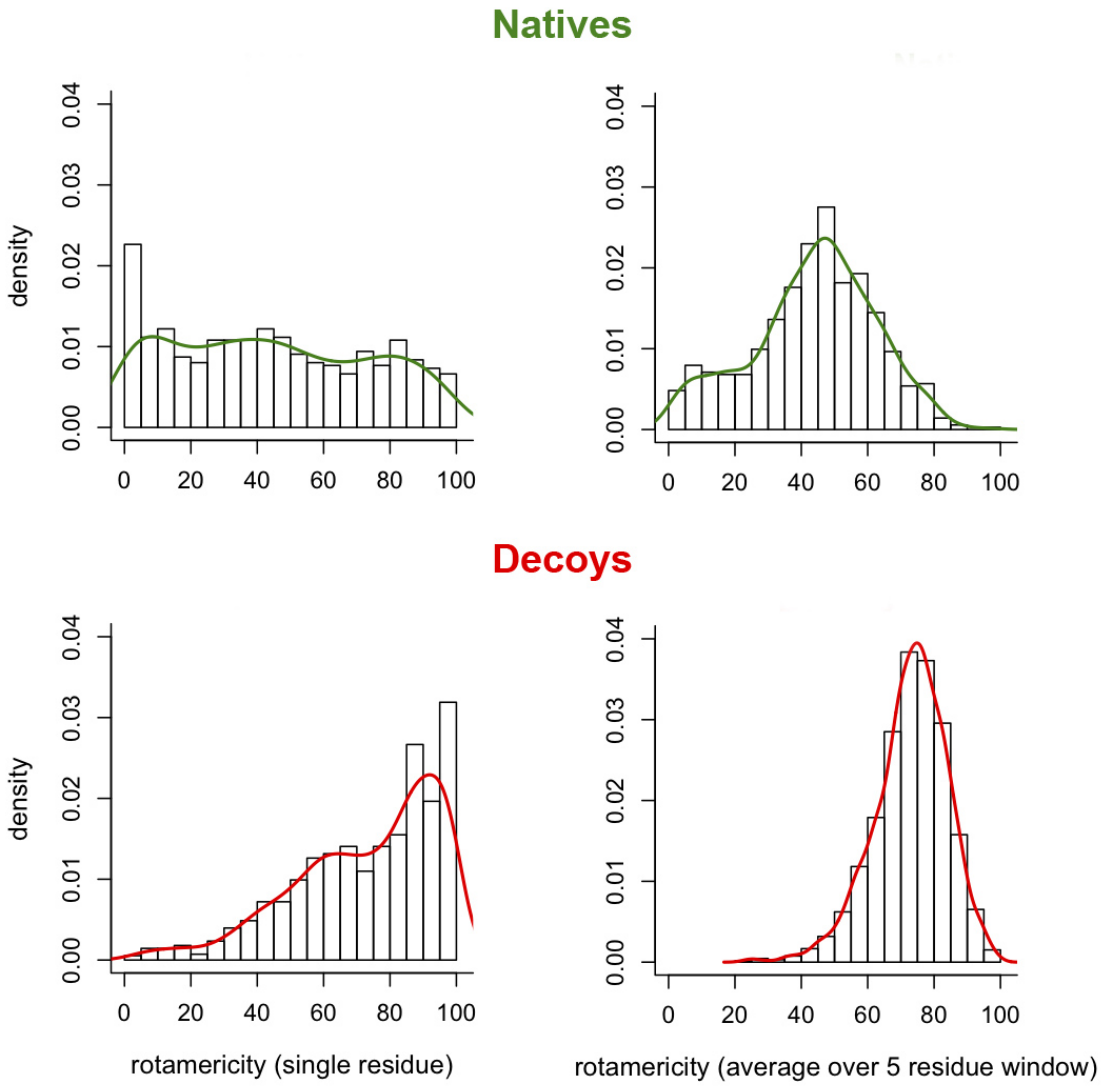

FIGURE 7.8: Rosetta models have excessively high "rotamericity" scores. Histograms for the 14 native structures in the data set (top) show that rotamericity is symmetrically distributed, whether for single residues or averaged across five-residue windows. For Rosetta decoy models (bottom), on the other hand, rotamericity is shifted strongly to higher values, which correspond to more probable rotamers.

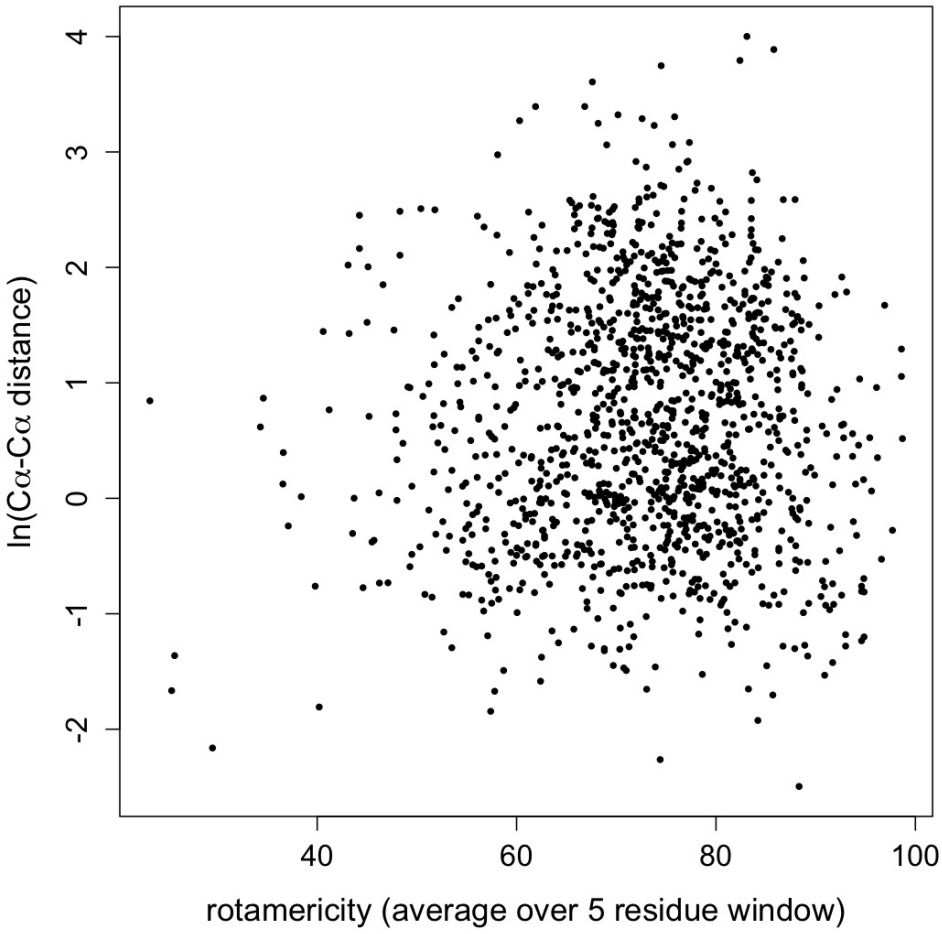

FIGURE 7.9: Excessive "rotamericity" is globally uncorrelated with Cα inaccuracy. Rotamericity averaged across five-residue windows fails to predict Cα-Cα model-target distance.

On a whim, I also examined the percentage of all-atom contacts (Word et al., 1999b) in each of three categories: sidechain-sidechain, sidechain-mainchain, and mainchain-mainchain. I suspected that the computed models had "un-protein-like" relative amounts of these contact types. However, the global percentages in each category were similar for native structures and Rosetta models (Figure 7.10).

In addition to Probe dot counts, I pursued a complementary approach that focused on the extent of "interdigitation" of sidechain contacts. First, for each residue I defined a sidechain axis $\vec{x}$ from the Cα to the centroid of all sidechain atoms (including hydrogens). Next, for each sidechain atom I defined a vector $\vec{n}$ to the nearest neighbor atom that was both in a different residue and $\leqslant$ 4 Å away. I then very simplistically calculated the angle θ between $\vec{x}$ and $\vec{n}$ for each sidechain atom. The average θ for each residue indicates its degree of interdigitation with its surroundings, where 90° is most interdigitated and 0° is least interdigitated. I excluded "surface residues", defined in a kludgy fashion based on combinations of several factors: absence of near neighbors, contact with one or more waters, distance from protein centroid, and relative orientation of sidechain axis and vector from Cα to protein centroid. This approach fails to account for sequence differences, since proteins with longer sidechains have statistically more possibilities for orthogonal contacts as opposed to end-on contacts, but comparison of interdigitation for computed models vs. experimental structures of the same protein should be valid. With this data set, however, I found that for each protein the average θ for the decoy models was very similar to the average θ for the native structures, with all differences $< 5°$.

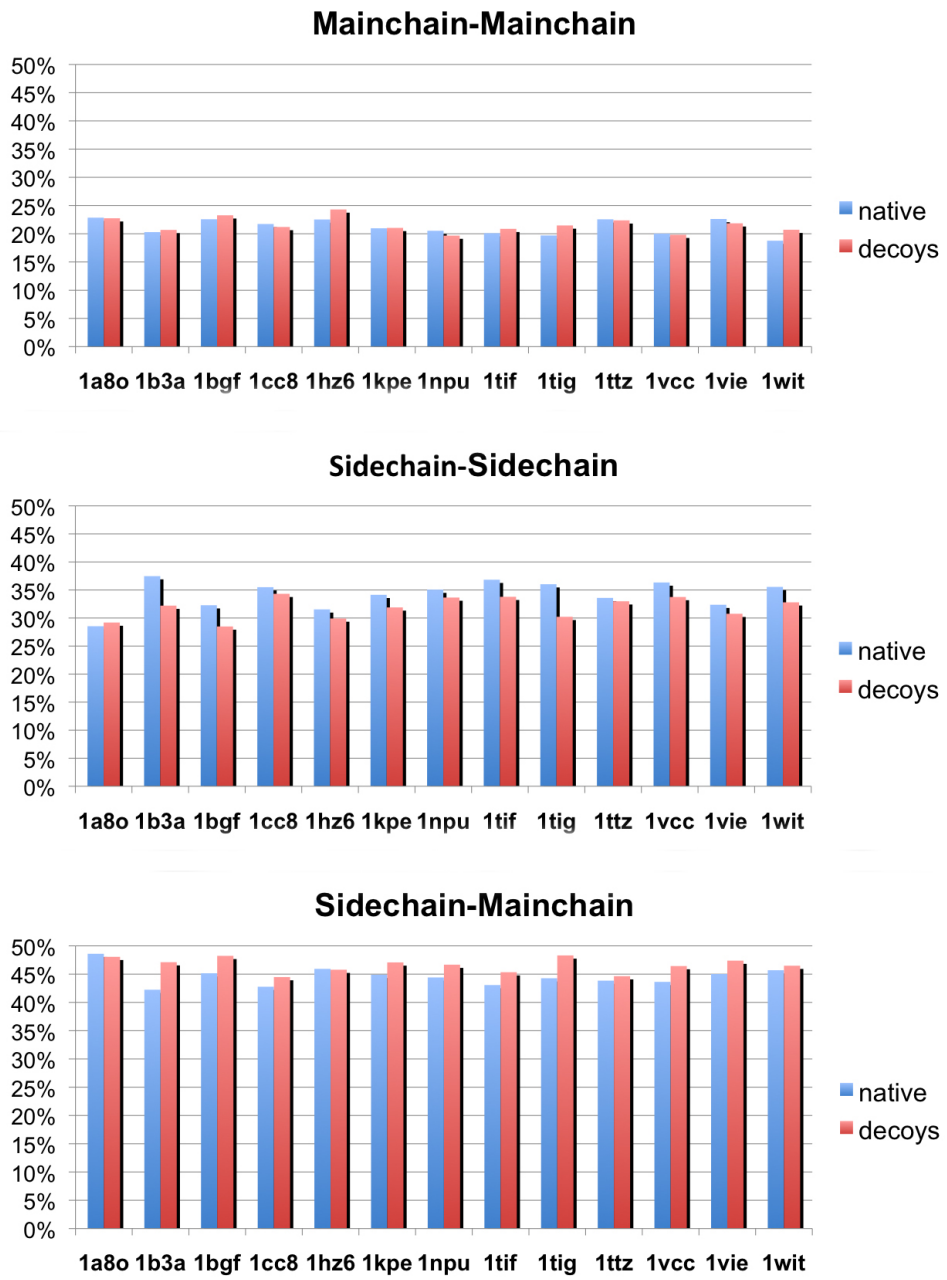

FIGURE 7.10: Decoys and natives have similar all-atom packing. The percentage of Probe dots that fall into each of the three major categories above are similar for decoy models and native structures.

7.3.2 Failed arginine rotamer predictions

The second data set, provided by Chris King, exhibited Rosetta failures that were inherently local in nature. For each individual example in a large set of arginines at protein-DNA interfaces, Chris substituted each rotamer in Rosetta's library and energy-minimized its χ angles in the context of the fixed protein and DNA surroundings. He provided us with a list of 13 examples for which the lowest-energy conformer derived from this procedure, the "repacked" conformation, did not coincide with the native rotamer well. As a point of comparison, for each case he also supplied coordinates for the "minimized" conformation, resulting from energy minimization of the χ angles of the native rotamer (i.e. without explicit rotamer sampling to escape the native rotamer well). My motivation for examining these cases was to determine whether some aspect of the Rosetta energy function caused the error, or whether instead the native sidechain had the wrong conformation in the crystal structure or actually had multiple conformations.

Electron density was available from the EDS for only 6 of these 13 examples. In 5 of those 6 examples Rosetta's neglect of ordered waters appeared to be at fault, and in 4 of the 6 its drive toward higher rotamericity at the expense of H-bonds also contributed to the error. Thus in almost all cases Rosetta prefers a statistically more common rotamer at the expense of H-bonds to ordered waters and even polar groups on DNA bases. A representative example can be seen in Figure 7.11.

In the remaining case, 1bl0 Arg46, a steric clash to the DNA indicates that the native rotamer is indeed misfit in the crystal structure (Figure 7.12). The weak electron density suggests that multiple conformations may be possible in this region, though the deposited structure is not one of them. Rosetta's minimized version of the native rotamer cleans up this clash by slightly tweaking its χ angles and fits the density as well as anything could in this region; it is a rather reasonable conformer.

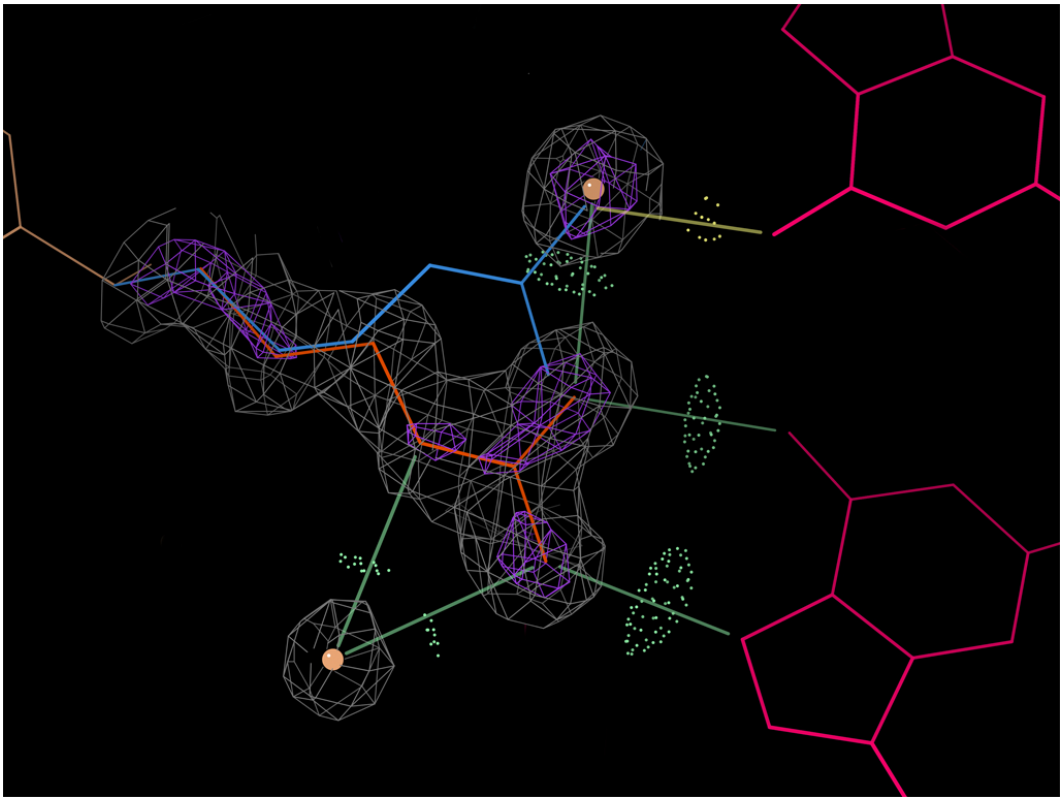

FIGURE 7.11: Rotamer prediction failure at an arginine-DNA interface. The native Arg42 in 1zs4 (orange) forms multiple H-bonds (light green dots and lines) to the adjacent DNA (pink) and well-ordered waters (peach balls). This deposited conformation is well supported by the 1.2 σ (gray mesh) and 3.0 σ (purple mesh) electron density. The "repacked" Rosetta rotamer (blue), predicted in the absence of the waters but in the presence of the DNA, instead chooses a more common rotamer (82.1% instead of 45.2%) that eschews the native H-bonds, allowing only one weak H-bond (light yellow dots and line) to a different DNA base.

Rosetta's repacked conformer, on the other hand, reaches up to the next DNA base and forms nice H-bonds while remaining clash-free, but fits the density perhaps less well. Interestingly, in this case the repacked conformer has a lower rotamericity than the native sidechain, 18.2% instead of 39.4%. Ultimately, it seems possible that both the minimized and repacked conformations coexist in solution, but there is insufficient evidence to be sure. If anything, the results from the other cases cast doubt on the repacked conformation.

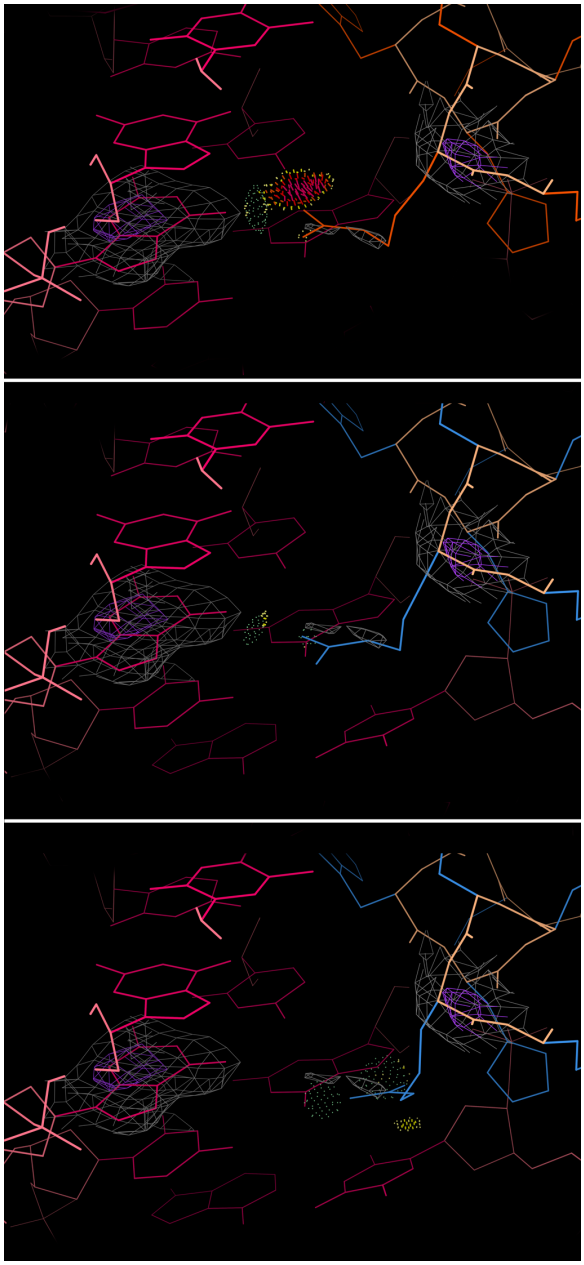

FIGURE 7.12: Possible rotamer prediction success at arginine-DNA interface. Top: The deposited Arg46 in 1bl0 (orange) forms an H-bond (green dots) to an adjacent DNA base (pink), but also clashes (pink spikes) to a second base. The 1.2 σ (gray mesh) and 3.0 σ (purple mesh) electron density supports the DNA and protein backbone conformations well, but is weak for the Arg sidechain. Middle: The Rosetta minimized sidechain (blue) preserves the original H-bond, forms a new small H-bond to the second base, eliminates the clash, and fits the density approximately equally well. Bottom: The Rosetta repacked sidechain (blue) forms two H-bonds to the second DNA base, but is slightly less well supported by the density.

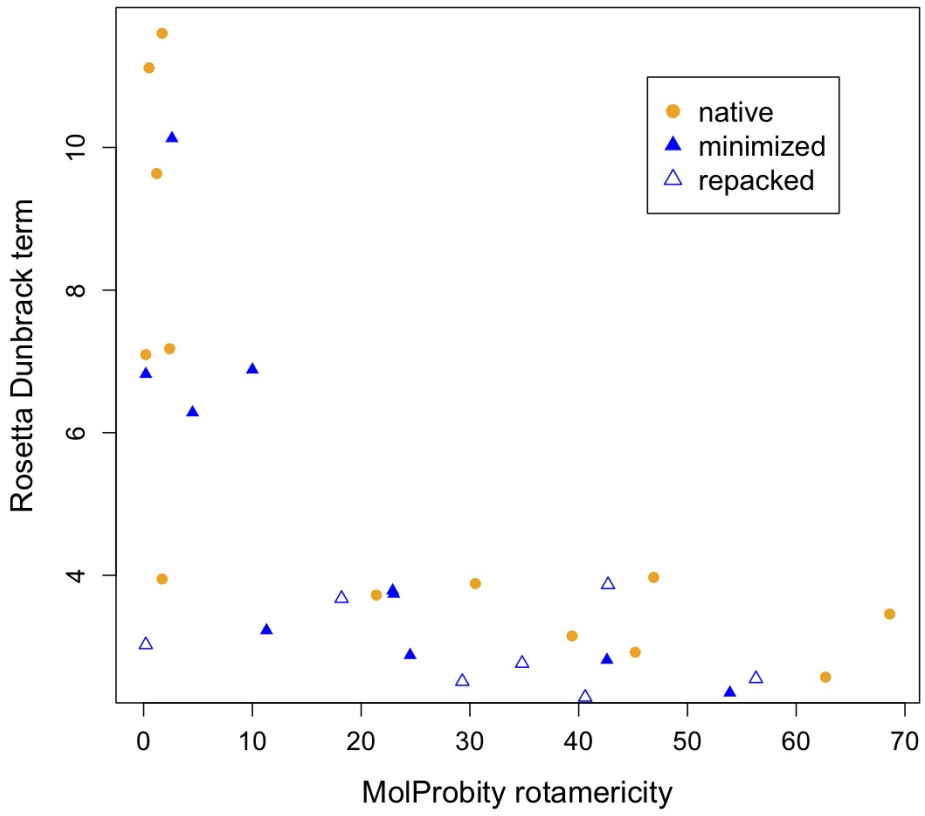

FIGURE 7.13: Relationship between Rosetta and MolProbity rotamer terms for arginine predictions at DNA interfaces. The energy-like Rosetta term is related to the probability-like MolProbity rotamer term by roughly a negative logarithmic transformation. Examples from the native, minimized, and repacked subsets all follow the same trend.

The root causes of misprediction appeared to be similar in the cases for which electron density was not available, but it was difficult to be sure given the lack of direct experimental corroboration of the native structure.

Conclusions: False Rosetta minima

I spent a not insignificant amount of time investigating these two sets of disappointing models, one with wrong global folds and the other with wrong arginine rotamers, and ultimately decided that Rosetta does two things noticeably wrong.

First, it over-values "rotamericity" (i.e. rotamer probability) relative to forming good hydrogen bonds. (The energy-like Rosetta rotamer term, called the Dunbrack or `fa_dun` term, is related logarithmically to the probability-like MolProbity rotamericity term (Figure 7.13); the relationship is imperfect because the Rosetta term is based on different χ angle distributions (Shapovalov and Dunbrack Jr, 2011) than we use in MolProbity.) This phenomenon highlights a fundamental limitation of a hybrid energy function incorporating both physics terms based on local interactions (e.g. van der Waals, H-bonds) and statistical/empirical terms based on global statistics (e.g. rotamer and Ramachandran probabilities). The problem is that statistical terms drive individual examples to adhere to the most common global features in spite of local, context-specific effects. For example, in the rotamer predictions discussed above, H-bonds were undervalued as individual sidechains gravitated to statistically more populated regions of χ angle space.

One solution may be to impose a constraint favoring a protein-like *distribution* of rotamers; a downside is that this approach would require "communication" between non-interacting sidechains during the simulation, which defies real-world intuition and could be computationally difficult. Another solution may be to down-weight statistical terms when local physics terms are pertinent, e.g. when multiple potential H-bond partners are present, rather than predetermining relative global weights for the statistical vs. physics terms (as is typically done).

Of course, given that the interactions that dominate protein energetics (van der Waals forces, hydrogen bonds, electrostatics) fall off relatively rapidly as a function of interatomic distance, it would be desirable for statistical properties like rotamer probabilities to simply "emerge" from simulations. However, statistical terms are arguably necessary for capturing subtle energetic effects that would be difficult or computationally expensive to model otherwise. Given this necessity, the question remains how to best coordinate the information content from these disparate sources.

Second, Rosetta mispredicts many local conformations because its implicit solvent model cannot place well-ordered, structurally integral waters. This problem can be more straightforwardly addressed – in principle, at least – by modeling waters explicitly instead of implicitly. Indeed, members of the Rosetta community have defined "solvated rotamers" in an attempt to address this problem (Jiang et al., 2005). Despite this study's admirable pioneering spirit, its implementation of solvated rotamers relied on defining large numbers of rotamer variants with different waters "attached" to sidechain polar groups, which results in combinatorial difficulties for most practical applications. A later study took an alternative approach by explicitly coupling waters to chemical groups on a ligand of interest instead of to sidechains in the main protein, and also incorporating a provably accurate algorithmic framework (Huggins and Tidor, 2011). However, this approach is specific to modeling applications involving ligands, such as drug design. Ultimately, additional work will be necessary to generically and accurately model water molecules near the protein surface, where they transition from disordered bulk solvent to being integral components of stable and unique protein structures.

7.4 Predicting linchpins: critical structural checkpoints for folding

In Section 7.2 above, aggressive conformational sampling seeded with native and homologous fragments was necessary to identify the near-native region of conformational space. In general, for a prospective structure prediction problem, massive amounts of sampling may be needed to fortuitously stumble upon this low-energy basin. However, as members of the Baker lab recently discovered (Kim et al., 2009), certain specific "features" – individual torsion bins, secondary structure assignments, or residue-residue β-strand pairings – are often bottlenecks to Rosetta structure prediction: when just one of these features is constrained to its native value, the computational time needed to identify the native energy well is reduced by many orders of magnitude. In many cases these features have slightly unusual or strained geometry and localize to functional regions or regions experimentally known to form late in folding; the authors therefore speculate that Rosetta simulations may capture some aspects of real-life protein folding.

Unfortunately, it is infeasible to commonly sample unusual features throughout a protein during prediction: because most regions adopt more probable local structures, so the overall process would be hampered. However, if it were possible to identify regions that may house a linchpin features before or during a simulation, *de novo* structure prediction could be vastly accelerated, marking a major step toward robust prediction of protein structures directly from sequences. In our view, such an ability can best be obtained from detailed visual analysis and local structure validation tools. To that end, for several examples from the paper (Kim et al., 2009), I examined low-energy models with relatively low global RMSD to the native structure, but lacking linchpins – in other words, models that were close but "not quite there". My goal was to be able to identify specific problems in these regions, and by extension to be able to suggest to Rosetta more precisely when and where unusual

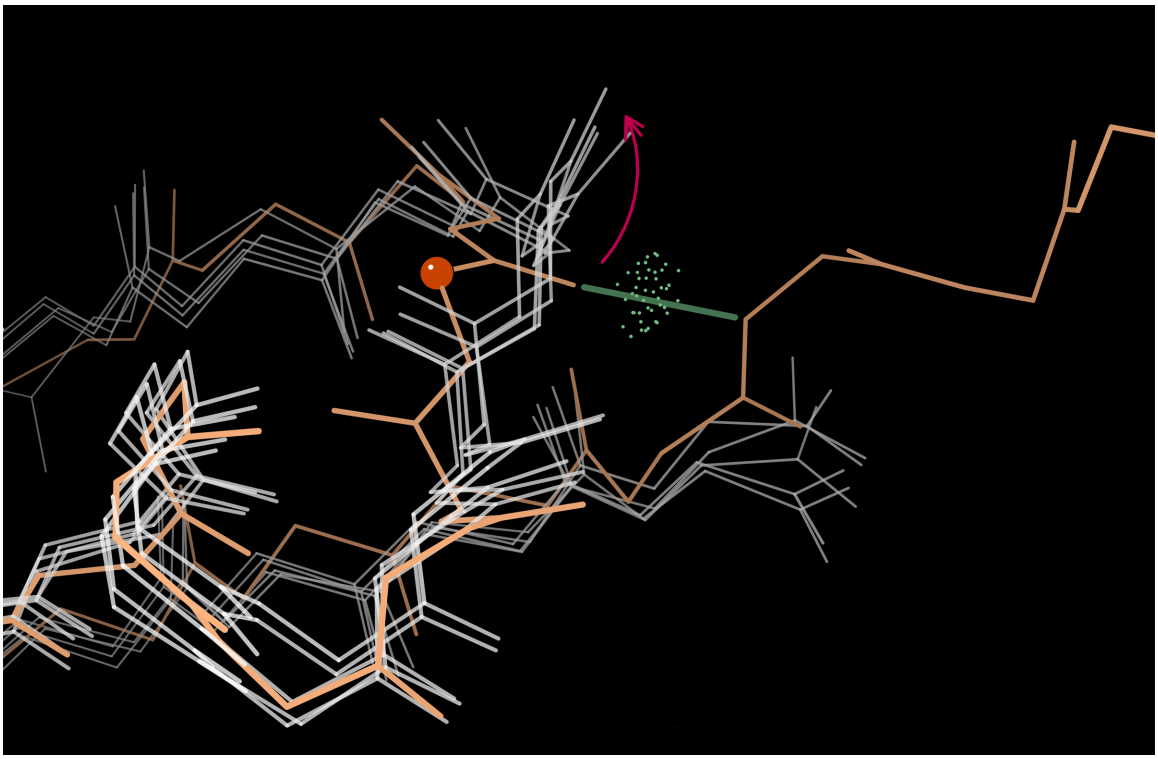

FIGURE 7.14: A truncated C-terminus leads to a false Rosetta folding bottleneck. Asp53 (peach ball) in 1pgx (peach) forms a mainchain-mainchain H-bond via its carbonyl to the amide of Met70. In the Rosetta models (white), on the other hand, the Asp53 carbonyl not only lacks its H-bonding partner, but also is likely repelled by the presence of a negatively charged truncated C-terminus *in silico*.

features should be given a fair shake.

In some cases I studied, the linchpin feature was rarely sampled for a trivial reason that was not biologically interesting. For example, Rosetta seldom samples the β ϕ,ψ bin for 1pgx Asp53, but that is simply because its H-bond partner, Met70, was trimmed during the simulation (Figure 7.14). In fact, the carboxyl group simulated at the *in silico* C-terminal residue, Glu69, likely *repels* the carbonyl of Asp53, causing it to rotate away by about 90° by adopting different ϕ,ψ. It should be noted that the decision to trim the tail was perfectly reasonable since it's surely disordered in solution, but one or two more residues should have been included to properly study this case.

In other cases, the region in question was modeled correctly, and the linchpin feature turned out to be quite thought-provoking. For example, 1di2 Ala134 adopts α ϕ,ψ to form a β bulge on a solvent-exposed edge strand; such irregularities are part of nature's way of avoiding aggregration via "negative design" (Richardson and Richardson, 2002). In contrast, Rosetta's models for 1di2 put the bulge at one end of the edge strand, and maintain highly regular β structure throughout the middle of the strand (Figure 7.15). A trained human eye would immediately judge the resulting edge strand to be too exposed and susceptible to pairing with an unwanted partner. One could imagine a heuristic that recognizes such features and attempts to translate the bulge more toward the center of the edge strand – even with a low success rate this could be useful. Of course, codifying "dangerous edge strand" is difficult, but strands with low curvature around the axis perpendicular to the strand (i.e. too straight) and/or low twist along the axis parallel to the strand (i.e. too flat), and also having a "regular" H-bond pattern between the edge strand and its partner, would often warrant concern.

Another interesting example was found in that most famous of computational protein targets, ubiquitin. (This protein had two linchpins, but they were sufficiently distal from one another that I treated them independently.) Rosetta models with the native β ϕ,ψ feature from 1ubi for Arg54 (8/10) matched the native 51-54 loop, but those with α ϕ,ψ (2/10) adopted a different loop conformation instead. Interestingly, the sidechains of models with the linchpin followed the loop trace of models without the linchpin and, to a lesser extent, *vice versa* (Figure 7.16). In other words, the two clusters of models were related by a "sidechain-mainchain swap".

Unfortunately, I was unable to find any distinguishing characteristics of the models without the linchpin that would flag them as needing some sort of change. Even if it were possible to define some such set of criteria, models meeting them may not always require sidechain-mainchain swaps specifically as opposed to some other type

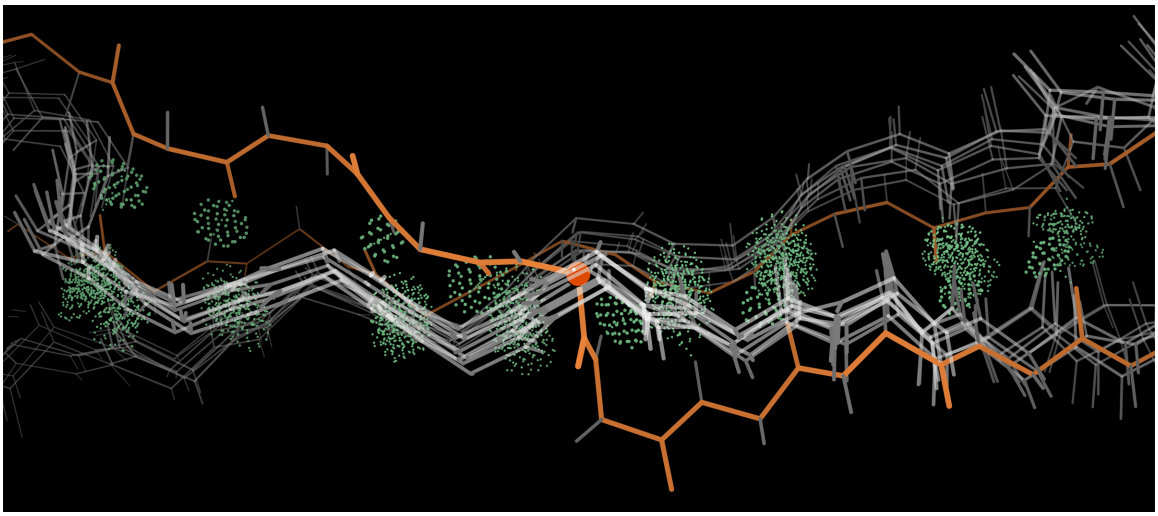

FIGURE 7.15: A β bulge as a Rosetta folding bottleneck. Ala134 (peach ball) in 1di2 (peach), in the middle of an edge β strand, adopts ϕ,ψ in the α region to form a β bulge. This residue fails to adopt the necessary torsions in most Rosetta folding simulations (white), instead maintaining regular β structure with regular β-sheet H-bonding (green dots) along the entire strand. View is from perspective of solvent.

of structural operation – that relationship here may be merely a coincidence.

Nevertheless, the general idea of sidechain-mainchain swaps is certainly intriguing, and may merit future investigation. For example, a new method for discovering diverse and potentially low-energy loops may be to search for ϕ,ψ changes that would redirect a model's backbone along a path already charted by its sidechain in the original conformation. Note that similar phenomena occur in natural proteins: N-caps (Section 2.3.1) and pseudo-turns (Section 3.1) use sidechains to replace backbone-only interactions in a type of structural mimicry. Certain misfittings at chain ends in crystal structures (e.g. Figure 3.3) are also reminiscent of this relationship among models produced entirely computationally.

In summary, I have proposed targeting enhanced sampling of rare features to specific regions such as overly exposed edge strands and potential sidechain-mainchain swap/branch points. It may be possible to test these ideas' utility by blind trials of

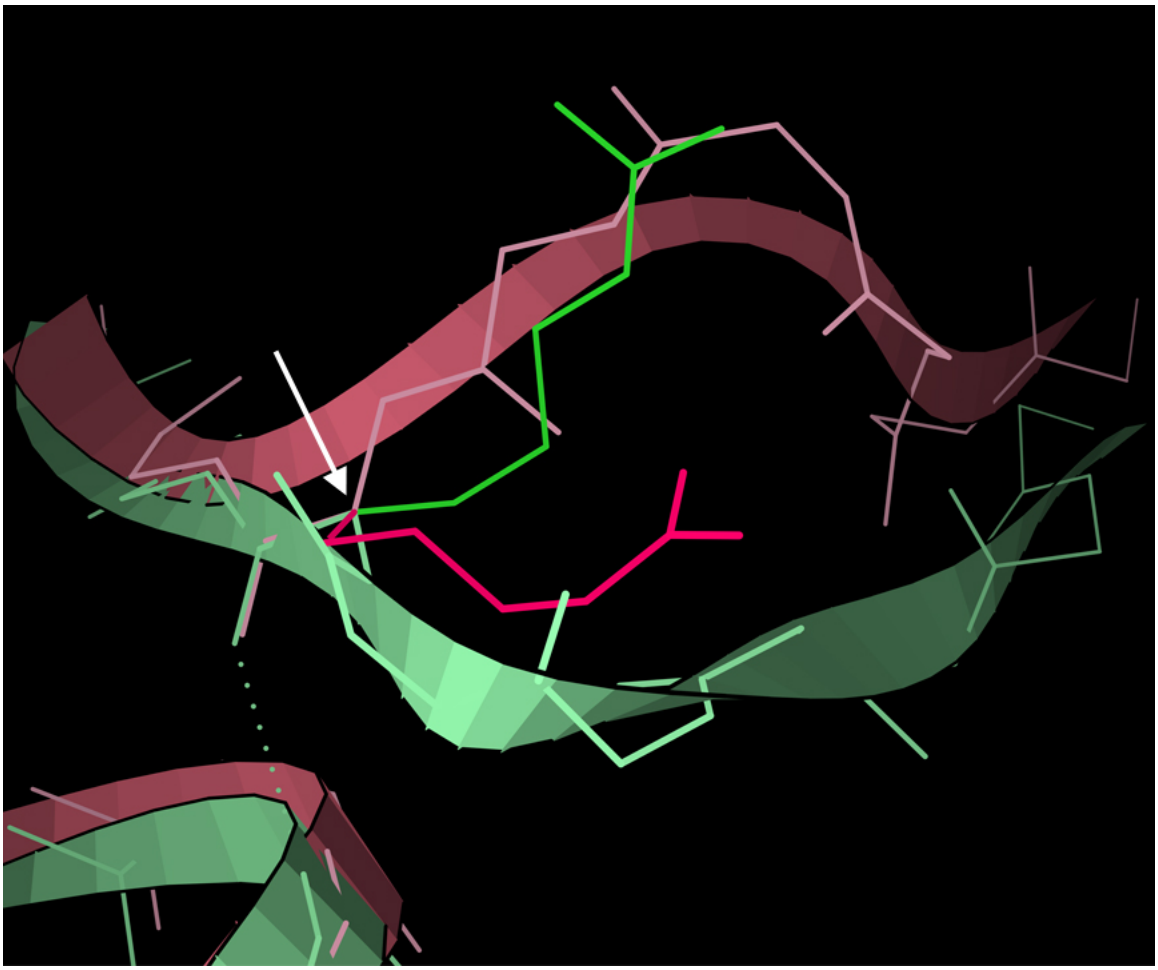

FIGURE 7.16: Sidechain-mainchain swap between Rosetta models with and without linchpin features. The mainchain of a Rosetta model for ubiquitin (1ubi) without the Arg54 β ϕ,ψ linchpin (pink) overlaps with the sidechain of a Rosetta model with the linchpin (green), and to a lesser extent *vice versa*, after co-centering (Block et al., 2009) (white arrow) on the Arg54 Cα.

discriminating regions that require linchpins vs. those that don't in larger data sets of Rosetta models. Ultimately, though, it may be necessary to implement them in Rosetta and empirically measure the speed-up in identifying native-like conformations.

7.5 Investigating the origins of strand-swaps in β-sheet designs

De novo design of protein folds – i.e. designing "from scratch" instead of modifying an existing protein – has seen various high-profile successes over the past 25 years (Hecht et al., 1990; Quinn et al., 1994; Harbury et al., 1998; Kuhlman et al., 2003). *De novo* design is epistomologically valuable because we often learn best by doing – only when we attempt to mimic nature by building new proteins are we brought face-to-face with the most fundamental underlying design principles.

In a final collaboration with the Baker lab, I examined *de novo* designed Rossmann folds that successfully adopt the desired global fold except for an unexpected central β strand swap, as seen in NMR structures of the designed sequences solved by the Montelione lab (Figure 7.17). The Rossmann architecture builds "outward" from the N-terminus, loops back in to the middle of the final protein, and proceeds "outward" again in the opposite direction, all the while alternating strands and helices. Thus the first strand of the N-terminal subdomain and the first strand of the C-terminal subdomain represent the junction point of the two halves of the final full sheet. Befuddlingly, these are the two strands that swap positions in 3 of the 4 designs (see Table 7.1).

This mysterious result indicated that we may be missing something fundamental about this class of structures. With this motivation, I used a cadre of structure validation and visualization tools and pursued several ideas in an attempt to identify a (singular) root cause of these bafflingly consistent swaps.

Degeneracy in branched-Cβ sidechains in central sheet

The prevalence of branched-Cβ sidechains (especially valines) in the sheet (Figure 7.18) probably facilitates a strand swap, since different strand pairings produce similar contacts. This sequence degeneracy lowers the activation barrier for a swap.

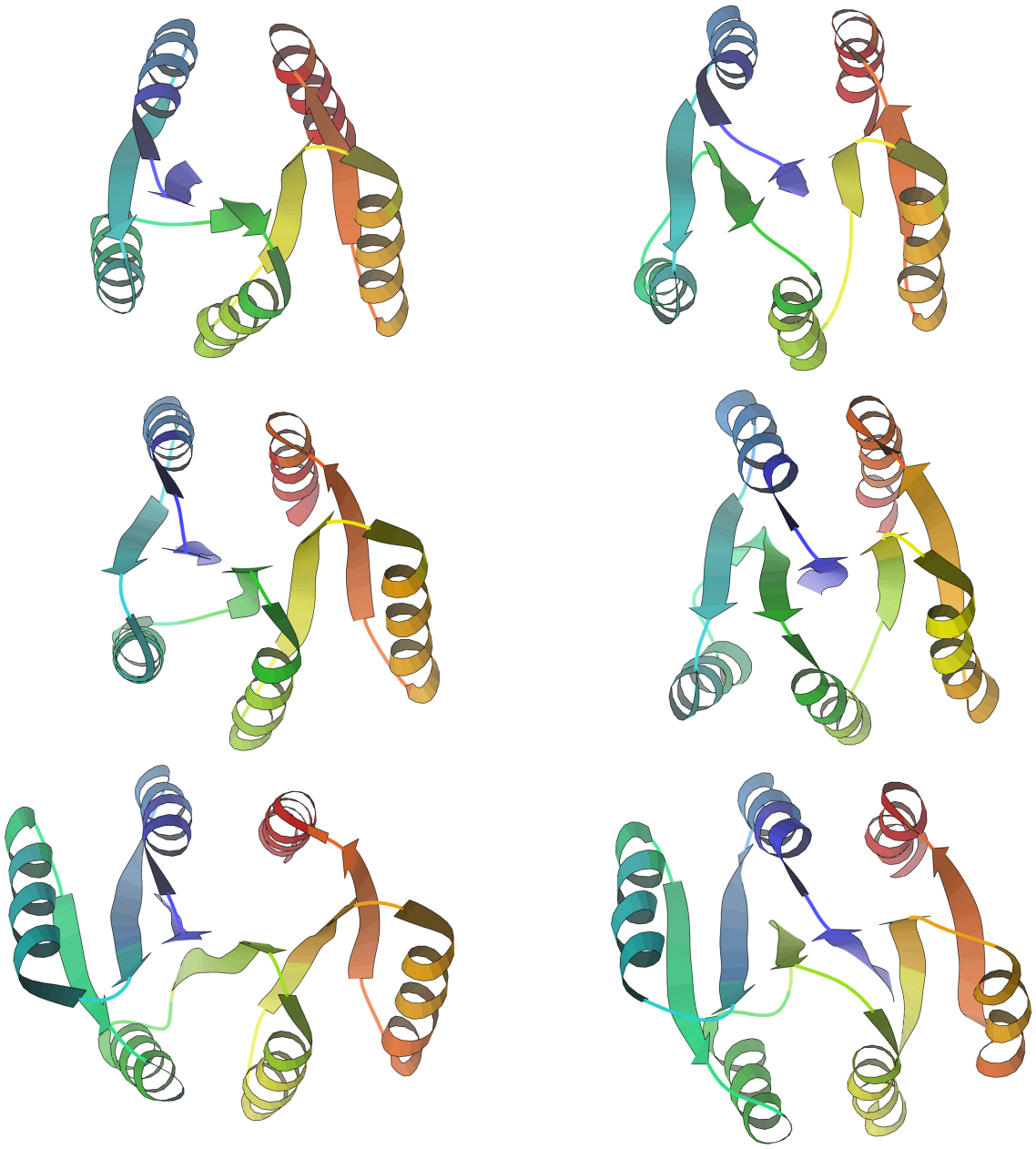

FIGURE 7.17: Flummoxing strand swaps in *de novo* designed Rossmann folds. Top: r2x3v1. Middle: r2x3v2. Bottom: r3x3. Left: The design models have classic Rossmann folds with the first strand (blue) to the left (from this perspective) of the third (fourth for r3x3) strand (green). Right: In model 1 of the NMR structures of these designs, those two strands swap places, but the rest of the protein is very close to its intended conformation. The r2x2 model (not shown) adopts the desired unswapped conformation.

Table 7.1: Descriptions of strand-swapping *de novo* Rossmann designs.

Design	NMR PDB code	Description	Swap
r2x2	2kpo	4-strand design	(no swap)
r2x3v1	2l69	first 5-strand design	β1-β3 swap
r3x3v2	n/a	second 5-strand design	β1-β3 swap
r3x3	n/a	6-strand design	β1-β4 swap

However, enhanced "swappability" doesn't say why a swap does occur, only why it can occur; an additional hypothesis was needed to explain why the swapped conformation is actually strongly preferred.

Note that a strand swap in the 4-stranded fold (r2x2) would constitute a much larger architectural change than in the 5-stranded (r2x3) or 6-stranded (r3x3) folds, in the sense that not just some but all $\beta\alpha\beta$ units would be disrupted. Edge strands are clearly distinguished by their hydrophilicity; since there are only two interior, hydrophobic strands in the r2x2 topology there is only one possible strand swap and it would affect all core interactions. This may be why r2x2 avoids a swap, adopting the desired conformation in the NMR structure.

Overly idealized rotamers

One readily apparent feature of the design models, especially for long sidechains, is that they over-value common rotamers. For example, many lysines switch from extended rotamers (with all or mostly trans χ dihedrals) pointing toward solvent in the design models to less extended rotamers making intramolecular interactions in the NMR structures (Figure 7.19). (The effect was less pronounced for other sidechains like arginine that aren't as "statistically simple" (Lovell et al., 2000); lysine rotamers are perhaps most emblematic of the general problem of statistically driven over-idealization in Rosetta.)

Apparently the energy function favors the rotamericity bonus of a "good" rotamer

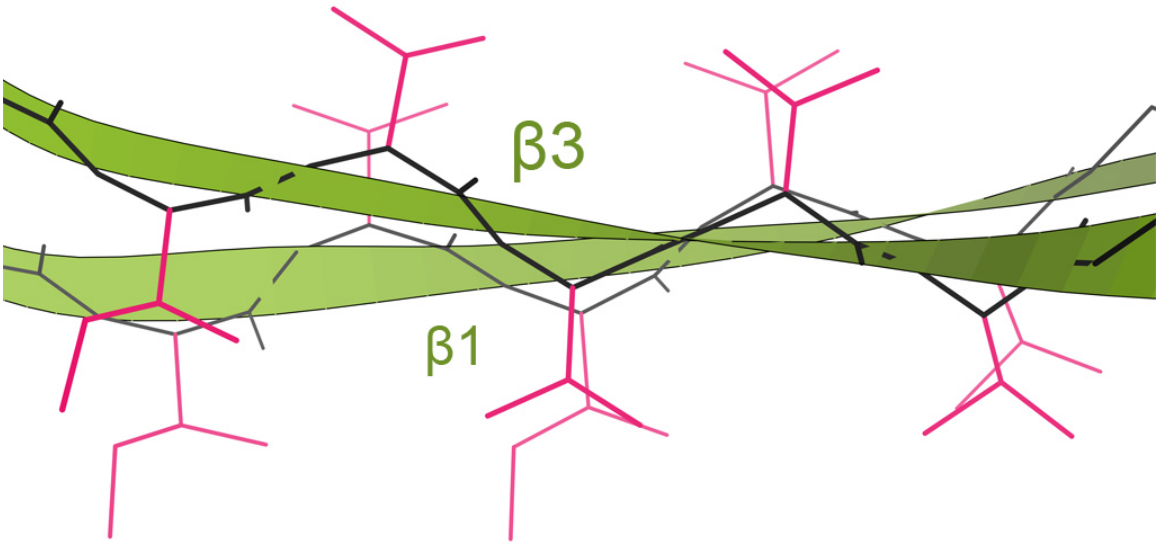

FIGURE 7.18: Over-representation of the branched-Cβ sidechains Val and Ile in the first and third strands of the design model of r2x3v1. These two strands swap in the NMR structure.

too much relative to charge-charge interactions or hydrogen bonds. Over-damping of charge-charge interactions by implicit solvation may play a role. This phenomenon has also been observed in past Rosetta models we have analyzed (see Section 7.3).

Although overly idealized lysine rotamers clearly occur quite often in these models, it is unclear by what general mechanism they could possibly be causing the observed strand swaps. That said, two specific cases offer some possible mechanistic insight involving unexpected attractive interactions (see Figures 7.20 and 7.21) – or, more colloquially, "ionic bondin' for your moronic ponderin'" (Deltron, 2000). However, these particular amino acids are not common to all the designs, so the effect cannot be a general explanation for the swaps.

Table 7.2: Overly idealized lysine rotamers in strand-swapping Rossmann designs.
* Lysine sidechains in an extended rotamer in the design, but in a non-extended rotamer with one or more intramolecular H-bonds in the NMR structure.
** Lysine sidechains in a non-extended rotamer with one or more intramolecular H-bonds in the design, but in an extended rotamer in the NMR structure.

Design	Forced Outward *	Forced Inward **
r2x3v1	46% (6/13)	0% (0/13)
r2x3v2	40% (8/20)	15% (3/20)
r3x3	17% (2/12)	8% (1/12)

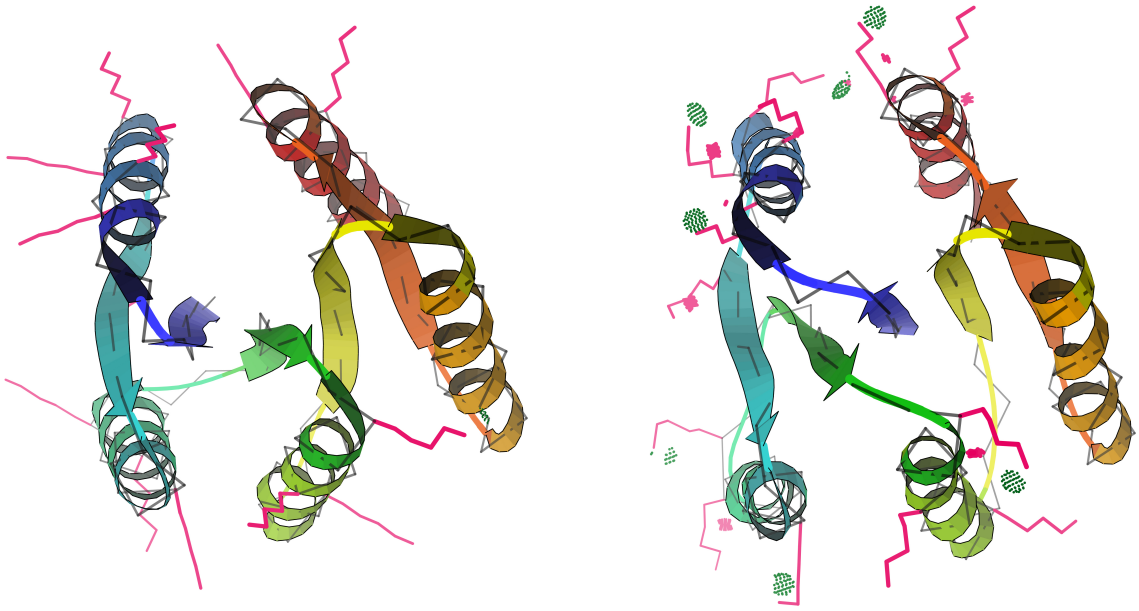

FIGURE 7.19: Overly idealized lysines in a strand-swapping Rossmann design. The design model for r2x3v1 (left) has 13 lysines, almost all of which adopt extended rotamers and extend into (implicit) solvent. In model 1 of the NMR structure (right), however, many of these lysines "tuck in" and form intramolecular H-bonds (green pillows).

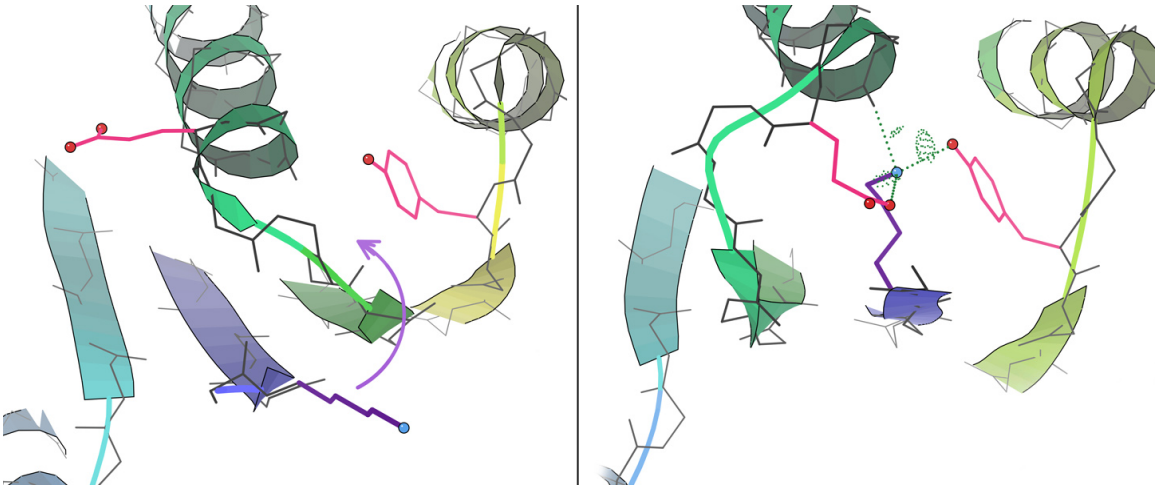

FIGURE 7.20: An N-terminal lysine forms unexpected H-bonds at the strand-swap locus. Lys2 (purple) in r2x3v2 is designed to extend toward solvent, but folds in to make hydrogen bonds and ionic interactions in the NMR structure. At the same time, strand 1 (blue) and strand 3 (green) swap positions. Notably, Lys2 can only access the carbonyls in the C-terminal turn of helix 2 (top of images) for hydrogen bonding when the α2-β2 loop moves due to the strand swap.

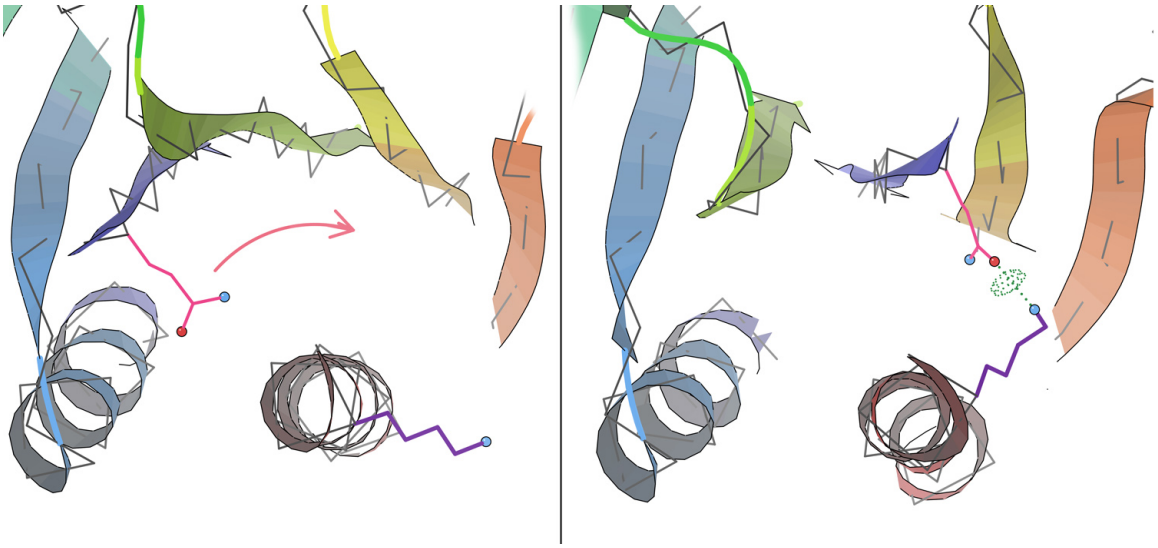

FIGURE 7.21: A C-terminal lysine forms an unexpected H-bond near the strand-swap locus. Lys150 (purple) in r3x3 is designed to extend toward solvent, but folds in to make a hydrogen bond with Gln2 (pink) in the NMR structure. For this to occur, strand 1 (blue) and strand 4 (yellow) must swap positions; meanwhile the flanking strands remain in place.

Kinetic trap caused by unnatural secondary structure propensities

With other explanations, the swapped state must have a lower free energy of folding than the designed state for the swap to actually occur. However, the swapped state could also be the dominant species if an unusual folding pathway funneled it in that direction. Indeed, Rosetta *ab initio* structure predictions of the designed sequences, performed by collaborators in the Baker lab, provided evidence in favor of this idea: the design model's energy basin, obtained in simulations restrained only by chemical shifts, was definitively lower in computed energy than the strand-swapped structure's energy basin, which required additional NOE restraints to robustly identify (Figure 7.22). Therefore, I investigated the possibility that a "kinetic trap" mechanism could explain all three strand swaps.

First, I identified a diverse set of real Rossmann structures with Dali (Model, 1996) (using r2x3v1 and r3x3) and SCOP (Murzin et al., 1995) searches. I then computed average sequence-based PSIPRED (McGuffin et al., 2000) secondary structure prediction confidence levels (from 0 to 9) for each secondary structural element (SSE), strand, or $\beta\alpha\beta$ unit for the design models and natural structures. For the latter, only the residues corresponding to actual secondary structure in the design model – based on a structure-based sequence alignment – were used. If the prediction matched the DSSP-determined (Kabsch and Sander, 1983) secondary structure designation, the confidence level was used for the average. If it did not match, half that number was used instead; this is because PSIPRED gives a confidence level for each residue only for the top prediction (H for helix, E for extended, or C for coil), so if the prediction is wrong, one must assume that its confidence for the true secondary structure type is somewhere between 0 and its incorrect confidence level. To the extent that secondary structure propensities correlate to rates of folding, these averages predict (or serve as proxies for) the relative rates of folding of different protein regions.

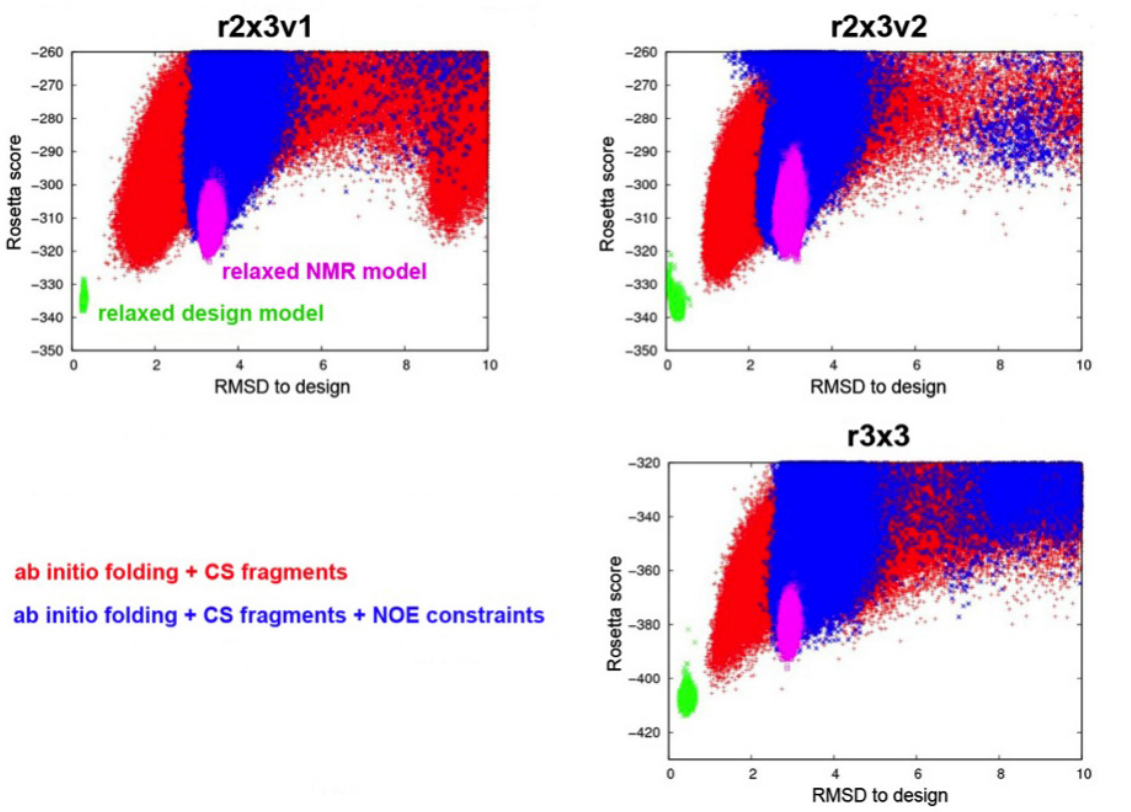

FIGURE 7.22: Rosetta simulations reveal an energy gap between the swapped and unswapped states. For each designed sequence, the strand-swapped NMR structure relaxed (i.e. energy-minimized) in the Rosetta energy function (magenta) had higher energy than the design model relaxed in the Rosetta energy function (green). This energy gap was confirmed by CS-Rosetta (Shen et al., 2008) *ab initio* structure prediction simulations, which use chemical shift restraints to reduce the amount of conformational sampling required. Simulations with only chemical shifts (red) produced conformations very similar to the design model. To reach conformations very similar to the NMR structure, it was necessary to supplement the chemical shifts with additional NOE restraints (blue). Even then, the computed energy of the swapped state was higher. Each panel plots Rosetta energy (y-axis) against Cα RMSD to the stated design model (x-axis). Credit: Nobuyasu Koga (Baker lab).

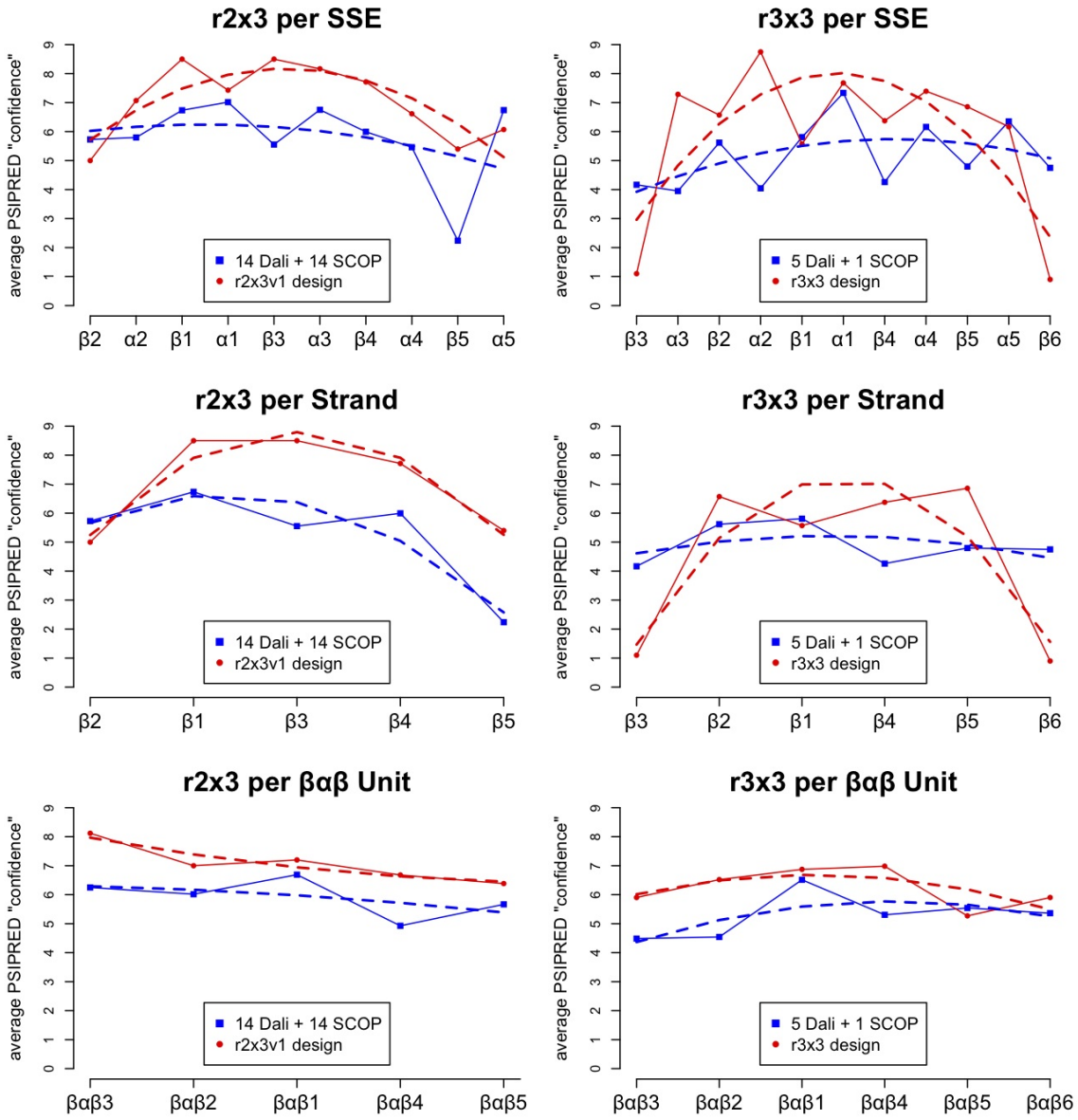

FIGURE 7.23: Average PSIPRED confidence level per structural unit in natural vs. designed Rossmann folds. These propensities are grouped in three ways: by SSE (secondary structural element) (top), by β strand (middle), and by $\beta\alpha\beta$ unit (bottom). The x-axis is in structural order (edge strand to edge strand) instead of sequence order (N- to C-terminus) to better reflect the topology of the Rossmann fold. The r2x3v1 design is used to represent the r2x3 topology. Dotted lines are polynomial fits to the data. Despite some noise, and a possible dip in propensity for natural proteins for r2x3 strand 5, the natural and designed proteins follow quite similar patterns in all cases.

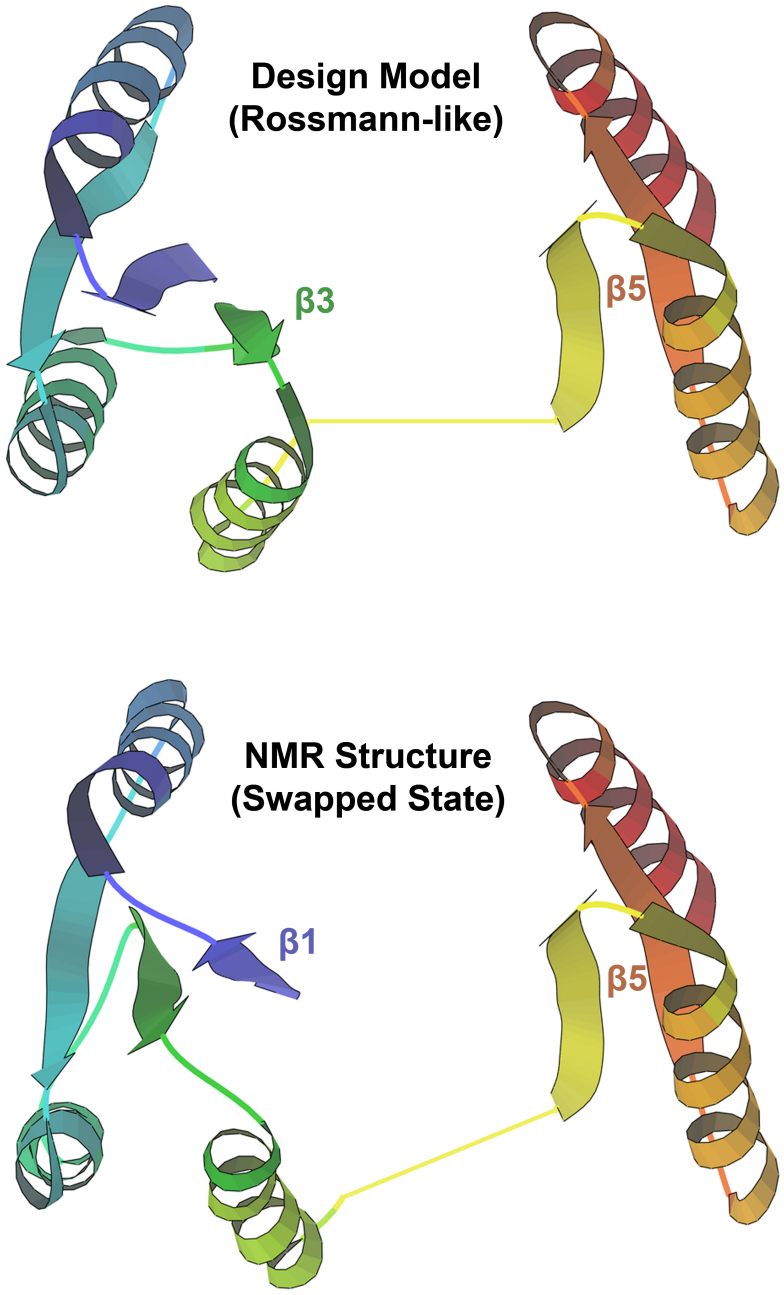

FIGURE 7.24: Putative kinetic trap explanation for strand swaps. Top: If folding of the $\beta 5$ unit (right) is delayed, the $\beta 1$ unit (left) is free to form "naturally". Bottom: If the $\beta 5$ unit folds precociously, $\beta 1$ may be presented prematurely. The two halves may then be pulled together if $\alpha 1$ and $\alpha 5$ (blue and red helices at top of panel) interact; this is at least plausible since helices tend to form faster than sheets.

We initially noticed that for r2x3, the natural structures (but not the designs) are predicted to delay folding of their last strand relative to the other SSEs in the protein (see dip in blue line in top left panel of Figure 7.23). Thus, it seemed possible that in real structures the C-terminal region could be unfolded early, so the N-terminal region would be free to fold independently and correctly (Figure 7.24). On the other hand, in the designs the final strand is predicted to fold just as fast as any other strand. So the C-terminal region could fold precociously and provide a docking interface for the nascent first strand, perhaps facilitated by nascent α1-α5 interactions (assuming that individual helices form relatively quickly), thus pulling β1 into its swapped position next to β4 in the folded C-terminal half (Figure 7.24).

An argument against this hypothesis is that the swapped structures have higher contact order (Plaxco et al., 1998) – i.e. average sequence separation between contacting residues – than the designed models, so the folding of the swapped structures should be slower in general. Also note that proteins are synthesized from the N-terminus, so the folding of the N-terminal half tends to occur first; this may mean that the natural proteins do not in fact need to delay folding of their C-terminal halves since the N-terminal halves will fold first anyway. Furthermore, there is no such drop in propensity for β5 in real r3x3 folds, so this explanation (if true at all!) cannot be universal.

Finally, taking the PSIPRED data more generally (Figure 7.23), the trend for SSEs and strands for both natural and artificial structures resembles an inverted parabola: secondary structure propensities are low at the edges and high in the middle. The $\beta\alpha\beta$ trend is also similar for both natural and artificial structures, though this time it is flatter. Of course, the design confidence levels are higher on average, but this is not unexpected from a Rosetta methodology that attempts to match observed distributions and thus may over-emphasize certain aspects (see Sections 7.2 and 7.3), in this case secondary structure occurrence probability by

residue type. Ultimately, given this context of similar overall propensity profiles, it seems likely that the kinetic trap mechanism involving a fast-folding last strand discussed above is based on an anomalous statistical fluctuation, and the designs probably fold at a similar rate and with a similar series of collapse events as real folds.

Note that fewer real r3x3 structures than real r2x3 structures were available through Dali and SCOP searches. Apparently this is because there are actually many r3x3 folds in the PDB, but they commonly have potentially functional extensions or insertions in their intra-SSE loops, as opposed to the short, minimalistic, utilitarian loops in the designs. Incorporating individual substructures from these structures, e.g. individual $\beta\alpha\beta$ units into the $\beta\alpha\beta$ unit averages (bottom row of Figure 7.23), could provide more data and perhaps reveal a new trend. However, (1) this seems unlikely because the r2x3 data is more ample and nevertheless shows a relatively flat line, and (2) it is possible that extended loops, perhaps even with extra full domains attached, would affect the folding kinetics of the r3x3 units, in a way that would be difficult to account for in my analysis here.

Neglected interactions involving C-terminal His tag and N-terminal Met

All the design models lack a C-terminal His tag (GSLEHHHHHH) that is present in the NMR structures. Furthermore, r2x3v1 and r3x3 lack an N-terminal Met residue that is present in the NMR structures. Interestingly, the N- and C-termini are close in space for both the designed and strand-swapped folds, so these extensions have the opportunity to interact with one another *in vitro* (Figure 7.25). If this interaction is more favorable in the swapped state than in the designed state, the swapped fold may actually be preferred overall.

Also, according to our collaborators in the Baker lab, preliminary results from MD simulations using the NMR model of r2x3v1 imply that the His tag can insert

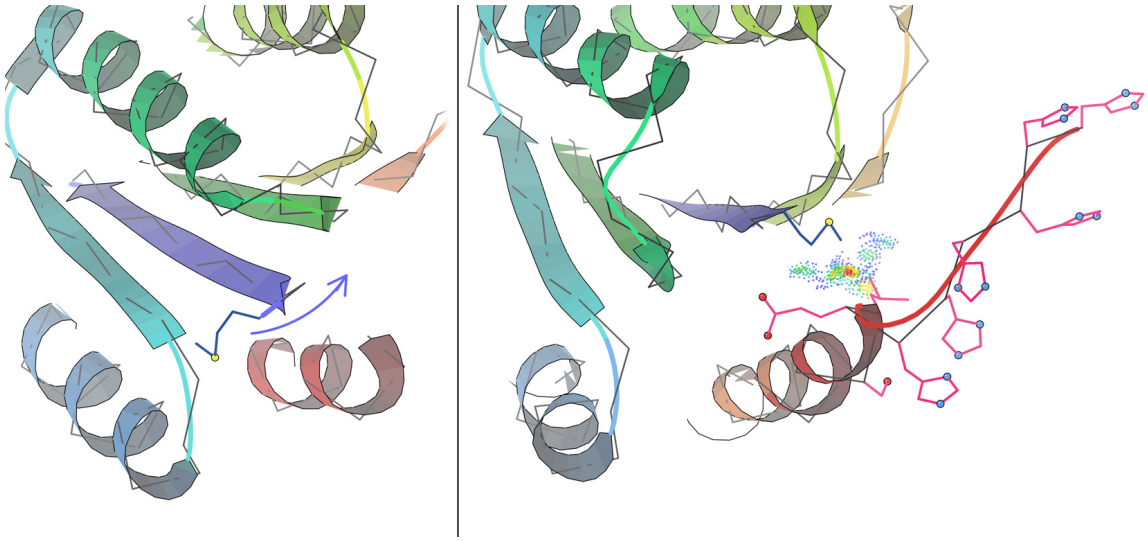

FIGURE 7.25: C-terminal His tag may stabilize swapped strands. In the design model for r2x3v2 (left), there are few interactions between the first strand in its designed position (blue) and the last helix (red). In model 1 of the NMR structure (right), however, the presence of a C-terminal His tag extension (GSLEHHHHHH) adds more turns of helix and introduces additional sidechains (pink), including Glu128, that interact (colored dots) with the first strand, but now in its swapped position relative to the third strand (green).

itself into the hole between the second and third helices, which could potentially stabilize the swapped state.

Members of the Baker and Montelione labs are purportedly testing this idea *in vitro* by using (1) a different C-terminal tag that can be cleaved between protein purification and structure determination and/or (2) an N-terminal tag that would have at least different interactions.

Note that the r2x2 construct used for NMR also has a His tag, yet a strand swap does not occur. However, we believe that a strand swap is fundamentally much less likely for the 4-strand topology for the reason described above, so despite any neglected His tag interactions, it is not surprising that the non-swapped fold is still overwhelmingly preferred.

Conclusions: Strand-swapped designs

After significant effort, I have failed to uncover a fundamental cause for these flummoxing swaps. The truth is probably less satisfying: a number of subtle but consistent factors likely conspire to alter the topology. To the extent that one or more of the above phenomena make significant contributions, the following suggested methodological improvements may prove useful for future designs.

First, it would be good to achieve a more protein-like rotamericity distribution through whatever means are feasible – as mentioned above (Section 7.3), perhaps by imposing statistical terms based on global observations only when more pressing local interactions are absent. However, it should be noted that the high rotamericity observed in Rosetta models illustrates an adherence to physically realistic geometry that is a very positive trait in general!

Second, negative design could be used to avoid undesired states. For example, alternate topologies, including various strand-swapped variants, could be explicitly instantiated and designed against. One limitation of such explicit negative design is that enumeration of all possible undesired states is often combinatorially infeasible; instead, heuristic negative design techniques, which implicitly avoid large numbers of undesirable states, may prove more useful (Fleishman and Baker, 2012). For example, methylene-Cβ sidechains could be interspersed with branched-Cβ sidechains in the core to simultaneously avoid many strand-swapped variants.

7.6 Discussion

Throughout the various projects involving Rosetta laid forth here, my goal was to investigate rare or unexpected conformations using all-atom validation and graphical inspection.

With the energy landscape mapping project (Section 7.2), I determined that some conformational alternatives likely reflected macromolecular reality in the cellular milieu in a way that traditional structural biology cannot, but that others were demonstrably false. With the false minima projects (Section 7.3), I found many more false conformational alternatives identified by Rosetta, and hypothesized energy function deficiencies that may explain the failures. With the linchpins investigations (Section 7.4), I interrogated models that failed to adopt specific *desired* conformations that would seed subsequent collapse events along the folding pathway. Finally, with the strand-swapping design project (Section 7.5), I investigated root causes for an *undesired* conformational change – in some ways the inverse of the first two projects.

Thus, in various ways, all these examinations involved the interplay of a protein's conformational alternatives: conformations it assumes *in silico* and likely *in vivo* but doesn't *in vitro*, conformations it assumes *in silico* but doesn't *in vitro* or likely *in vivo*, conformations it should assume *in silico* but doesn't, and conformations it shouldn't assume *in silico* but does *in vitro*.

The *Zeitgeist* surrounding Rosetta seems destined to intensify as its high-profile successes continue to roll in – yet significant weaknesses remain. A recently published study tells a cautionary tale: using more sophisticated sampling algorithms coupled to highly parallel computer architectures, the authors (Tyka et al., 2012) found lower energies for many of the same proteins from Section 7.2, but also flatter score-vs.-RMSD energy landscapes. These results indicate that there is still room for improvement in both sampling methodologies and scoring functions. Ultimately, the

type of detailed visual and bioinformatic analysis I performed here will be increasingly valuable as a reality check on Rosetta's performance.

8

Conclusions and Future Directions

Taken together, the work described in this thesis illustrates that we're still in a stage of learning from nature rather than modeling biological systems from first principles. Throughout the various examples presented here of validating alternate conformations – performing alternate confirmation, if you will – a unifying theme was the Bayesian-like idea of "protein-likeness". In essence, conformations unlike those commonly observed in real proteins are treated skeptically, and only confirmed as valid if there are significant extenuating factors (H-bonds, van der Waals packing, etc.).

This paradigm will likely be perpetuated in the biological information age we are entering because of continuing high-throughput structure determination, most notably by the Protein Structure Initiative. This influx of new protein structures will provide additional fodder with which to refine our empirically based concept of protein-like features. Simply a greater quantity of structures won't be enough, however: better refinement techniques will be necessary to ensure all pertinent conformations are properly modeled (Chapter 4), and stringent structure validation tools will be critical for ensuring only the highest-quality input data is used for crafting

measures of protein-likeness (Chapter 5). This is where the PDB Validation Task Force (VTF) comes into play: by standardizing metrics of structural quality and making them publicly available alongside the atomic coordinates (and experimental data), the VTF will play an important role in streamlining the deployment of large-scale structural information for more prospective uses in protein engineering and drug design.

Despite substantial efforts at "structural genomics", genome sequencing remains vastly more efficient. There are consequently orders of magnitude more sequence information than structural information, and current technological trends suggest that gap will only widen in the years to come. Some progress can likely be made by carefully mining this evolutionary "fossil record"; indeed, a new algorithm based on sequence covariation alone recently succeeded in determining the global folds of several membrane proteins (Hopf et al., 2012). Although direct structural information on every protein of possible interest would of course be preferable, beggars can't be choosers – and genome sequencing has made a generous donation.

It should be remembered that protein structure modelers resort to such empirical methods as opposed to well-established, higher-level theory such as quantum mechanics only because computational power is limiting. In the longer term, the roadblocks to continuation of Moore's law of exponential growth in processing power may be removed, and/or quantum computers may succeed in revolutionizing computational capabilities. There may then be a shift away from indirect empirical approaches and back in favor of approaches that search for the global free-energy minimum using more fundamental theories of physics and chemistry. Of course, vastly expanded computational power could also/instead be applied to making sense of the even more unmanageable amounts of structure and especially sequence data available at that time.

Yet all of this pertains to the *methods* for evaluating protein conformations; my

work more importantly highlights the relevance of conformational heterogeneity to biology. My studies of alternate conformations that are detectable in crystallographic electron density maps (Chapters 2, 3, and 4) emphasized the point that protein structures are not strictly unique, but rather exhibit minor excursions around the single most populated conformation. In addition to harmonic vibrations or fluctuations of individual bonds or angles within energy wells, it is important to note that many discrete changes (primarily sidechain rotamer jumps) occur. These dynamics are every bit as much a substrate for natural selection as is rigid structure, and thus are likely to be coupled to function in many cases. NMR can be used to study the kinetics of transitions between substates – which may be quite pertinent to function, especially for catalytically coupled heterogeneity – but cannot provide direct structural information about the multiple conformations involved. The surprise to many is that X-ray crystallography, a technique which admittedly is getting a bit long in the tooth, can fill in this gap (provided the data has high enough resolution). In my opinion, studying near-native ensembles in atomic detail will reveal an entire new level on which evolution operates, and will ultimately be key to designing man-made enzymes on par with their natural counterparts (pending improvements in many other areas as well: solvation, polarizability, etc. (Baker, 2010)).

In light of the growing appreciation for the functional relevance of near-native conformational heterogeneity, a related question is the role of conformational entropy in dictating protein stability and binding. Indeed, a recent study of alternate conformations "hidden" in electron density found that conformational heterogeneity disappears upon ligand binding in at least some systems (Lang et al., 2010), which is in line with other work suggesting that protein conformational entropy can contribute significantly to binding free energy (Frederick et al., 2007). The K* ensemble-based protein design algorithm has also had great success by computing ratios of bound and unbound partition functions to capture entropic changes (Chen et al., 2009a;

Frey et al., 2010).

The Rosetta scoring function (Chapter 7) gets a lot of entropic information "for free" – in its implicit solvation, residue-pair, and rotamer probability terms, for example – but that information is convolved with other statistical correlations, making it difficult to isolate the effects of entropy. Because of such convolution, this function has proved sufficient for predicting the structures of many natural proteins (Section 7.2), which thanks to natural selection have large free energy gaps between the native state and states that are structurally somewhat similar, but is less adept at selecting sequences with similarly pronounced folding funnels for the purposes of design (Fleishman and Baker, 2012).

Ultimately these types of statistical/physical hybrid functions may need to eliminate implicit entropic contributions in order to enumerate low-energy conformations that contribute entropically to near-native ensembles. To wit, some have attempted to preserve the hybrid model by carefully subtracting the double-counted interactions in various statistical terms while maintaining the corresponding physical terms (Song et al., 2011). However, this approach is fraught with difficulties: manual intervention is necessary to determine which statistical term is responsible for double-counting which physical term. There are fewer answers for the perhaps more insidious problem of *each* individual residue being held to the standard of a global distribution (Section 7.5) that is based on an essentially comprehensive set of examples (Chapter 5). Frankly, I see no easy solutions – but in many ways the benefits of modeling innumerable complex interactions implicitly and rapidly (albeit imperfectly) outweigh the costs of somewhat obscuring the contributing physicochemical phenomena, and thus for the time being Rosetta's continuing success is ensured.

In the bigger picture, molecular biology is still mostly reductionist, but is undergoing a philosophical transition from documentation to manipulation, from bird-watching to tinkering. Proteins are an inspiring example: despite being made up of

only 20 subunits, they are among the most complex objects in the universe, with nearly infinite potential for chemical function that biological evolution has only begun to explore. Today, the broadly defined field of synthetic biology is in the nascent stages of harnessing natural mechanisms for molecular production and expanding into engineering pseudo-biological systems. The first important steps are being taken now, such as creating variants of existing proteins with altered enzymatic properties. These initial strides put us on a trajectory toward a new world of novel chemical systems. In some sense "artificial", these molecules will owe a debt of inspiration to "natural" proteins, but will have unique capabilities selected on the merit of their benefit to our species. Perhaps, then, the term "protein-like" will one day be rendered obsolete by our improved understanding of the chemical physics of molecules.

Appendix A

Digital resources

This thesis is supplemented by a CD or DVD with "bonus material", including my Java code, PDB coordinate and kinemage graphics files for various studies, raw wet lab data, miscellaneous useful scripts, ..., and a summary file explaining it all. If you didn't receive such a disk, feel free to pester me (daniel.keedy@duke.edu at the time of this writing) or Dave and Jane Richardson (dcrjsr@kinemage.biochem.duke.edu). If you've read this far, you've earned it!

Bibliography

Adams, P., Afonine, P., Bunkoczi, G., Chen, V., Davis, I., Echols, N., Headd, J., Hung, L., Kapral, G., Grosse-Kunstleve, R., et al. (2010), "PHENIX: a comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D: Biological Crystallography*, 66, 213–221.

Allen, F. (2002), "The Cambridge Structural Database: a quarter of a million crystal structures and rising," *Acta Crystallographica Section B: Structural Science*, 58, 380–388.

Anfinsen, C. (1973), "Principles that Govern the Folding of Protein Chains," *Science*, 181, 223–230.

Arendall, W., Tempel, W., Richardson, J., Zhou, W., Wang, S., Davis, I., Liu, Z., Rose, J., Carson, W., Luo, M., et al. (2005), "A test of enhancing model accuracy in high-throughput crystallography," *Journal of Structural and Functional Genomics*, 6, 1–11.

Baker, D. (2010), "An exciting but challenging road ahead for computational enzyme design," *Protein Science*, 19, 1817.

Bell, J., Becktel, W., Sauer, U., Baase, W., and Matthews, B. (1992), "Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of six amino acid substitutions at Thr 59," *Biochemistry*, 31, 3590–3596.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000), "The protein data bank," *Nucleic Acids Research*, 28, 235–242.

Block, J., Zielinski, D., Chen, V., Davis, I., Vinson, E., Brady, R., Richardson, J., and Richardson, D. (2009), "KinImmerse: macromolecular VR for NMR ensembles," *Source Code for Biology and Medicine*, 4, 1–14.

Bondi, A. (1964), "van der Waals Volumes and Radii," *The Journal of Physical Chemistry*, 68, 441–451.

Bouvignies, G., Vallurupalli, P., Hansen, D., Correia, B., Lange, O., Bah, A., Vernon, R., Dahlquist, F., Baker, D., and Kay, L. (2011), "Solution structure of a minor and transiently formed state of a T4 lysozyme mutant," *Nature*, 477, 111–114.

Bradford, M. et al. (1976), "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding," *Analytical Biochemistry*, 72, 248–254.

Bradley, P., Misura, K., and Baker, D. (2005), "Toward high-resolution de novo structure prediction for small proteins," *Science*, 309, 1868–1871.

Branden, C., Tooze, J., et al. (1991), *Introduction to protein structure*, vol. 2, Garland New York.

Chen, C., Georgiev, I., Anderson, A., and Donald, B. (2009a), "Computational structure-based redesign of enzyme activity," *Proceedings of the National Academy of Sciences*, 106, 3764.

Chen, V., Davis, I., and Richardson, D. (2009b), "KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program," *Protein Science*, 18, 2403–2409.

Chen, V., Arendall, W., Headd, J., Keedy, D., Immormino, R., Kapral, G., Murray, L., Richardson, J., and Richardson, D. (2009c), "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, 66, 12–21.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., et al. (2010), "Predicting protein structures with a multiplayer online game," *Nature*, 466, 756–760.

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

Crick, F. et al. (1970), "Central dogma of molecular biology," *Nature*, 227, 561–563.

Cruz, J., Blanchet, M., Boniecki, M., Bujnicki, J., Chen, S., Cao, S., Das, R., Ding, F., Dokholyan, N., Flores, S., et al. (2012), "RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction," *RNA*.

Das, R. (2011), "Four small puzzles that Rosetta doesn't solve," *PLoS ONE*, 6, e20044.

Davis, I., Arendall III, W., Richardson, D., and Richardson, J. (2006), "The backrub motion: how protein backbone shrugs when a sidechain dances," *Structure*, 14, 265–274.

Davis, I., Leaver-Fay, A., Chen, V., Block, J., Kapral, G., Wang, X., Murray, L., Arendall III, W., Snoeyink, J., Richardson, J., et al. (2007), "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids," *Nucleic Acids Research*, 35, W375–W383.

Deltron (2000), "Deltron 3030," Album.

DePristo, M., de Bakker, P., and Blundell, T. (2004), "Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography," *Structure*, 12, 831–838.

DiMaio, F., Terwilliger, T., Read, R., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H., et al. (2011), "Improved molecular replacement by density-and energy-guided protein structure optimization," *Nature*, 473, 540–543.

Donald, B. (2011), *Algorithms in Structural Molecular Biology*, MIT Press.

Dunkle, J. and Cate, J. (2010), "Ribosome structure and dynamics during translocation and termination," *Annual Review of Biophysics*, 39, 227–244.

Eiben, C., Siegel, J., Bale, J., Cooper, S., Khatib, F., Shen, B., Players, F., Stoddard, B., Popovic, Z., and Baker, D. (2012), "Increased Diels-Alderase activity through backbone remodeling guided by Foldit players," *Nature Biotechnology*.

Engh, R. and Huber, R. (2001), *International Tables for Crystallography, Vol. F*, Dordrecht: Kluwer Academic Publishers.

Fleishman, S. and Baker, D. (2012), "Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution," *Cell*, 149, 262–273.

Fraser, J., Clarkson, M., Degnan, S., Erion, R., Kern, D., and Alber, T. (2009), "Hidden alternative structures of proline isomerase essential for catalysis," *Nature*, 462, 669–673.

Fraser, J., van den Bedem, H., Samelson, A., Lang, P., Holton, J., Echols, N., and Alber, T. (2011), "Accessing protein conformational ensembles using room-temperature X-ray crystallography," *Proceedings of the National Academy of Sciences*, 108, 16247–16252.

Frederick, K., Marlow, M., Valentine, K., and Wand, A. (2007), "Conformational entropy in molecular recognition by proteins," *Nature*, 448, 325–329.

Frey, K., Georgiev, I., Donald, B., and Anderson, A. (2010), "Predicting resistance mutations using protein design algorithms," *Proceedings of the National Academy of Sciences*, 107, 13707–13712.

Friedland, G., Linares, A., Smith, C., and Kortemme, T. (2008), "A simple model of backbone flexibility improves modeling of side-chain conformational variability," *Journal of Molecular Biology*, 380, 757–774.

Friedland, G., Lakomek, N., Griesinger, C., Meiler, J., and Kortemme, T. (2009), "A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family," *PLoS Computational Biology*, 5, e1000393.

Gainza, P., Roberts, K., Georgiev, I., Lilien, R., Keedy, D., Chen, C.-Y., Reza, F., Anderson, A., Richardson, D., Richardson, J., and Donald, B. (2012), "OSPREY: Protein Design with Ensembles, Flexibility, and Provable Algorithms," *Methods in Enzymology*, submitted.

Georgiev, I. and Donald, B. (2007), "Dead-end elimination with backbone flexibility," *Bioinformatics*, 23, i185–i194.

Georgiev, I., Keedy, D., Richardson, J., Richardson, D., and Donald, B. (2008a), "Algorithm for backrub motions in protein design," *Bioinformatics*, 24, i196–i204.

Georgiev, I., Lilien, R., and Donald, B. (2008b), "The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles," *Journal of Computational Chemistry*, 29, 1527–1542.

Goodsell, D. and Olson, A. (2000), "Structural symmetry and protein function," *Annual Review of Biophysics and Biomolecular Structure*, 29, 105–153.

Grigoryan, G., Zhou, F., Lustig, S., Ceder, G., Morgan, D., and Keating, A. (2006), "Ultra-fast evaluation of protein energies directly from sequence," *PLoS Computational Biology*, 2, e63.

Haber, E., Anfinsen, C., et al. (1962), "Side-chain interactions governing the pairing of half-cystine residues in ribonuclease." *Journal of Biological Chemistry*, 237, 1839.

Hallen, M., Keedy, D., and Donald, B. (2012), "Dead-End Elimination with Perturbations ("DEEPer"): A provable protein design algorithm with continuous sidechain and backbone flexibility," *Proteins: Structure, Function, and Bioinformatics*, accepted.

Harbury, P., Plecs, J., Tidor, B., Alber, T., and Kim, P. (1998), "High-resolution protein design with backbone freedom," *Science*, 282, 1462–1467.

Harper, E. and Rose, G. (1993), "Helix stop signals in proteins and peptides: the capping box," *Biochemistry*, 32, 7605–7609.

Headd, J., Immormino, R., Keedy, D., Emsley, P., Richardson, D., and Richardson, J. (2009), "Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place," *Journal of Structural and Functional Genomics*, 10, 83–93.

Hecht, M., Richardson, J., Richardson, D., and Ogden, R. (1990), "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence," *Science*, 249, 884–891.

Hofmann, B., Tölzer, S., Pelletier, I., Altenbuchner, J., Van Pee, K., and Hecht, H. (1998), "Structural investigation of the cofactor-free chloroperoxidases1," *Journal of Molecular Biology*, 279, 889–900.

Hopf, T., Colwell, L., Sheridan, R., Rost, B., Sander, C., and Marks, D. (2012), "Three-dimensional structures of membrane proteins from genomic sequencing," *Cell*.

Hu, X. and Kuhlman, B. (2006), "Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences," *Proteins: Structure, Function, and Bioinformatics*, 62, 739–748.

Huggins, D. and Tidor, B. (2011), "Systematic placement of structural water molecules for improved scoring of protein–ligand interactions," *Protein Engineering, Design, and Selection*, 24, 777–789.

I. Asimov, J. A. (1993), *Frontiers 2: More Recent Discoveries About Life, Earth, Space, and the Universe*, Truman Talley Books/Dutton.

Janin, J. (2002), "Welcome to CAPRI: a critical assessment of predicted interactions," *Proteins: Structure, Function, and Bioinformatics*, 47, 257–257.

Jiang, L., Kuhlman, B., Kortemme, T., and Baker, D. (2005), "A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein–protein interfaces," *Proteins: Structure, Function, and Bioinformatics*, 58, 893–904.

Jiang, L., Althoff, E., Clemente, F., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J., Betker, J., Tanaka, F., Barbas III, C., et al. (2008), "De novo computational design of retro-aldol enzymes," *Science*, 319, 1387–1391.

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

Kapp, G., Richardson, J., and Oas, T. (2004), "Kinetic role of helix caps in protein folding is context-dependent," *Biochemistry*, 43, 3814–3823.

Karplus, P. (1996), "Experimentally observed conformation-dependent geometry and hidden strain in proteins," *Protein Science*, 5, 1406–1420.

Keedy, D., Williams, C., Headd, J., Arendall III, W., Chen, V., Kapral, G., Gillespie, R., Block, J., Zemla, A., Richardson, D., et al. (2009), "The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models," *Proteins: Structure, Function, and Bioinformatics*, 77, 29–49.

Keedy, D., Georgiev, I., Triplett, E., Donald, B., Richardson, D., and Richardson, J. (2012), "The Role of Local Backrub Motions in Evolved and Designed Mutations," *PLoS Computational Biology*, 8, e1002629.

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., Phillips, D., et al. (1958), "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, 181, 662–666.

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011), "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nature Structural and Molecular Biology*, 18, 1175–1177.

Kim, D., Blum, B., Bradley, P., and Baker, D. (2009), "Sampling bottlenecks in de novo protein structure prediction," *Journal of Molecular Biology*, 393, 249–260.

Kleywegt, G. (1999), "Experimental assessment of differences between related protein crystal structures," *Acta Crystallographica Section D: Biological Crystallography*, 55, 1878–1884.

Kleywegt, G. and Jones, T. (1996), "Efficient rebuilding of protein structures," *Acta Crystallographica Section D: Biological Crystallography*, 52, 829–832.

Kleywegt, G., Harris, M., Zou, J., Taylor, T., Wahlby, A., and Jones, T. (2004), "The Uppsala electron-density server," *Acta Crystallographica Section D: Biological Crystallography*, 60, 2240–2249.

Kopp, J., Bordoli, L., Battey, J., Kiefer, F., and Schwede, T. (2007), "Assessment of CASP7 predictions for template-based modeling targets," *Proteins: Structure, Function, and Bioinformatics*, 69, 38–56.

Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B., and Baker, D. (2003), "Design of a novel globular protein fold with atomic-level accuracy," *Science*, 302, 1364–1368.

Lakomek, N., Carlomagno, T., Becker, S., Griesinger, C., and Meiler, J. (2006), "A thorough dynamic interpretation of residual dipolar couplings in ubiquitin," *Journal of Biomolecular NMR*, 34, 101–115.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001), "Initial sequencing and analysis of the human genome," *Nature*, 409, 860–921.

Lang, P., Ng, H., Fraser, J., Corn, J., Echols, N., Sales, M., Holton, J., and Alber, T. (2010), "Automated electron-density sampling reveals widespread conformational polymorphism in proteins," *Protein Science*, 19, 1420–1431.

Lange, O., Lakomek, N., Farès, C., Schröder, G., Walter, K., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C., and de Groot, B. (2008), "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution," *Science*, 320, 1471.

Laskowski, R., MacArthur, M., Moss, D., and Thornton, J. (1993), "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of Applied Crystallography*, 26, 283–291.

Lazaridis, T. and Karplus, M. (1999), "Effective energy function for proteins in solution," *Proteins: Structure, Function, and Bioinformatics*, 35, 133–152.

Leaver-Fay, A., Tyka, M., Lewis, S., Lange, O., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P., Smith, C., Sheffler, W., et al. (2011), "ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules," *Methods in Enzymology*, 487, 545–574.

Lilien, R., Stevens, B., Anderson, A., and Donald, B. (2005), "A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme," *Journal of Computational Biology*, 12, 740–761.

Lindorff-Larsen, K., Best, R., DePristo, M., Dobson, C., and Vendruscolo, M. (2005), "Simultaneous determination of protein structure and dynamics," *Nature*, 433, 128–132.

Lovell, S., Word, J., Richardson, J., and Richardson, D. (1999), "Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering," *Proceedings of the National Academy of Sciences*, 96, 400.

Lovell, S., Word, J., Richardson, J., and Richardson, D. (2000), "The penultimate rotamer library," *Proteins: Structure, Function, and Bioinformatics*, 40, 389–408.

Lovell, S., Davis, I., Arendall III, W., de Bakker, P., Word, J., Prisant, M., Richardson, J., and Richardson, D. (2003), "Structure validation by Cα geometry: ϕ, ψ and Cβ deviation," *Proteins: Structure, Function, and Bioinformatics*, 50, 437–450.

MacCallum, J., Hua, L., Schnieders, M., Pande, V., Jacobson, M., and Dill, K. (2009), "Assessment of the protein-structure refinement category in CASP8," *Proteins: Structure, Function, and Bioinformatics*, 77, 66–80.

MacCallum, J., Pérez, A., Schnieders, M., Hua, L., Jacobson, M., and Dill, K. (2011), "Assessment of protein structure refinement in CASP9," *Proteins: Structure, Function, and Bioinformatics*.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011), "Assessment of template based protein structure predictions in CASP9," *Proteins: Structure, Function, and Bioinformatics*.

McGuffin, L., Bryson, K., and Jones, D. (2000), "The PSIPRED protein structure prediction server," *Bioinformatics*, 16, 404–405.

Merkel, J. and Regan, L. (1998), "Aromatic rescue of glycine in [beta] sheets," *Folding and Design*, 3, 449–456.

Model, J. (1996), "1. Holm L, Sander C: Dali: a network tool for protein structure comparison." *Trends in Biochemical Sciences*, 20, 478–480.

Mowbray, S., Helgstrand, C., Sigrell, J., Cameron, A., and Jones, T. (1999), "Errors and reproducibility in electron-density map interpretation," *Acta Crystallographica Section D: Biological Crystallography*, 55, 1309–1319.

Murshudov, G., Grebenko, A., Brannigan, J., Antson, A., Barynin, V., Dodson, G., Dauter, Z., Wilson, K., and Melik-Adamyan, W. (2002), "The structures of Micrococcus lysodeikticus catalase, its ferryl intermediate (compound II) and NADPH complex," *Acta Crystallographica Section D: Biological Crystallography*, 58, 1972–1982.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., et al. (1995), "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, 247, 536–540.

Plaxco, K., Simons, K., and Baker, D. (1998), "Contact order, transition state placement and the refolding rates of single domain proteins1," *Journal of Molecular Biology*, 277, 985–994.

Presta, L. and Rose, G. (1988), "Helix signals in proteins," *Science*, 240, 1632–1641.

Quinn, T., Tweedy, N., Williams, R., Richardson, J., and Richardson, D. (1994), "Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein," *Proceedings of the National Academy of Sciences*, 91, 8747.

Ramachandran, G., Ramakrishnan, C., and Sasisekharan, V. (1963), "Stereochemistry of Polypeptide Chain Configurations," *Journal of Molecular Biology*, 7, 95–99.

Raman, S., Lange, O., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T., Eletsky, A., Szyperski, T., et al. (2010), "NMR structure determination for larger proteins using backbone-only data," *Science*, 327, 1014–1018.

Read, R. and Chavali, G. (2007), "Assessment of CASP7 predictions in the high accuracy template-based modeling category," *Proteins: Structure, Function, and Bioinformatics*, 69, 27–37.

Read, R., Adams, P., Arendall, W., Brunger, A., Emsley, P., Joosten, R., Kleywegt, G., Krissinel, E., Lütteke, T., Otwinowski, Z., et al. (2011), "A new generation of crystallographic validation tools for the Protein Data Bank," *Structure*, 19, 1395–1412.

Rees, D., Lewis, M., and Lipscomb, W. (1983), "Refined crystal structure of carboxypeptidase a at 1· 54 Å resolution*," *Journal of Molecular Biology*, 168, 367–387.

Richardson, J. and Richardson, D. (1988), "Amino acid preferences for specific locations at the ends of alpha helices," *Science*, 240, 1648–1652.

Richardson, J. and Richardson, D. (2002), "Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation," *Proceedings of the National Academy of Sciences*, 99, 2754.

Richardson, J., Richardson, D., Tweedy, N., Gernert, K., Quinn, T., Hecht, M., Erickson, B., Yan, Y., McClain, R., Donlan, M., et al. (1992), "Looking at proteins: representations, folding, packing, and design. Biophysical Society National Lecture, 1992." *Biophysical Journal*, 63, 1185.

Richter, B., Gsponer, J., Várnai, P., Salvatella, X., and Vendruscolo, M. (2007), "The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins," *Journal of Biomolecular NMR*, 37, 117–135.

Rohl, C., Strauss, C., Misura, K., and Baker, D. (2004), "Protein structure prediction using Rosetta," *Methods in Enzymology*, 383, 66–93.

Röthlisberger, D., Khersonsky, O. Wollacott, A., Jiang, L., DeChancie, J., Betker, J., Gallaher, J., Althoff, E., Zanghellini, A., Dym, O., Albeck, S., Houk, K., Tawfik,

D., and Baker, D. (2008), "Kemp elimination catalysts by computational enzyme design," *Nature*, 453, 190–195.

Salmon, L., Bouvignies, G., Markwick, P., and Blackledge, M. (2011), "NMR Provides a Quantitative Description of Protein Conformational Flexibility on Physiologically Important Timescales," *Biochemistry*.

Sela, M., White Jr, F., Anfinsen, C., et al. (1957), "Reductive cleavage of disulfide bridges in ribonuclease," *Science*, 125, 691–692.

Serrano, L. and Fersht, A. (1989), "Capping and α-helix stability," *Nature*, 342, 296–299.

Shapovalov, M. and Dunbrack Jr, R. (2007), "Statistical and conformational analysis of the electron density of protein side chains," *Proteins: Structure, Function, and Bioinformatics*, 66, 279–303.

Shapovalov, M. and Dunbrack Jr, R. (2011), "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions," *Structure*, 19, 844–858.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J., Liu, G., Eletsky, A., Wu, Y., Singarapu, K., Lemak, A., et al. (2008), "Consistent blind protein structure generation from NMR chemical shift data," *Proceedings of the National Academy of Sciences*, 105, 4685.

Siegel, J., Zanghellini, A., Lovick, H., Kiss, G., Lambert, A., St.Clair, J., Gallaher, J., Hilvert, D., Gelb, M., Stoddard, B., Houk, K., Michael, F., and Baker, D. (2010), "Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction," *Science*, 329, 309–313.

Simons, K., Kooperberg, C., Huang, E., and Baker, D. (1997), "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions1," *Journal of Molecular Biology*, 268, 209–225.

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

Smith, C. and Kortemme, T. (2008), "Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction," *Journal of Molecular Biology*, 380, 742–756.

Smith, C. and Kortemme, T. (2011), "Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design," *PLoS ONE*, 6, e20451.

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. (2011), "Structure-guided forcefield optimization," *Proteins: Structure, Function, and Bioinformatics*.

Stevens, B., Lilien, R., Georgiev, I., Donald, B., and Anderson, A. (2006), "Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity," *Biochemistry*, 45, 15495–15504.

Tainer, J., Getzoff, E., Beem, K., Richardson, J., and Richardson, D. (1982), "Determination and analysis of the 2 A-structure of copper, zinc superoxide dismutase." *Journal of Molecular Biology*, 160, 181.

Takano, T. (1977), "Structure of myoglobin refined at 2-0 A resolution. I. Crystallographic refinement of metmyoglobin from sperm whale." *Journal of Molecular Biology*, 110, 537.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M., and Dunbrack, R. (2010), "Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model," *PLoS Computational Biology*, 6, e1000763.

Tjandra, N., Feller, S., Pastor, R., and Bax, A. (1995), "Rotational diffusion anisotropy of human ubiquitin from 15N NMR relaxation," *Journal of the American Chemical Society*, 117, 12562–12566.

Tokuriki, N. and Tawfik, D. (2009), "Protein dynamism and evolvability," *Science*, 324, 203–207.

Tress, M., Ezkurdia, I., and Richardson, J. (2009), "Target domain definition and classification in CASP8," *Proteins: Structure, Function, and Bioinformatics*, 77, 10–17.

Tyka, M., Keedy, D., André, I., DiMaio, F., Song, Y., Richardson, D., Richardson, J., and Baker, D. (2010), "Alternate states of proteins revealed by detailed energy landscape mapping," *Journal of Molecular Biology*.

Tyka, M., Jung, K., and Baker, D. (2012), "Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers," *Journal of Computational Chemistry*.

van den Bedem, H., Dhanik, A., Latombe, J., and Deacon, A. (2009), "Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers," *Acta Crystallographica Section D: Biological Crystallography*, 65, 1107–1117.

Venkatachalam, C. (1968), "Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units," *Biopolymers*, 6, 1425–1436.

Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., et al. (2001), "The sequence of the human genome," *Science*, 291, 1304.

Vriend, G. (1990), "WHAT IF: a molecular modeling and drug design program." *Journal of Molecular Graphics*, 8, 52.

Warren, W., Hillier, L., Graves, J., Birney, E., Ponting, C., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A., et al. (2008), "Genome analysis of the platypus reveals unique signatures of evolution," *Nature*, 453, 175–183.

Word, J., Lovell, S., Richardson, J., and Richardson, D. (1999a), "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1," *Journal of Molecular Biology*, 285, 1735–1747.

Word, J., Lovell, S., LaBean, T., Taylor, H., Zalis, M., Presley, B., Richardson, J., and Richardson, D. (1999b), "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms1," *Journal of Molecular Biology*, 285, 1711–1733.

Zemla, A. (2003), "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, 31, 3370–3374.

Biography

Daniel Austin Keedy was born September 25, 1984 in Nashville, Tennessee. He has one mother, one father, one sister, and one brother.

Daniel has always had an interest in biology; over the years, that interest has narrowed to ever-smaller spatial scales. As a youngster, he loved animals, especially reptiles. This passion was evident from the frequent presence of captive snakes and lizards in the Keedy house. His parents were kind enough to treat him (and subject his siblings) to regular visits to zoos – especially reptile houses.

Daniel attended public elementary, middle, and high schools in Williamson County, Tennessee, a rural area south of Nashville. He graduated as valedictorian from Fred J. Page High School and was a National Merit Scholar.

He wished to remain somewhat close to home and attend a smaller, more intimate college, and fortunately received a generous scholarship from Rhodes College in Memphis, Tennessee. During this time, Daniel took advantage of several exciting opportunities, including a visit to Washington, D.C. to lobby for federal funding of undergraduate research and an unforgettable voyage aboard NASA's "Vomit Comet" to demonstrate an entirely electrostatic macroscopic orbit. He graduated Phi Beta Kappa with a B.A. (!) in Biochemistry and Molecular Biology and a minor in Spanish in May 2006.

In August 2006, Daniel enrolled in the Structural Biology and Biophysics graduate program at Duke University, hoping to combine his long-held love of biology

with a newfound appreciation for physics. Fortune smiled upon him as he stumbled upon two wonderful advisors in David and Jane Richardson and a supportive collaborator in Bruce Donald. He has co-authored 7 scientific papers in grad school: "Algorithm for Backrub Motions in Protein Design" (Bioinformatics, 2008), "Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place" (J. Struct. Funct. Genomics, 2009), "The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models" (Proteins, 2009), "MolProbity: all-atom structure validation for macromolecular crystallography" (Acta Cryst. D, 2010), "Alternate states of proteins revealed by detailed energy landscape mapping" (J. Mol. Biol., 2011), "The Role of Local Backrub Motions in Evolved and Designed Mutations" (PLoS Comp. Biol., 2012), and "Dead-End Elimination with Perturbations ("DEEPer"): A provable protein design algorithm with continuous sidechain and backbone flexibility" (Proteins, 2012).

Daniel plans to graduate from Duke with a Ph.D. in Biochemistry and a certificate in Structural Biology and Biophysics in September 2012. Afterwards he will join James "Jaime" Fraser's laboratory at the University of California, San Francisco to pursue further research in structural biology (and delicious Mission burritos).

He enjoys short walks on the beach, long walks in the mountains, and racquetball.