AN ABSTRACT OF THE DISSERTATION OF

Andrew E. Brereton for the degree of Doctor of Philosophy in Biochemistry and Biophysics presented on August 11, 2017.

Title: Exploring Protein Structure: Seeing the Forest and the Trees.

Abstract approved: _____
P. Andrew Karplus

Life on Earth intimately depends on the function of countless proteins. For the majority of studied proteins, function absolutely depends on conformation (*i.e.* 3-dimensional shape in solution). The exact nature of how a protein goes from an unfolded linear polypeptide chain to an organized folded molecule is still not known, and there is still much uncertainty about the details of folded protein structures. Answering both questions fully, especially by being able to accurately predict protein structures from sequence alone, is known as the protein folding problem. In this dissertation, the nature of protein structure research, especially as it pertains to model-building, entropy, and Boltzmann's principle, is discussed. Original work is presented in three chapters in the form of primary research reports. In chapter 2, high resolution protein crystal structures are used to describe the details of a high-energy transition conformation that occurs during protein folding. These native structures were found to have stabilized individual residues in conformations that represented "snapshots" along the transition pathway, and could be used to model the transition. In chapter 3, the extent and reliability of observed non-planarity of the peptide bond is assessed, using ultra-high resolution protein crystal structures. This work continues a discussion on peptide planarity that exists in the literature, and sets the record straight on the occurrence of this "non-ideal" geometry. In chapter 4, the "Ensemblator" is described and demonstrated. This software package is capable of comparing and analyzing large numbers of related protein models simultaneously, and represents an invaluable tool to protein structure researchers. Lastly, in chapter 5, impacts and

highlights of the reported work are discussed, along with directions for future work. The dissertation concludes with a reflection on the nature of problems being researched in protein structure, and a discussion on how the included original work relates to Boltzmann's principle and model building in general.

Exploring Protein Structure: Seeing the Forest and the Trees

by
Andrew E. Brereton

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

August 11, 2017
Commencement June 2018

Doctor of Philosophy dissertation of Andrew E. Brereton presented on August 11, 2017

APPROVED:

Major Professor, representing Biochemistry and Biophysics

Chair of the Department of Biochemistry and Biophysics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Andrew E. Brereton, Author

ACKNOWLEDGEMENTS

There are many people I need to acknowledge, and no way that the full depth of my appreciation to everyone can be sufficiently expressed.

First, I would like to thank my mentor, Dr. Andy Karplus. Andy, most of all, thank you for the freedom and encouragement you gave me to pursue my interests in the lab. Even when we would have a meeting in which I would obsessively discuss only some off-topic thing I was working on (*the reassigner…*), you would say "I'm encouraging you, this is me encouraging you!", and it made a huge difference. You would always give me as much help and guidance as I asked for, but you never micromanaged me, and I recognize now that this more than anything else has allowed me to grow as a scientist and become independent and capable. You always encouraged my good habits, and never hesitated to teach me about my bad habits (at least some of which I hope I have put behind me). You taught me the value of: clear, concise language; the need for precision and accuracy in speech and in practice (*I'll keep watching for flippant exaggerations*); an inordinate amount of good science; and that all the interesting and crazy ideas in the world aren't worth much if you can't express and share them. Thank you for all the opportunities you gave me (conferences, award nominations, letters of rec, *etc.*). There are a million more things that I am grateful for, but I'll let it be enough to say that I will always feel this gratitude for the gifts you gave me and I hope that I can live up to it. Also, let the record officially show that I liked all the puns, even the bad ones, and even when I pretended not to.

Next, I would like to thank the members of my lab, without whom I am sure I would not be even remotely the same person as I am now. I need to thank Dr. Dale Tronrud, for being a good teacher. Dale, you taught me that X-ray crystallography isn't magic (almost though). You taught me that choosing the right word doesn't mean just considering the meaning, but also the baggage (cultural or otherwise) that comes with it. You also taught me that convictions should only be as strong as your ability to defend them rationally. Thank you for the hundreds of long conversations and arguments, and for never telling me I was asking a stupid question. I also need to

thank my two graduate student mentors, Dr. Arden Perkins and Dr. Camden Driggers. Arden, you taught me to always be ready to face failure (esp. crystallization), but to never be unwilling to face it. You showed me that any goal can be achieved through determination and effort. I wouldn't have been making and submitting cover images with my manuscripts if I hadn't seen you doing it first. Camden, you taught me (explicitly) that sometimes you need to do things your own way, regardless of what others might say. Your positive and relaxed attitude made every day pleasant, and you *always* helped me when I needed it, without hesitating. Lastly, I need to thank Kelsey Kean, my friend and peer. Kelsey, firstly, thanks for putting up with me. You probably saw me more than almost anyone, and it was usually while you were trying to work. You have helped me with a countless number of things, in lab and outside of it, and never asked for anything in return. You also always encouraged me to do better, and if I did, I think it is partly because of that.

I would like to thank the members of my committee. Each of you gave me advice, guidance, conversation, and encouragement. Your questions during my prelim directly led to my own improved understanding of the work I hoped to do, and has resulted in me learning things I never thought I would understand. I'd like to especially thank Dr. Victor Hsu, for all the in-depth conversations, about science, careers, or even just shared interests, and for all the advice. I'd like to thank the department of biochemistry, for the opportunity, the funding, the support, and for the welcoming and open atmosphere of collaboration. I'd especially like to thank Dr. Michael Freitag, who I really feel always had my best interests at heart, and looked out for me. I'd also like to thank the NIH, Oregon State University, and the Department of Biochemistry and Biophysics for funding and awards I received during my studies.

I thank my family, for always telling me to follow my interests, even if it took me very far away from them. Thanks Mom, for always encouraging my love of science, and for getting me hooked on science fiction as a kid.

Lastly, I need to thank all the friends I have made here, without whom I would have failed more than just my studies. Especially Nathan Jespersen and Andrew Popchock, for always being there when I needed them, and for just getting it.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

Dedicated to:

To Aradine Dulake, for always laughing.

Exploring Protein Structure: Seeing the Forest and the Trees

Chapter 1

Introduction

The Importance of Protein Structure

Any understanding of biology must include some understanding of proteins. The reason for this is clear: the mind-numbing range of functions for proteins in life, and their abundance in living organisms. For example, proteins account for ~20% of the mass of a typical human cell, second only to water[1]. These abundant proteins can act as enzymes, catalyzing the chemical reactions that enable life, as structural scaffolding, providing a matrix for bone growth or building hair or shells, and even as sensors, reporting the presence or absence of a huge variety of signals. In these and most other studied examples, the function of the protein is critically dependent on its conformation (*i.e.* its three-dimensional structure in solution), whether that be one or a few well defined folded conformations, or a mixture of less structurally well-defined conformations. As such, to understand how a protein works (or why it does not work), the nature of protein structure must be understood and described.

A Limited Primer on Protein Structure

In this section of the introduction, I will briefly describe some of the critical "textbook" concepts of protein structure and biochemistry that are required to understand the more complex concepts in the rest of this dissertation.

Describing Protein Structure

A protein is a biological macromolecule, composed of one or more long polypeptide chains of covalently linked amino acid residues. All of the 20 canonical amino acids that occur in natural proteins have unique chemical and structural properties. The exact sequence of the specific amino acid residues that compose a given protein is said to be its primary structure. It is the primary structure of the protein that will determine the final structure, through the interaction of these residues with each other and with the environment, in a process called folding.

The strongest energetic contributor favoring the folding of proteins is thought to be "the hydrophobic effect"[2–4], named for the tendency of the less polar sidechains of some amino acids to pack together and bury surface area away from the water in

which the protein is solvated. To anthropomorphize a bit, the more holistic "goal" of the folding protein can be described as being to pack these hydrophobic sidechains together into a hydrophobic core, surrounded on the outside by residues more "comfortable" interacting with water. While it's hard to overstate the importance of the hydrophobic effect to protein folding, it is not the only major contributor to the folding process. Another major player in this process is the energy involved in forming hydrogen bonds[3]. Since the unfolded protein has almost every residue exposed to solvent, it can be easily imagined that almost every hydrogen bond capable of being formed can be satisfied by the abundant water molecules solvating the protein. Thus, once the protein folds and many of these atoms are no longer exposed to solvent, they must hydrogen bond with each other. Of the atoms that can form hydrogen bonds, both side chain atoms and backbone atoms play a role in protein folding, though the hydrogen bonding patterns formed by the backbone atoms tend to be much more regular and consistent.

These consistent hydrogen bonding patterns are a large part of what defines the secondary structure of a protein. The secondary structure is defined as the 3-dimensional conformations of small segments of a protein (*i.e.* its local conformation). These simple elements of structure tend to repeat along the backbone, have consistent hydrogen bonding patterns, and can form as an intermediate step during folding. The two most common defined elements of secondary structure, by far, are the α-helix and the β-strand. Once the protein has folded, the elements of its secondary structure interact with each other, (*i.e.* they are packed together), to form the tertiary structure of the protein: the final folded conformation. Many proteins also contain a final level of structure, called the quaternary structure, which is defined as the number and positioning of multiple folded polypeptide chains that come together to create a larger, more complex macromolecule.

Finally, two more terms to consider that are very useful when discussing protein structure are protein "domains", and protein "folds". A protein domain is typically a part of its structure that can fold independently of the rest of the protein, and could be stable on its own. While some proteins consist of only a single domain, many consist of multiple domains that fold together. This definition of a domain may

sound fairly unambiguous, however, it is often not easy to define domains in practice. The "gold standard" for defining a domain is still said to be the trained eye of an experienced investigator of protein structure[5]. A protein's fold or "topology" is the pattern made by considering the order of its secondary structure elements along the chain and how they pack in space. For example, a protein composed of four *anti-parallel* α-helices in sequence would be said to have a different fold than a protein composed of four *parallel* α-helices, even though the overall shape of the two proteins could be relatively similar.

Experimental Methods to Obtain Structures

Only a few experimental methods exist that can yield enough information to accurately determine the 3-dimensional structure of a folded protein. Of these, the most used and well known are: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). Each of these methods has its own associated advantages and disadvantages, which is why the most complete structural information often comes from using a combination of the methods. X-ray crystallography depends on growing crystals from proteins, measuring the diffraction patterns produced when those crystals are exposed to an X-ray beam, and performing complex data analysis and refinement to produce a model that is consistent with the underlying diffraction data and a calculated electron density map. Depending on myriad factors, (such as the data collection temperature, the size and level of orderliness of the crystal, the mobility of the protein, *etc.*), the diffraction will be of higher or lower quality. Higher quality diffraction will often lead to better resolution, meaning more details will be visible in electron density map, and the final model will be of higher quality. One great strength of X-ray crystallography is its ability to precisely determine the positions of atoms (at high resolutions). One weakness is that it is often difficult to obtain crystals of sufficient quality, especially in the case of very dynamic proteins or membrane proteins.

NMR spectroscopy is a method that excels at informing about features of protein structure like dynamics and flexibility. NMR spectroscopy depends on, ideally, many nuclei in the protein structure having a distinct chemical environment.

NMR spectroscopy is used to obtain multiple spectra that each contain different types of information about the protein structure. For example, NOESY experiments can provide information about which protons are close to each other in space, regardless of whether they are covalently bonded. By combining the data generated during these experiments, and doing what is known as resonance assignment to associate each observed peak with specific nuclei in the protein, the data can be used to restrain simulations of protein models. These simulations are run to generate ensembles of models (commonly using a method called molecular dynamics). Alone, these simulations would almost certainly not produce accurate models. However, by restraining features of the possible models using the experimental information gained in the NMR experiments, it is possible to create an ensemble of models that accurately represents the structure of a protein (defined as models that satisfy both the experimental restraints as well as assumed geometric restraints of protein structure). A great strength of this method is its ability to describe flexible, dynamic proteins, while a weakness is its dependence on simulation to generate models; the parts of the models that depend on simulation can only be as accurate as the algorithms used to produce them.

Cryo-EM is a much more direct imaging method than both NMR and X-ray crystallography. In cryo-EM, proteins or protein complexes in solution are frozen in a monolayer on a surface, and imaged directly using electron microscopy. Each of these images has noise that interferes with naive visualization of the structure, but the images can be combined (with similar conformations/views being placed together) and analyzed to greatly improve the quality of the final model. With cryo-EM, it is difficult to get extremely precise data for the positions of each atom, however, it is possible to get structural information about very large, complex proteins.

The Protein Folding Problem

As discussed above, the tertiary structure of any given protein is determined by the interplay of its specific sequence of amino acids and its environment during folding which typically (but not always) occurs during or directly after translation. It is well known that various environmental factors such as temperature, pH,

concentration, and the presence or absence of binding partners can influence (*i.e.* favor or disfavor) the folding of proteins[3]. Typically, all these factors will continue to play a role in shaping the conformation of a protein even after it has finished being translated and folded. This is because proteins are dynamic objects, constantly moving and sampling different available conformations. Thus, the entire 'native state' of a protein is not only one single conformation (a common misconception), but is better described as the equilibrium of all the productive conformations (*i.e.* those that enable it to complete its function) that a protein adopts.

The equilibrium of conformations available to a protein, sometimes also including the unfolded or aggregated conformations, is often referred to as an energy landscape[6]. This typically funnel-shaped landscape has multiple minima that correspond to various stable conformations that the protein could adopt, both during and after folding, and a much larger number of possible unfolded conformations at the top of the funnel. Occasionally, the most stable possible conformation will be some sort of partially unfolded aggregate, which can have disastrous results for the health of the organism[7]. In the context of the energy landscape, this would be the lowest minimum, and would (hopefully) be restricted by high energy barriers surrounding it that reduce the chances of spontaneous aggregation. In contrast to this, the desired native state of the protein will ideally be accessible via lower energy barriers, through a series of intermediate semi-stable conformations; this is referred to as the folding pathway[8].

It is the loosely defined goal of "solving" the protein folding problem to be able to predict the final folded conformation of a protein, especially from its sequence alone (called *de novo* prediction), and to fully understand how folding pathways enable the great speed at which proteins fold[6]. Aside from the amazing practical benefits this ability would provide to biochemists, pharmacologists, and bioengineers, an important reason we care about this milestone achievement, currently only imperfectly obtained[6], is that we can use *de novo* prediction of structure as a metric to assess the quality of our understanding of protein structure[9].

An Interesting Way to Think About Protein Structure

Boltzmann's Principle, Entropy, and Protein Structures

I find that it can be interesting to think statistically about the conformations that punctuate the energy landscape. One can imagine that each of these stable or semi-stable conformations can be represented with a certain probability of existing (in solvent, typically water), and transitions between conformations also occur at certain frequencies that depend on a large variety of factors (*e.g.* intermolecular forces, intramolecular forces, temperature, *etc.*). The most exciting result of thinking about protein conformational energy landscapes in this way is that it enables us to take advantage of Boltzmann's principle, and to attempt the arduous task of predicting native state conformations *de novo*.

So how does Boltzmann's principle relate to our attempts to predict protein structure? In rough terms, the principle states that the entropy of a system at equilibrium is a measure of the number of *microstates* of the system that are consistent with the observed *macrostate* of that system. This is best understood using the classic example of a noble gas in a container: properties like the position, velocity, and momentum of each molecule of gas together define the unobservable microstate of the system; larger-scale properties like the temperature, pressure, and volume define the macrostate of the system. Thus, a perfect understanding of the microstate would contain everything needed to perfectly describe the macrostate. Interestingly, this is not true in reverse: knowledge of the macrostate does not provide enough information to describe the true microstate, instead there may be incredibly large numbers of possible microstates that will satisfy the same macrostate. Boltzmann's principle is an equation that links the energy terms of the individual molecules in the system, the free energy of the system, and the probabilities of the possible microstates.

A remarkable aspect of Boltzmann's principle is that it can be applied in a variety of ways to achieve the same aim (an accurate description of the system being studied). In an example that is more related to protein structure, an *inductive* approach to understanding structure could begin with quantum-mechanical calculations and

observations from very simple systems, which are then extrapolated up to the macro-scale in an attempt to create an accurate forcefield (defined as the derivative of the energy functions) that can describe more complex systems. The *deductive* approach attempts to solve the same problem from the opposite direction. In this methodology, details of structure are extracted from experimentally solved structures of entire proteins and used to infer a forcefield that could, ideally, describe unobserved structures as well as the observed structures (*i.e.* every observed folded structure is treated as a single microstate that satisfies the macrostate of 'stable folded protein conformation'). So, it is the striking property of Boltzmann's principle that it not only links the probability of multitudes of microstates to the energy of the system, but that it also links the inductive and deductive approaches to understanding such a system, through a single common element: entropy[*]. For me, this observation is the conceptual key of this dissertation.

Descriptions of Protein Structure Require Scale and Context

Before I continue, I need to define new uses for two terms that I will use throughout this introduction. These two terms are *holistic* and *atomistic*, and they are conceptually related to *deductive* and *inductive*. As we saw, a deductive understanding of protein structure might begin by looking at solved protein structures, and extracting general principles to describe the forces that led to those structures. Likewise, inductive reasoning is its counterpoint, beginning with simpler and more ideal systems, and scaling up to a better understanding of more complex systems and eventually the big picture. The reason I want to avoid using only the terms inductive and deductive is that I think there is a key component missing in how they are generally used. Inductive and deductive reasoning are often seen as two complementary tools aimed at solving the same problem, with each having its own strengths and pitfalls. Here, I want to use the words holistic and atomistic to introduce a link between the two conceptual frameworks for understanding protein structure,

---

[*] For a more detailed description of the relationship between Boltzmann's principle and protein structure, see: Sippl, 1993[9]

and to encourage appropriate thinking about scale, more in line with the flexibility inherent to Boltzmann's description of entropy.

Merriam-Webster (2017) defines holistic as: "relating to or concerned with wholes or with complete systems rather than with the analysis of, treatment of, or dissection into parts", and atomistic as: "composed of many simple elements; also: characterized by or resulting from division into unconnected or antagonistic fragments". Based on this, we can imagine that a holistic understanding of protein structure might concern itself with broad, macro-scale details (*e.g.* native state structures, folding pathway analysis, protein evolution, etc.). However, predictions on this scale are often extremely difficult to test; it is preferable to 'dissect into parts' and create independent, testable, atomistic predictions. These individual atomistic details of protein structure all must be accounted for and satisfied by any holistic description, if that holistic description is to have any hope for accuracy. This is conceptually similar to treating the holistic understanding as a macrostate, and the myriad combinations of various compatible atomistic observations as different microstates. A holistic framework needs to be informed by accurate atomistic information, but it is only by discarding the exact details of the atomistic information (*i.e.* summarizing) that the holistic model can be created (otherwise, it would only be a collection of atomistic pieces).

Just as the number of possible microstates for a given macrostate is often not enumerable, the number of ways to atomistically describe protein structure is limitless, and will vary depending on context; one simply needs to imagine reducing the scale of observations smaller and smaller, *ad infinitum* (*e.g.* quaternary structure to tertiary, to secondary, to primary, to individual residues, to individual atoms, to quarks and leptons, and so on). As these mostly independent observations grow in number, it becomes clear that a holistic framework is needed to provide context and meaning. On their own, independent atomistic observations tend to have limited utility. Describing something at a holistic scale by only using atomistic elements would be like trying to describe the *Mona Lisa* to someone else by only listing the exact locations, chemical composition, and density of the pigments on the canvas.

Lastly, it is critical to note that the context and scale are never absolute, but both must be defined when defining something as holistic or atomistic. By this I simply mean that any holistic description can be made atomistic by considering it to be an atomistic piece of a "larger" holistic theory, and any atomistic description can be seen as holistic by considering the discarded information that created it (*e.g.* describing a residue as a *serine* rather than explicitly defining all of its atoms, their bonds, and their geometry). If one is interested in the forest, then seeing the forest is the holistic understanding, and seeing the trees is the atomistic understanding. If one is interested in the trees however, then maybe the description of the trees is holistic, and to describe the trunk and branches and individual leaves is atomistic. Thus, holistic and atomistic understanding depend totally on one another, change into one another, and unlike inductive and deductive reasoning, they don't merely work toward describing the same thing, but, they are simply the same system described with differing amounts of entropy. With these terms defined, the rest of this introduction will be concerned with describing our understanding of protein structure, and establishing the scale on which we can consider observations and methods to be mostly holistic or atomistic.

Atomistic Understanding of Protein Structure

To begin to set the stage for understanding how my work fits into the larger picture of research into protein structure, I want to first describe some of the atomistic research that provided the background and the context for all the research described within this dissertation. This is the research that is concerned with getting the details correct, and with precision to the highest degree possible. The work I will discuss in this section together constitute many pieces of the puzzle that are placed into our larger understanding of protein structure.

Early Work

In the simplest sense, the early atomistic approaches to understanding protein structure began by asking the question "How do amino acids interact to produce a final folded protein?". This approach is intuitively pleasing, and seems to make sense;

since amino acids are the basic constituent of all proteins, in some way their interactions must define protein structure. For example, to understand the contributions of the individual types of amino acids to the overall energy of protein folding, researchers attempted to quantify the "hydrophobicity" of each type of amino acid occurring in proteins. Many hydrophobicity scales for the 20 canonical amino acids were proposed and estimated[2,10]. While the details of these scales differ (especially depending on the specific solvent used to measure them), the general trends are mostly the same. Amino acids with larger hydrophobic sidechains and especially those lacking the possibility of hydrogen bonding, such as phenylalanine, tend to contribute most to the hydrophobic effect[2].

No structural biology dissertation at Oregon State University would be complete without a cameo from the ubiquitously influential Linus Pauling. This is where he makes his appearance in our story, with his critical description of the α-helix. Pauling understood that for a protein to fold most of the hydrogen bonds being made with water must remain satisfied in the folded state (for the overall energy changes to balance out). When the protein folds, the backbone atoms that are capable of hydrogen bonding (carbonyl oxygens and amide nitrogens) interact with each other in consistent and repetitive ways to form the secondary structure of the protein. The first such units of secondary structure described correctly were the α-helix and the β-strand, as modelled by Linus Pauling[11,12]. Eventually other elements of secondary structure would be described (P_{II}-helix, hydrogen bonded turns, etc.), but Pauling's original description of the most abundant elements of secondary structure (the α-helix and β-strand) set the stage for understanding these simple repeating units of protein structure, which together contribute to the final structure of the protein.

The majority of this work on modelling the protein was enabled by extensive research into the structure of amino acids and small peptides[13]. Why these particular studies fit into the atomistic umbrella should be intuitively clear: each of them independently provides information only for the small systems they described, but together, they could begin to inform a larger picture, and in turn lead to better atomistic studies (*e.g.* Pauling and Corey's work on the α-helix[11]). In fact, even Pauling's work in describing the α-helix immediately led to experimental work with

small peptides to test and verify his prediction about this important element of protein structure[12]. Over the years, this work of characterizing the minute details of peptide (and thus protein) structure has continued. A landmark in this direction of research was the work of Engh and Huber[14,15], producing and compiling a series of tables containing details about the bond angles and lengths in protein structures. The values for these parameters would go on to inform model refinement in X-ray crystallography and NMR-based model building, and also protein structure prediction. As will be discussed later, though valuable, it's possible that the protein structure community became somewhat complacent regarding the tables produced by Engh and Huber: spending many years refining and updating the individual values when perhaps a new paradigm was called for. As interesting and useful as it has been to describe the details of how individual amino acids are structured within a protein, thinking about the ways that individual residues *cannot* be structured has perhaps led to one of the most important insights in all of protein structure research.

In 1963, Ramachandran and his graduate student Sasisekharan described the famous graph we now refer to as the Ramachandran Plot[16,17], which is simply a plot comparing the two main variable backbone dihedral angles, φ and ψ. By considering the individual amino acids in a polypeptide chain, and asking the question "What values of φ and ψ will produce steric clashes?", they were able to describe a simple range of sterically allowed conformations for residues within a protein. This way of thinking about conformation profoundly impacted our understanding of protein structure. The sterically allowed conformations on the plot included the two most common elements of secondary structure (α-helix and β-strand), and refinement, model building, and other aspects of protein structural research could be greatly improved by eliminating the need to consider conformations in the sterically disallowed regions of the plot.

Piece by piece, a larger picture of protein structure was beginning to be built up from these many atomistic results. The shape of the whole was beginning to come clear.

Contemporary Work

The contemporary atomistic research into protein structure is a clear continuation of the research described above, and could be mostly summarized as "improvement". In each case as the descriptions have gotten more detailed, the precision of the results has improved, and the accuracy of the predictions is presumably higher. In this section, I will outline some of that work.

As discussed earlier, early forcefield parameterization for proteins was done mostly using details from small peptides and amino acids. As time has passed, however, increasingly inexpensive computation has provided opportunity to make some of the limitations of these forcefields more obvious, demonstrating the need to improve the accuracy and enable longer simulations of proteins[6]. A lot of the work done to improve the quality of forcefields involved updating or correcting the values previously derived for the atomic details of protein structure, or in some cases, even providing a totally new paradigm. Consider bond angles for example. After being updated by Engh and Huber in 1999[15], the bond angles for amino acids within a protein were considered to be more accurate, though only small incremental changes had been made. In contrast, it was understood by some that the paradigm of treating each backbone angle as if it had an ideal value was perhaps deeply flawed[18,19]. Instead, it seemed that the "ideal" value of any angle would change depending on φ and ψ, indicating a dependence on local conformation for these atomic-level details of protein structure. This meant that work was needed to update the paradigm that was used to inform refinement of experimental structures[20], as well as used to predict protein structure[21,22].

In 2009, Berkholz *et al* described a conformation-dependent geometry in which the target values of the backbone bond angles along the protein chain change depending on the values of φ and ψ at that position[23]. This new paradigm brings context into the equation when considering the details of local conformation. For example, the N-Cα-C angle is expected to have values near 108° in some conformations, but closer to 114° in others. In contrast to this, in the old paradigm, the single value estimate for this angle would simply be 111° (per Engh and Huber[15])

in any conformation. The discrepancy is clear, and it's apparent that the single value is an artificial result of averaging over an uneven landscape. Switching over to the new paradigm of conformation dependent geometry has led to considerable improvement of models during crystallographic refinement with "no disadvantages... apparent" compared to the old single-value paradigm[24],.

The considerable interest in revisiting and reanalyzing the established details of protein geometry has not only been restricted to the bond angles within the backbone. A partial explanation for this renaissance in protein geometry lies in the vastly increasing number of structures, especially at high resolution, deposited in the Protein Data Bank[25] (PDB). Since the first protein structure was solved at atomic resolution by John Kendrew in 1958[26], the number of solved protein structures in the PDB has now swelled past 130,000[27], and with methods for solving structures being continuously developed or improved, that growth shows no signs of slowing[27]. Recently the Protein Geometry Database (PGD) was made available[28], an auxiliary tool for searching among structures deposited in the PDB for specific features of protein geometry. In this database, it is possible to quickly identify regions of proteins having very particular features (*e.g.* all three-residue segments that end in a proline and have a *cis*-peptide bond), and to filter the results using various quality control methods (*e.g.* a cutoff for sequence similarity, for resolution, for crystallographic R-factor, etc.). Throughout my own research, I made extensive use of the PGD; the value of the PGD is that it enables further research into protein geometry, and it greatly simplifies the work of finding the protein structures that a given researcher might be interested in researching.

Another important feature of protein geometry that continues to be studied is the planarity (or lack thereof) of the peptide bond. In the past, this feature of local conformation was anticipated even before the first protein structure was solved; it was the assumption that the peptide bond forming the protein backbone was roughly planar that allowed Linus Pauling to predict elements of secondary structure, by only considering the effect of the φ and ψ angles on the conformation of the protein backbone (as the story goes, he folded a paper helix while sick in bed after becoming bored of reading science fiction and detective novels). It is also important to note that

this planarity was not thought to be overly strict: even Pauling and Corey estimated that deviations of ~10° from planarity would only carry a cost of ~1 kcal/mol[12]. Despite this, in the past few years there has been some debate on this subject, with some claiming that estimates of non-planarity are errors in modelling that need to be "corrected" back to the ideal[29], and others quantifying the extent of observed non-planarity in deposited protein structures[30] and asserting its legitimacy. In this latter study, 0.5% of observed residues from very well defined protein crystal structures deviated more than 20° from planarity, and the electron density of these extreme outliers suggested they were modeled correctly. Overall, the results painted a picture of a flexible protein backbone that never quite conforms to "ideal" geometry.

Ramachandran's early predictions about the regions of φ,ψ space that are sterically allowed and disallowed have proven over time to be fairly accurate, and have since been reinforced by further investigations based on solved protein structures[31–34] and detailed modern simulations[35]. While the basic understanding of φ,ψ-space has remained more or less the same, the depth of that understanding has increased as researchers have investigated the role backbone conformation plays in protein structure. So-called $(\varphi,\psi)_2$-motifs[36], for example, provide a more discretized and complex view of backbone conformation than the simply defined allowed and disallowed regions of φ,ψ space. These motifs were characterized by considering φ and ψ for two adjacent residues at the same time, and clustering the resulting 4-dimensional space into significant subgroups, each of which is considered a conformational motif. This is conceptually similar to the use of fragments to build predicted structures in the Rosetta software package[22], as well as the use of naive structural alphabets to describe secondary structure, rather than terms like α-helix or β-strand[37,38]. The main difference between these various approaches to describing backbone conformation tends to be scale (*e.g.* considering 2 vs 12 residues at a time), while the main similarity is that in each case an attempt is being made to transcend the simplistic single-residue model originally characterized by Ramachandran.

Holistic Understanding of Protein Structure

Compared to atomistic research into protein structure, holistic research tends to be much more varied. This work is concerned with the big picture, and creating an accurate context that both accounts for existing atomistic work and guides future atomistic studies. As increasingly better atomistic details become available, the accuracy and completeness of holistic theories improve significantly, in turn leading to more holistic research and more guided atomistic research. In this section I will describe some examples of this type of research to provide the context for my own work.

Early Work

Evolution is the core process that underlies all of biology. Every living thing has been shaped by it. Unsurprisingly, since proteins are the tools by which most living organisms build themselves and interact with the environment, they are one of the key substrates in which the process of evolution can take place. As such, there is much to be learned both about biology and about proteins from the study of the evolution of proteins. Perhaps one of the earliest examples of this was the comparison of the sequence of insulin from a few different species, by Sanger in 1956[39]; though an important in stepping stone, it's not clear that this directly or indirectly improved our understanding of protein structure, due to its extremely limited scope. Over time, however, this research subject matured into a source of rich information about protein structure, as much more complex analyses became possible (*e.g.* an early comparison of protein sequences between cyanobacteria and eukaryotes[40]). The study of protein evolution led to insight into substitution rates for various amino acids, providing evidence that amino acid chemistry, size, hydrophobicity, and other such factors, generally play more of a role in protein evolution, over the long term, than simple codon swapping frequencies[41,42]. It's intuitively clear how this could guide further atomistic research: all of these features are clearly important to protein structure or function and warrant further investigation. This insight also led to the development of important tools for studying proteins, called substitution matrices[43], which can be

used for assessing the relative similarity or difference of proteins based on their sequence.

The development of improved substitution matrices also allowed for more accurate alignment of protein sequences, and the beginnings of the separation of known proteins into large protein families[44]. Independently, it also become possible to align proteins using their full structure, rather than just their sequences, and to use that information to describe protein families[45]. This work was not just done for the sake of improving taxonomy, but by including proteins with known structures and/or functions into families, the function and structure of otherwise uncharacterized proteins could often be inferred. Furthermore, when proteins in the same family differ in function, it becomes very interesting to investigate what structural changes led to that difference in function. The work to answer this question is, again, more atomistic, requiring comparison of details of the structures.

While useful information often came from comparing structures in detail, it was also necessary to compare structures at a broader level. This was a requirement of both time (of the researchers comparing large numbers of structures manually), and computational difficulty (detailed atomic comparisons are costly compared to simpler comparisons). To suit these purposes, protein folds were used to describe the overall structure of the protein in a more general sense. Databases were developed to track and organize various protein folds[46,47], to facilitate research into how folds evolve and the relationship between function and fold[48–50], as well as efforts to find new folds or predict folds accurately[51,52], and even attempts to determine the total number of unique folds that exist in nature[53]. While it has been productive to describe protein structure in this way, it is important to remember describing proteins at the level of folds is a compromise, and as such comes with drawbacks. One such drawback is that depending on the exact methodology used to define a protein's fold, a single model of a structure can be interpreted as having different folds. Furthermore, while not necessarily common, some proteins can have multiple stable conformations that are so different as to represent completely different folds (so-called "metamorphic proteins"[54–56]). The fold is also necessarily limited to describing only the smallest stably folding unit of the protein, the domain.

Consistent with my overall theme in this dissertation, a whole protein structure can be broken up into parts, which themselves can be broken up into parts, and so on, until a point where perhaps little useful information can be gained (in fact some proposed using similar hierarchies in reverse to describe the sequence of events that lead to folded proteins, from residue to secondary structure, to super secondary structure, to sub-domain, and so on[5]). One of earliest recognized and largest of these essential parts to any protein structure is the domain. Over the years, there has been considerable effort to design computational methods to detect and define domains, with varying success. One particularly interesting example conceived by George Rose in 1979 involved projecting the 3-dimensional coordinates of the protein backbone onto a 2-dimensional plane, and finding the line that divides the plane into two halves with as few cuts through backbone as possible[57]. This solution is very simple in its essence, but impressively useful for finding domains. A neat feature of this approach is that it can be repeated iteratively, to find ever smaller "domains" within the protein; so, it is possible to define multiple hierarchical domains for the protein, or even sub-domains, or sub-sub-domains, *etc*. Since the maturation of the many databases of protein families, domains, and their folds, domains are typically (but not always) identified by matching patterns to known and already identified domains[58].

Another critical early work that I would be remiss if I did not mention is the creation of the DSSP algorithm for automatically detecting secondary structure (named for a Pascal program created by the original authors, called *Define Secondary Structure of Proteins*)[59]. Like Rose's approach to identifying domains, DSSP is beautiful in its hierarchical organization and its extreme simplicity (a common requirement in an era with less ubiquitous computing power). Since a measure of all the specific elements of protein geometry (φ, ψ, Cα to Cα distance, *etc*.) would be both time consuming, and not necessarily simple to use to define secondary structure (*i.e*. the values of these elements exist along a spectrum, and cutoffs must be defined separately for each), the approach used by DSSP instead is to detect the presence or absence of main chain hydrogen bonds, which "can be characterized by a single decision parameter, a cutoff in the bond energy"[59]. Once the locations of the backbone hydrogen bonds have been established, a growing hierarchy of hydrogen

bonding patterns are identified, and the final result is the definition of the main elements of main chain hydrogen-bonded secondary structure for that protein. Overall the method works quite well, and I would be hard pressed to identify any single program that feels more omnipresent in protein structure research, even today.

Contemporary Work

The contemporary holistic research on protein structure is, in my opinion, some of the most interesting research taking place today. The breadth of this topic is such that it spans all the way from trying to understand the role(s) of misfolded and aggregating proteins in disease (*e.g.* the link between amyloid fibrils and Alzheimer's disease[7,60]), to trying to describe networks of intricately related motions throughout a folded protein[61–64]. Each of the topics I hope to cover in this section are examples of what can be discovered when experiments inform theories, and when experimental data is abundant and of high-quality. Even as I earlier chose the word "improvement" to describe the contemporary atomistic work relative to the early work, to summarize how contemporary holistic compares with early holistic work, I would choose the word "expansion". While the accuracy and richness of our holistic understanding of protein structure has been improving over time, the most impressive feature of this research is the expansive breadth of the problems being investigated.

As discussed previously, protein evolution has long been a critical contributor to the holistic understanding of protein structure. Recent work, however, has benefitted greatly from the massive increase in the number and quality of structures deposited in the PDB. A primary concern in protein structure research is attempting to quantify the dynamics and range of conformations in the native state. To this end, recent work has been done to answer the question: "Does the variability in the conformations of a protein family match the variability in the conformations of a single member of the family?" In other words, can the dynamics of a protein family be used to approximate the dynamics of the native state for a single protein? It has been reported that the rate at which backbone flexibility diverges over time is low[65], and that similar levels of estimated flexibility can be used to identify distant relationships between proteins[66]; both of these findings suggest a link between

familial variability and individual variability. Others have investigated this computationally, using a method that relies on "backrub"[67] motions to create backbone variability in line with that observed in solution[68]. These authors found a link between both the solution state dynamics observed in NMR experiments and the conformational ensemble of multiple X-ray structures, as well as a link between the variation within a family and within the conformational ensemble of a single member of the family[68]. It has also been suggested that where the conformational variability differs between a family and a single member of the family, it could be due to selection for function, and thus valuable to understand in detail if one is interested in the function of that particular protein[69].

With the knowledge that the ensemble of solved structures for a given protein can begin to approximate its dynamics in solution[70], it has become more important than ever that related sets of structures be easily accessible; this requires consistent sorting and handling of meta-data about deposited structures. While the PDB has been a fantastic boon to the protein structure research community, it is not a perfect resource. In particular, it can be difficult to locate and obtain all the structures deposited for a single protein, and time consuming to quantify what makes them different in how they were determined. The Conformational Diversity of Native State (CoDNaS) database was created partially to address this issue[71]. This database reorganizes the PDB by protein, so one can easily obtain all known structures of a given protein. Helpful flags are associated with each model indicating the method it was determined by, if it has any mutations, if it was post-translationally modified, and if a ligand is bound. This database facilitates investigations into how a given structure fits into the larger context of all the other known conformations of that protein. It is also a great resource for researchers hoping to understand the native state and the extent to which dynamics and conformational variability can be sampled by our current methods for obtaining structural information.

As I have discussed above, thinking about the native state of a protein has moved towards considering ensembles of conformations, rather than just single conformations. For example, the refinement of NMR ensembles against data collected in solution can be improved by this principle. More accurate results can be obtained

by refining the properties of the entire ensemble against the experimental data, rather than an ensemble built from single models that were each refined against the experimental data one at a time[72,73]. Even building an ensemble from pairs of models (models refined two at a time) offers substantial improvements in accuracy compared to the more traditional approach, and can be a good compromise in light of the increased risk of overfitting that comes from multiple simultaneous model refinement[74]. The reality is that proteins are dynamic, complex molecules, and are better described by ensembles than single models. Unfortunately, even now, most experimental approaches still do not reflect that fact.

One branch of protein structure research where generating ensembles has never been a problem is in protein structure prediction; often, the problem lies in filtering the generated ensembles to only contain the structures likely to have high accuracy against experimental data (of course, one must be able to even generate *any* accurate models in the first place). Prediction can be done using homologous proteins with known structures as templates, as is often useful for solving the phase problem in X-ray crystallography[75], or *de novo;* this latter goal is one that if achieved would indicate a high level understanding of protein structure. The critical assessment of methods of protein structure prediction (CASP) initiative[76,77] has been instrumental in testing the ability of researchers to predict protein structures, with many different research groups coming together every other year to test their methods against a set of truly unknown structures[78–82]. The power of the CASP competition comes from the fact that each year its targets are selected from structures that have been (or will soon be) solved, but have not yet been shared with the community. This provides blind targets of varying degrees of difficulty, and thus a true test of a methodology's ability to accurately predict a structure. The results show a gradual but substantial improvement over time in our ability to predict protein structure[77]. The methods themselves are often quite varied, and can take inspiration from diverse sources (*e.g.* one successful method for predicting structures that have no solved homologous structure uses information about co-evolving residues extracted from large multiple sequence alignments[82,83]). Improvements to forcefields and sampling methods have also contributed greatly to successful prediction, often through the incorporation of

information from high-quality atomistic experimental data[84–87]. Lastly, protein structure prediction has found great utility in a fairly new field, protein design, where it has become possible to create structures that have never before existed in nature[88–92]. This is already leading to new insight into protein folding and stability, as principles of protein structure can be tested directly in a context free from evolutionary biases[93].

My Work

My own work on protein structure began with research that is not included in this dissertation. During my rotation in Dr. Andy Karplus' lab, I solved a crystal structure of a peroxiredoxin protein, taking the project all the way from purified protein to refined model. I did this work under the guidance of Dr. Arden Perkins, and the structure was recently included in a paper he co-authored[94]. This project not only gave me a positive first taste for structural biology, it also gave me a firm experimental background in crystallographic structure determination that was invaluable throughout the rest of my studies.

Within this dissertation, chapters 2 and 3 pertain to atomistic research into protein structure, and build heavily on the topics discussed briefly above. In chapter 2, "Native Proteins Trap High-Energy Transit Conformations", I discuss the atomic details of extremely unlikely backbone conformations observed in ultra-high resolution protein structures. Each of these conformations has a φ value that is close to 0, and as such, it is deep in a classically disallowed region of the Ramachandran plot. We determined that these stabilized high-energy conformations provide a series of "snapshots" showing the details of a conformational transition presumed to be ubiquitous during protein folding and conformational changes. These details would have otherwise been inaccessible, and in fact, even computational simulations do not accurately describe this transition.

In chapter 3, "On the Reliability of Peptide Non-Planarity Seen in Ultra-High Resolution Crystal Structures", I continue the conversation about peptide planarity present in the recent literature. In this chapter, I address each of the concerns raised in a report by Chellapa and Rose[29] that called into question the validity of the reported

structures having very non-planar residues, and I demonstrate the flaws present in that analysis. Furthermore, I describe strong evidence to further support the conclusion that most residues deviate from planarity in some way.

Chapter 4, I describe the main holistic research into protein structure that I have completed. This chapter is titled "Ensemblator v3: Robust Atom-level Comparative Analyses and Classification of Protein Structure Ensembles." It is difficult to overstate the importance of ensemble descriptions of protein structure. Not only is the ensemble a more accurate way of describing the native state of a protein, it is also much more information-rich than any single model can ever be. Despite this, ensembles tend only to be used until detailed analysis is required; at that point, researchers typically select "the best representative" single model, or the lowest-energy model, or even create an average model, to make comparisons and perform analysis on the structure. This is not due to any limitation inherent to ensembles, rather, it is a limitation inherent to the software that has been available and the conceptual approach taken by many researchers. In chapter 4, I address this problem. We have developed software, called the Ensemblator, that greatly expands the possibilities for analyzing and comparing ensembles of protein structures. Not only can it automatically find significant conformational subgroups within an ensemble, but it can also often pinpoint the exact locations of regions of difference or similarity between these subgroups that are the most significant, and thus potentially interesting to consider. In chapter 4 I describe this software in detail, as well as demonstrate its utility with a few case studies showcasing its most valuable features.

Finally, in chapter 5, I will reinforce the themes and main points of this dissertation, discuss the impact and future directions of my work, and conclude with some final remarks on our understanding of protein structure.

Chapter 2

Native Proteins Trap High-Energy Transit Conformations.

Andrew E. Brereton and P. Andrew Karplus

Published in *Science Advances*, 1, e1501188 (2015). 10.1126/sciadv.1501188

Abstract

During protein folding and as part of some conformational changes that regulate protein function, the polypeptide chain must traverse high-energy barriers that separate the commonly adopted low-energy conformations. How distortions in peptide geometry allow these barrier-crossing transitions is a fundamental open question. One such important transition involves the movement of a non-glycine residue between the left side of the Ramachandran plot (that is, $\phi < 0°$) and the right side (that is, $\phi > 0°$). We report that high-energy conformations with $\phi \sim 0°$, normally expected to occur only as fleeting transition states, are stably trapped in certain highly resolved native protein structures and that an analysis of these residues provides a detailed, experimentally derived map of the bond angle distortions taking place along the transition path. This unanticipated information lays to rest any uncertainty about whether such transitions are possible and how they occur, and in doing so lays a firm foundation for theoretical studies to better understand the transitions between basins that have been little studied but are integrally involved in protein folding and function. Also, the context of one such residue shows that even a designed highly stable protein can harbor substantial unfavorable interactions.

Introduction

Proteins carry out myriad functions that are enabled by their three-dimensional structures, and decades of research have led to over 100,000 structures in the Protein Data Bank (PDB[95]) and substantial understanding of protein folding and dynamics (e.g. [6,96]). In pioneering work, Ramachandran and coworkers[16] introduced the φ and ψ torsion angles to describe protein backbone conformations (see Fig. 2.1A,B), defining some conformations as "allowed" and others as "disallowed" due to collisions between atoms. Now, state-of-the-art energetics calculations[97] and the distributions of φ,ψ-angles seen in high-resolution protein structures[98,99] recapitulate remarkably well the main features of the original φ,ψ-plots. For alanine-like residues (Fig. 2.1D), these include two well-populated low-energy regions – typically called the α and β basins –

on the left-hand side of the plot (having φ<0°) and a single smaller reasonably populated low-energy basin – called α_L – on the right-hand side (having φ~+60°).

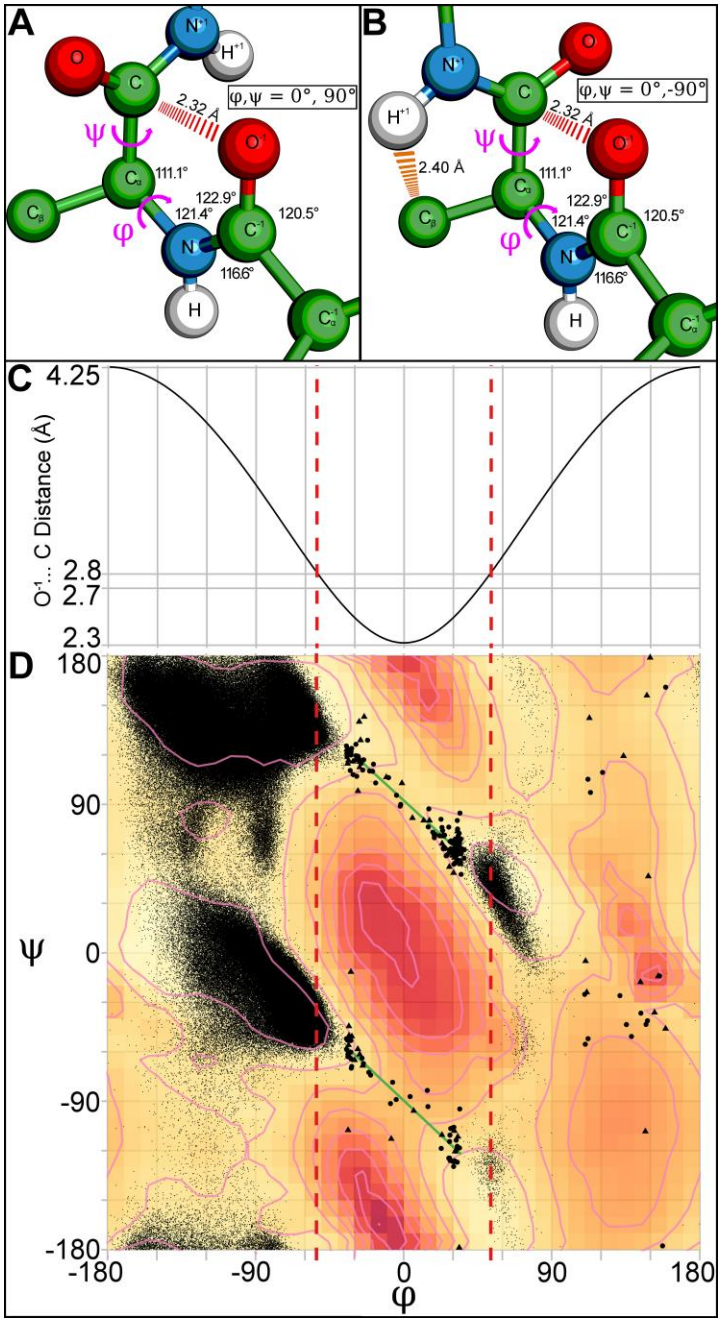

Figure 2.1 The populated high energy passes for transitions between φ<0° and φ>0° conformations.

(A) A standard geometry[15] alanine dipeptide with φ,ψ =0°,+90°. Indicated are the positive rotation direction for the φ and ψ torsion angles (magenta), the standard values for the five

backbone bond angles not involving Cβ (black), and the O^{-1}...C clash (red dashes with distance). **(B)** An alanine dipeptide, as in panel A, but with φ,ψ=0°,-90°. The H^{+1}…Cβ approach (orange dashes with distance) also shown matches the "normal" close approach limit of 2.4 Å for these atoms[17] and causes the ψ~-90° transition track to be somewhat more unfavorable than the ψ~+90° track (~7 vs ~5 kcal/mol as seen in panel D). **(C)** O^{-1}...C distances as a function of φ for standard geometry alanine dipeptides. The expected "normal" (2.8 Å) and "extreme" (2.7 Å) approach limits[17] are indicated; red dotted lines at φ=±53° mark where the "normal" approach limit is crossed. **(D)** A Ramachandran plot with energy contours for the alanine dipeptide calculated using an adaptive force biasing algorithm[97] displayed in steps of 2 kcal/mol (pink). Also shown (small black dots) are 616,212 non-glycine residues from representative ≤1.5 Å resolution structures; of these 16,613 (or ~3%) have φ<0°. Reliable (large circles) and unreliable (large triangles) observations between -35°<φ<+35° and +110°<φ<+160° are highlighted. The best fit lines for reliable residues between -35° < φ < +35° are shown for both the ψ=+90° and ψ=-90° passes (green).

While much study has been devoted to the geometries and relative energetics of the well-populated basins (e.g.[100]), much more difficult to study (e.g.[86,97]), and still poorly understood, is how alanine-like residues cross the high-energy barriers near φ=0° or +135° (Fig. 2.1D) that match classically disallowed regions and separate the common conformations having φ<0° from those having φ~+60°. As estimated by Guvench *et al.*[97] the heights of the barriers between the basins are about ~5-7 kcal/mol (Fig. 2.1D). These barriers are much lower than the ~20 kcal/mol barrier associated with *cis-trans* isomerization of proline that can be rate limiting for folding[101], and so the transitions should not be rate limiting but rather common occurrences during protein folding. Such transitions also have been seen to be important for regulatory conformational switches that govern the function of certain proteins, such as modulating peptide binding by an SH2 domain[102], or switching between the low and high affinity states of the cell adhesion mediator CD44[103].

As noted above, to transition between the populated conformations having φ<0° or φ~+60°, a residue must cross one of the two high-energy swaths near φ=0° or +135°. These regions were classically described as disallowed because of collisions between the carbonyl carbon (C) or the Cβ-carbon, respectively, and the peptide

oxygen of the previous residue (O^{-1}). For example, using standard peptide geometry[15] the O^{-1}...C approach at $\varphi = 0°$ is 2.32 Å (Fig. 2.1A-C), much closer than the expected extreme contact limit of 2.7 Å[17]. Like all transition states, these high-energy transit conformations are expected to be only fleetingly populated and inaccessible to direct experimental characterization, so that there cannot be certainty about what the transition structures really look like. Contrary to this expectation, we have discovered and describe here high resolution observations of a series of conformations that that have been trapped in native protein structures deposited in the PDB and that cover the full range of the $\varphi \sim 0°$ transitions. The analysis of these observations provides an experimentally-derived detailed map of the geometric distortions that take place during these conformational transitions.

Results

Reliably-modeled residues exist in the two high-energy passes near $\varphi=0°$

While surveying the conformations of residues in high resolution ($\leq$ 1.5 Å) protein structures, we were surprised to discover two narrow strings of observations that span completely across the classically disallowed transition regions near $\varphi=0°$ (an upper one with $\psi \sim +90°$ and a lower one with $\psi \sim -90°$) as well as a few sporadic observations in the regions near $\varphi=+135°$ (Fig. 2.1D). The existence of residues adopting conformations in the two "mountain passes" through the $\varphi \sim 0°$ high-energy landscape can be seen in some previously published Ramachandran plots (eg.[104,105]), but the reliability and potential importance of these residues has to our knowledge not been investigated. Even a recent paper explicitly focused on describing residues in sparsely populated regions of the Ramachandran plot made no mention of these residues consistent with them not being considered as reliably observable[104].We carried out visual checks of each of the putative transition residues against its electron density (e.g. Figs. 2.2 and 2.S1) and found that the majority are reliably defined (Fig. 2.1D, circles). Importantly, the reliably defined residues all have φ,ψ-angles falling roughly within the predicted lowest energy passes through the high energy terrain (Fig. 2.1D). Furthermore, as might be anticipated, the observed residues having $\varphi \sim 0°$

that are not in the low-energy passes were found to be the result of incorrect or unreliable modeling (Fig. 2.1D, triangles). Since the reliably determined residues with high-energy conformations near $\varphi = 0°$ are real and relatively abundant (146 observations in the $-35° < \varphi < 35°$ transit zone; see Table 2.S1), they represent fortuitous "natural experiments" that provide an unprecedented ability to experimentally define at high resolution exactly how the standard peptide geometry becomes distorted as a residue passes through these highly strained conformational transition states. Although 15 residues in the passes near $\varphi \sim +135°$ are also well defined (Fig. 2.1D), those populations are not yet sufficiently large to enable an accurate description of the pathways they represent.

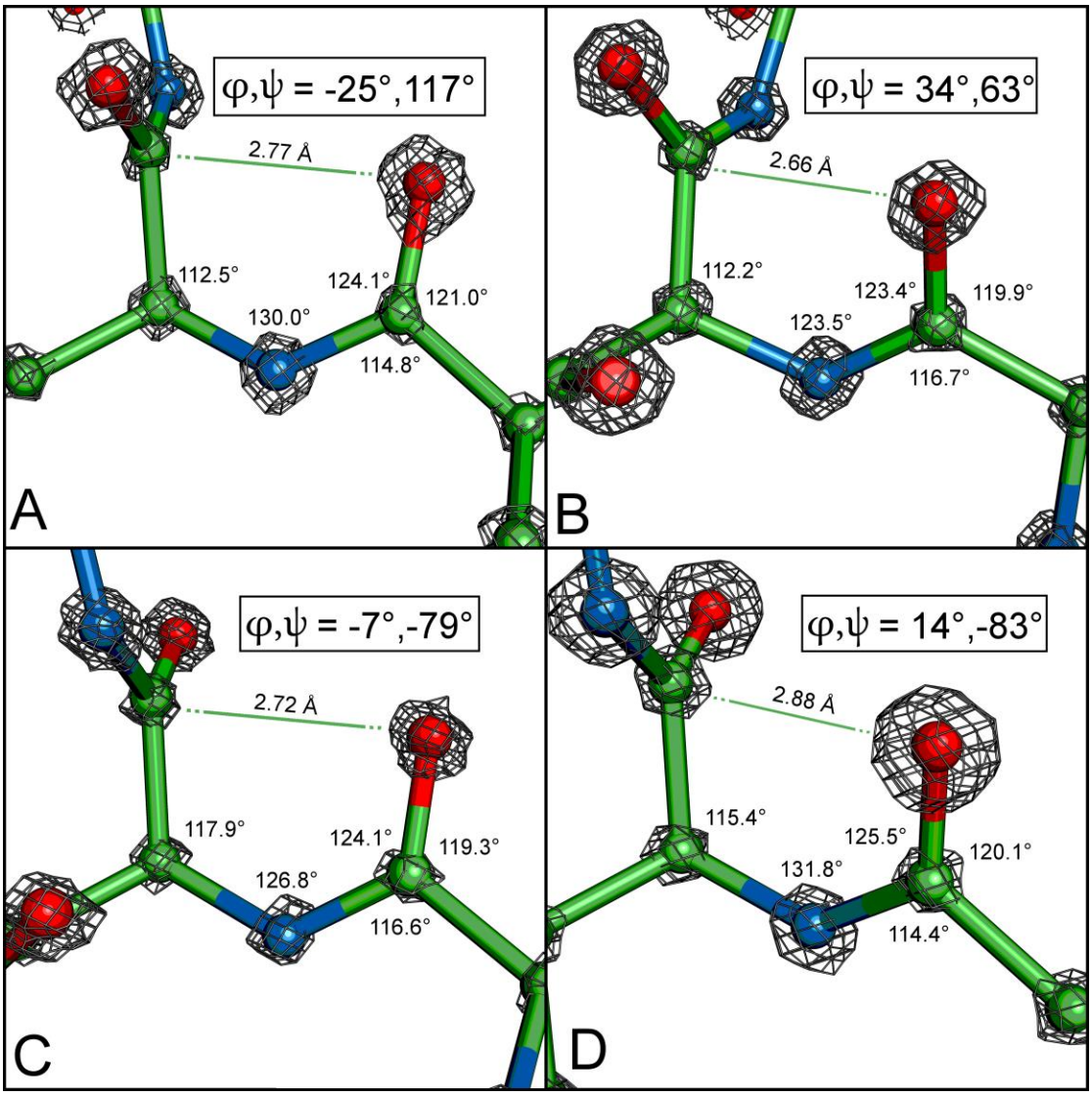

Figure 2.2 Electron density evidence for four residues adopting conformations in the -35° < φ < +35° range.

Each panel shows a residue with its $2F_O$-F_C electron density, its backbone bond angle values (black), its φ,ψ angles (inset box) and its O^{-1}...C approach (green line with distance). **(A)** His261from PDB entry 4N1I (1.0 Å resolution; contoured at 6.2 x root-mean-square electron density ($\boldsymbol{\rho}_{rms}$)). **(B)** Ser115 from PDB entry 2DDX (0.86 Å resolution; contoured at 7.0x$\boldsymbol{\rho}_{rms}$). **(C)** Asp249 from PDB entry 4AYO (0.85 Å resolution; contoured at 7.0x$\boldsymbol{\rho}_{rms}$). **(D)** Ile152 from PDB entry 3NOQ (1.0 Å resolution; contoured at 5.5x$\boldsymbol{\rho}_{rms}$).

The φ~0° transition residues exist in diverse contexts

The φ~0° transition residues trapped in native proteins exist in a variety of conformational contexts (Fig. 2.S2), and are distributed among 17 of the 20 standard residue types (Table 2.S2), implying they are not special cases, but represent realistic snapshots along a transition pathway. Many of these residues are present in or near active sites, but others are not (e.g. Fig. 2.S3). The cases occurring in two proteins are particularly instructional. In one case, the occurrence proves that even a small, highly stable, designed, helical bundle with a melting temperature of 105 °C can accommodate a residue with such high local strain energy (Fig. 2.S3A). In the second case, it has been shown that a simple Cys-to-Ala mutation that removes a single hydrogen bond in the active site of an isocyanide hydratase (Fig. 2.S3B) leads to a rearrangement of short backbone segment and the loss of the high energy conformation[106]. Furthermore, it was also shown that a Cys-to-Ser mutation that strengthened the hydrogen bond actually enhanced the stability of the segment in the native conformation[106]. This example implies that the energy cost for a residue adopting a high energy transition conformation can apparently be offset by the formation of a single hydrogen bond and the rearrangement of a few residues.

Mapping the φ-dependent distortions involved in the transitions

On a Ramachandran plot the strip of observations near ψ=–90° nearly perfectly matches through inversion symmetry that near ψ=+90° (see Fig. 2.1D, green lines, and Table 2.S3), making it reasonable to treat the two passes as a single phenomenon, roughly doubling the density of observations available for mapping the barrier crossing. In order to define the patterns of distortion that allow peptides to traverse this barrier, we calculated φ-dependent average values for the O^{-1}...C distance and all backbone bond angles. Given the diverse contexts of the residues, treating them as an aggregate should average out specific features due to each particular context and provide a view of the generic transition properties that are solely due to local factors, and are generally relevant. This is supported by previous studies showing that the average conformation-dependence of backbone bond angles

and planarity, found in ultra-high resolution protein structures, agrees well with those from quantum mechanics calculations of simple model compounds and those from structures of small peptides[19,23,107,108].

The behavior of the O^{-1}...C distance is quite striking (Fig. 2.3A). The average values near $\varphi=\pm60°$ track with the distance expected for standard geometry, until the distance reaches 2.8 Å (near $\varphi\sim\pm50°$), then distortion begins and the average distance decreases much less rapidly than predicted by standard geometry, until it reaches ~ 2.7 Å (near $\varphi=\pm25°$); then, between φ of -25° and +25°, the observed distance is remarkably flat, with the average distance of 2.68±0.02 Å over that range matching remarkably well with the 2.7 Å "extreme approach limit" for these atom types defined nearly 50 years ago[17].

The φ-dependent variations of the backbone bond angles are also quite systematic, with each angle roughly matching its standard value at $\varphi=\pm60°$ and varying smoothly to its maximal deformation at $\varphi=0°$. Only three bond angles expand substantially and they are the $\angle O^{-1}$-C^{-1}-N, $\angle C^{-1}$-N-Cα, and ∠N-Cα-C angles, which expand roughly 2°, 6°, and 4°, respectively (Fig. 2.3B). The lesser expansion of the $\angle O^{-1}$-C^{-1}-N angle is consistent with the expectation that as a purely sp^2-hybridized center it would have a higher force constant for resisting distortion. Given the expanding $\angle O^{-1}$-C^{-1}-N angle, to keep the C^{-1} carbonyl group largely planar the $\angle C\alpha^{-1}$-C^{-1}-O^{-1} and the $\angle C\alpha^{-1}$-C^{-1}-N angles decrease in a coordinated fashion by ~1° and 1.5°, respectively.

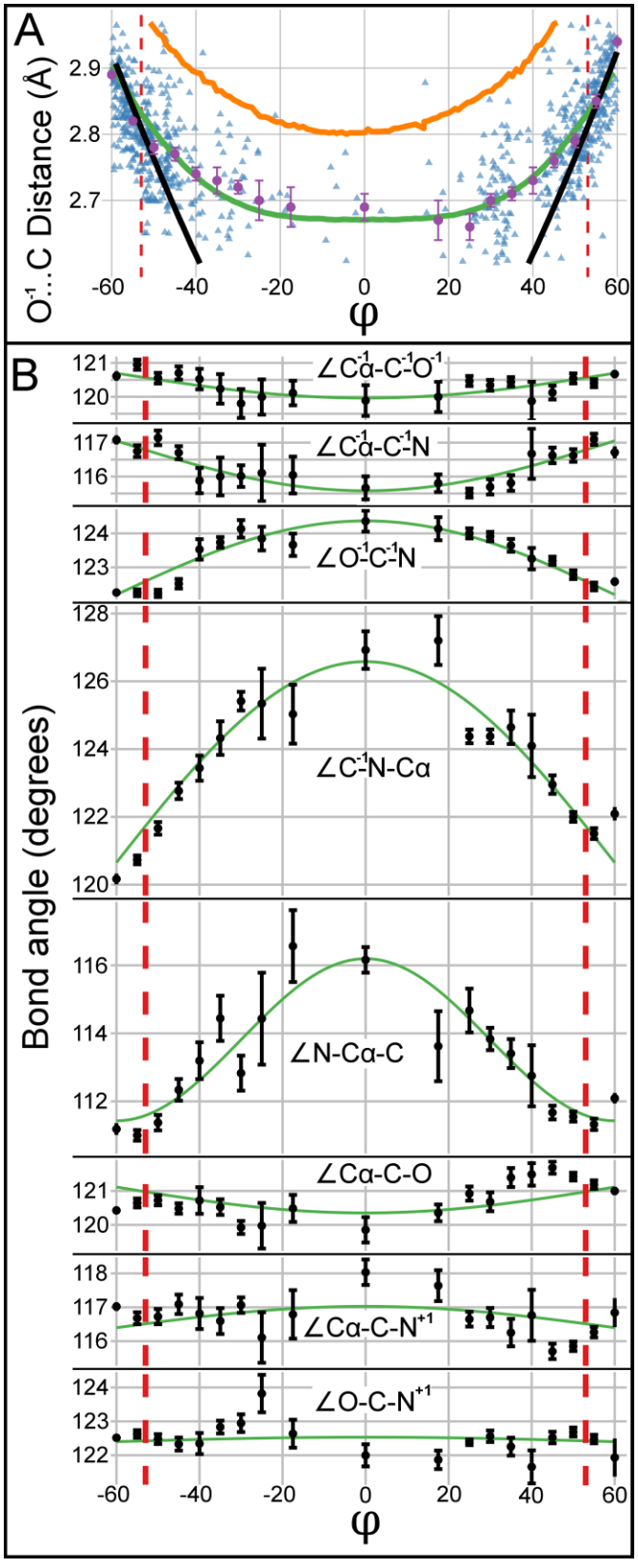
A
O-1...C Distance (Å)
2.9
2.8
2.7
-60
-40
-20
0
20
40
60
φ
B
Bond angle (degrees)
∠Cα-C-1O-1
∠Cα-C-1N
∠O-1C-1N
∠C-1N-Cα
∠N-Cα-C
∠Cα-C-O
∠Cα-C-N+1
∠O-C-N+1
φ

Figure 2.3 Systematic deformations of geometry associated with transition through the high energy φ~0° passes.

(A) Observed average O^{-1}...C distance (large purple dots and error bars) plotted as a function of φ (see Table 2.S4 for details), along with each data point (blue triangles), and the O^{-1}...C distance predicted by standard geometry (black curve), by the empirically defined φ-dependent geometry functions (green curve), and by the AMBER FF99SB force field (orange curve). Dotted red lines at φ=± 53° are as in Fig. 2.1B. **(B)** Average backbone bond angles (black dots with error bars) as a function of φ (see Table 2.S4 for details) along with cosine functions fit to the data (green curves; Table 2.S3 for the equations). All error bars are standard error of the mean.

To check the validity of treating the ψ~+90° and ψ~–90° passes as equivalent, we analyzed the data from the two passes separately, and found that all angles behaved similarly, except that for the ψ~–90° transition, the ∠Cα-C-N^{+1} angle also expands ~2° (Fig. 2.S4), as makes sense to minimize the clash between the N^{+1} hydrogen and Cβ (Fig. 2.1B). We note that for two reasons these empirical bond angle distortions may slightly underestimate the actual average distortions: first, structures in the 1.0 to 1.5 Å resolution range are still somewhat influenced by refinement restraints tethering them to the standard values[23,30], and second, at φ=0° – the point of expected maximal distortion – the empirical value is, due to limited data, an average over the rather broad φ range of –12.5˚ to +12.5˚ (Table 2.S4).

An analytical model for the transition and a comparison with molecular mechanics

These observed φ-dependencies of the backbone bond angles were modeled as a set of smooth conformation-dependent functions (Fig. 2.3B, green curves; Table 2.S3) that could be used to generate prototype models for the conformational transition. That these yield O^{-1}...C distances (Fig. 2.3A, green line) matching reasonably well with the empirical averages supports the validity of these functions as a capturing a realistic general model for how the φ~0° transition is traversed. As noted above, the variations of the individual observations from the average behavior (such as in the examples shown in Figure 2.2) are not primarily due to experimental uncertainty, but are real variations reflecting the forces caused by the unique tertiary environments that stabilize the transit conformations.

To assess how accurately a state of the art molecular mechanics force field handles these high energy transition conformations, AMBER and the FF99SB force field (recently demonstrated[109] to perform best in a protein modeling test) were used to minimize conformational energy while restraining φ and ψ to the values along the upper narrow transit path. The energy minimized O^{-1}...C separation distances (Fig. 2.3A, orange line) and backbone bond angles (Fig. 2.S5) showed qualitative similarity to the empirical variations, but were not in good quantitative agreement: the limiting O^{-1}...C approach was ~0.15 Å too high and four bond angles had notable systematic displacements from the empirical values, with the largest difference of ~4° occurring for the ∠C^{-1}-N-Cα angle (Fig. 2.S5). These discrepancies imply that the empirical conformational details defined here for the φ~0° high-energy conformations are not just confirming what is already well-understood, but represent a rare resource for enhancing force field parameterizations.

Discussion

The observation of these conformations and their conformation-dependent bond angles represent a remarkably detailed experimental characterization of these important conformational transition states that had not been thought to be accessible to direct observation. These are not transition state analogs, artificially held in place by a covalent modification that might alter the pathway, but they are authentic residues that are free to transition through the barrier, yet are stabilized part way through by non-covalent interactions with their environment. The fact that the φ,ψ angles of the observed transition residues match so well with the low energy pathway calculated for an isolated dipeptide (Fig. 2.1D) supports the conclusion that neither the specific protein environments nor the cryogenic temperatures at which most of the structures were determined has changed the nature of the pathway.

In one sense, these images contribute to our understanding how this transition occurs in the same way that Eadweard Muybridge's striking "series of instantaneous photographs" of horses provided information previously considered unobservable and showed "with absolute accuracy the motions of horses when walking, trotting, and running"[110]. These proved that all four legs of the horse are off the ground roughly

half of the time even during a trot. Similarly, the observations presented here provide indisputable evidence that proteins truly can adopt these unfavorable φ~0 conformations, and can, based on direct observation, reveal in high resolution detail the nature of the bond angle deformations that are involved. Just as Muybridge's photographs strung together could provide an observation-based movie of a horse in motion, our empirically-derived analytical functions allow us to generate such a movie of a peptide traversing the mountain pass (Movie 2.S1).

In contrast, although molecular simulations are powerful, if simulations were the only source of information, many uncertainties would remain. One illustration of this are the discrepancies between the approach distances and distortions observed here and those predicted by the AMBER force field (Figs. 2.3A and 2.S5). Another is a molecular dynamics study of the conformational switch in the SH2 domain for which a residue goes from the α_L basin (φ,ψ ~ +60°, +60°) to the β basin (φ,ψ ~ -60°, +120°). Acknowledging they could not be certain which was the preferred path, the authors proposed based on a lower predicted energy in their molecular mechanics calculations that the residue traversed the longer path through the high energy pass near φ~+135[102]. Our results suggest that the shorter path through the mountain pass at φ,ψ~+0°,+90° should be reconsidered as an *a priori* more likely path.

In terms of the larger picture of protein folding and function, these analyses bring a new clarity about how this fundamentally important transition occurs and the level of distortions that peptides are subject to. As such, they provide a foundation for future investigations of the important, but little studied area of high energy barrier crossings and open the door for a richer understanding of folding routes and conformational transitions. On a practical level, this work also provides conformation-dependent restraints, similar to those previously developed for well-populated areas of the Ramachandran plot[23,30] that can both guide force field development and enhance the accuracy that can be achieved in experimental (e.g.[24]) and predictive (e.g.[84,85]) modeling of proteins having residues adopting these rare but important conformations in the φ~0° transition region. Finally, this study holds the promise that other high energy transition conformations can be similarly characterized as the size of the PDB increases and more such observations

accumulate.

This work also provides some insight into the thermodynamics in native proteins. It is well-known that naturally occurring proteins are not optimized for stability, and this has recently been dramatically illustrated by the creation of a set of designed proteins adopting five different folds and having melting temperatures above 95 °C[88] and of similarly designed set of super-stable helical bundles, one of which had a stability of ~60 kcal/mol and a melting temperature above 135 °C[89]. In the latter report, it was concluded that "low-energy structures must have unstrained backbone conformations…"[89], but this is clearly not the case given that one of the proteins in our sample was a designed highly stable protein (Fig. 2.S3A). That example, and the other example noted above in which the high energy conformation was apparently stabilized by the folding of just a small segment of the protein (Fig. 2.S3B), emphasize two things. One is that the potential stability achievable by a folded protein is so high that even highly stable proteins may still contain many suboptimal and even some highly unfavorable interactions. A second is that suboptimal interactions (i.e. "frustration") present in native proteins need not only be present in the form of many slightly unfavorable interactions, but can also include individual interactions that are even as high as 5-7 kcal/mol destabilizing.

Materials and Methods

Protein geometry database searches

The dataset plotted as small dots in Figure 2.1D was created using the Protein Geometry Database (PGD)[28], and it includes 616,212 non-glycine residues. Each of these residues is at the center of a three residue segment that has backbone, average side chain, and γ-atom B-factors ≤ 25 Å^2, and is present in a protein crystal structure refined to $R_{work}/R_{free} \leq 0.2/0.25$ at a resolution of 1.5 Å resolution or better, and from a protein having $\leq 90\%$ sequence identity to any other structure in the set. Another, smaller dataset was generated using a $\leq 25\%$ sequence identity cutoff, in order to get amino acid frequencies that are representative diverse sets of proteins (Table 2.S2).

Manual curating of the observations in the high-energy passes

Based on the above search, all observations having their φ torsion angle in one of the high energy pass regions, either $-35° < \varphi < +35°$ or $110° < \varphi < 160°$, were manually curated as to the reliability of their conformation based on a visual assessment of the fit to their electron density map. Using conservative criteria, each residue was designated as either reliable (shown as large black dots in Figure 2.1D) or unreliable (shown as triangles in Figure 2.1D). Residues designated as reliable had to have strong, well-defined, and not highly anisotropic electron density and a model that was well-fit in that density. Observations leading residues to be deemed unreliable also included the presence of alternate conformations or a close association with uninterpreted density that might indicate alternate conformations. These criteria erred on the side of possibly excluding residues that may have been accurately modeled, rather than including any residues that might not be accurately modeled.

Generation of modeled peptide structures

All peptides were generated using the PeptideBuilder python program and library[111] which was slightly modified to be able to handle φ-dependent equations instead of single-value standard geometries.

Calculations of the protein geometries

The set of curated observed residues in the $-35° < \varphi < +35°$ range output by the PGD was used as input for a custom script written in R, which made use of the Bio3D[112] package to read PDB files, and then calculate specific geometric details for all residues of interest for each protein, excluding any residue not having at least two residues on both sides of it without a chain break. The quantities calculated included all of the relevant backbone torsion angles and bond angles and the $O^{-1}…C$ distances. Bond lengths were not analyzed as it has been shown that in crystal structures at these resolutions and conformation-dependent bond length variations are so small as to not be reliably determined and to not crucially impact modeling accuracy[23,24]. Even those quantities available from the PGD search were recalculated so that the information

could also be obtained for the non-crystallographic symmetry (ncs) mates of the PGD hits (which are not present in the PGD). This allowed the 100 unique and curated residues having φ in the -35° to +35° range to be expanded by the addition of 46 ncs chains (that were also manually curated and deemed as reliable) making for 146 total observations.

Statistical analyses and least squares modeling of the data

All averages and standard errors of the mean were calculated using conventional formulae written in R. Best fit lines in Fig. 2.1D were generated using principle component analysis to fit an orthogonal linear regression, due to experimental uncertainty in both x and y values. The function prcomp() in R was used, where:

```
> x = phi
> y = psi
> r = prcomp( ~ x + y )
> slope = r$rotation[2,1] / r$rotation[1,1]
> intercept = r$center[2] - slope*r$center[1]
```

Independent best fit lines were calculated for each transit region separately, and there was negligible difference between the two (Table 2.S3, Fig. 2.1D). The dependence of bond angles on φ were fit using the geom_smooth() function from the R package 'ggplot2', while specifying "formula = (y~ I(cos(x * pi/120)))", where x is the central value in the phibin, and y is the mean value of the bond angle.

AMBER minimizations

AMBER calculations were performed using AMBER12[113] and the FFSB99 force field. Peptides were capped with N-terminal acetyl (ACE) and C-terminal amide (NME) groups. SANDER minimizations were done every 1° in φ over the range of -60° to +60° with the φ and ψ torsion angles restrained using "NMR restraints" of 300.0 kcal/mol*rad and the ψ target value set according to the best fit line given in

Table 2.S3 for the ψ > 0 pass. Minimizations were carried out two ways, once starting from standard backbone bond angles and another time starting from the φ-dependent backbone bond angles as defined in the equations in Table 2.S3. For all calculations the dielectric constant was set to 80 and minimizations were run for 2000 cycles. The results based on both starting points were equivalent so only one is shown in Figures 2.3A and 2.S5.

Acknowledgements

This work was supported by NIH grant R01-GM083136 (to PAK). We thank Dr. Olgun Guvench for providing the data for the energy contours in figure 2.1D. The authors declare that they have no competing interests. All data is available in the Protein Data Bank, with specific details reported in supplementary table 2.S1. The project was conceived by PAK, experiments, analyses and figure preparation were done out by AEB, and writing was done by AEB and PAK.

Supplementary Materials

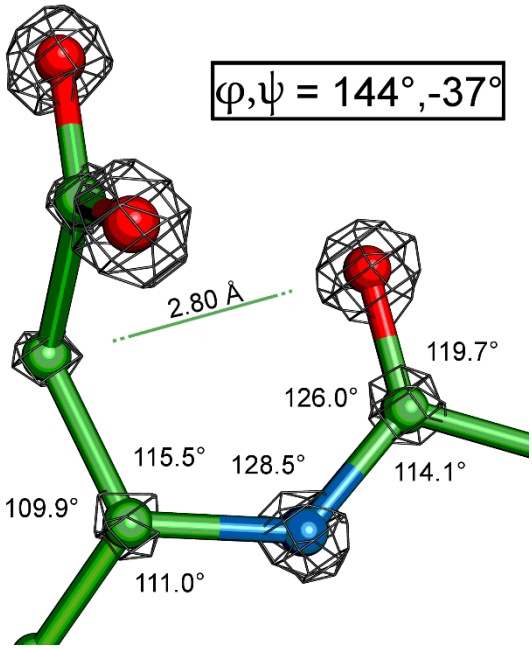

Figure 2.S1 Electron density evidence for a reliable residue adopting a conformation in the $+110° < \varphi < +160°$ range.

As in Figure 2.2, shown is residue Asp236 from PDB entry 2FVY, with resolution 0.92 Å, along with its $2F_O$-F_C electron density map contoured at 6.5 x $\boldsymbol{\rho}_{rmsd}$. As in Figure 2.2, the values of its φ,ψ angles, the backbone bond angles, and the O^{-1}... C approach (green line with distance) are also given.

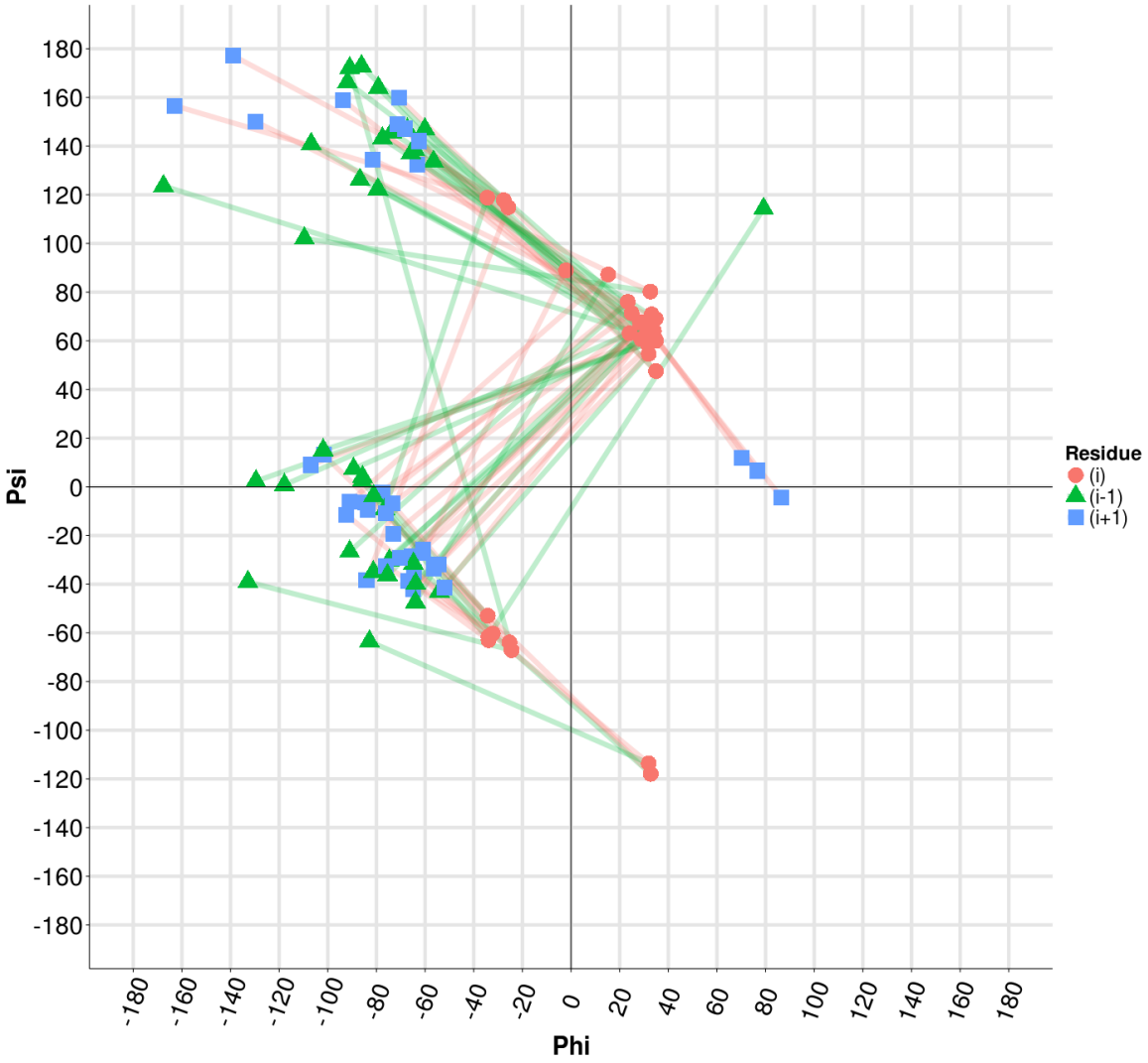

Figure 2.S2 φ,ψ angles describing the local conformational context of the mountain pass residues.

Shown are the φ,ψ angles for each residue of interest (residue i; red circles) and the previous (i-1; green triangles), and following (i+1; blue squares) residues. For simplicity, data are only shown for a representative subset of the structures analyzed, which are those from proteins

having ≤25% sequence identity with any other protein in the complete dataset. These residues are indicated with an asterisk (*) in the "Code" column of Table 2.S1. Green lines connect the (i-1) to (i) symbols, and red lines connect the (i) to (i+1) symbols. Although the observations are diverse, there are three common patterns: firstly, those somewhat similar to type II turns with residues (i-1) to (i) going from the P_{II} region (broadly φ = -60˚:-100˚ and ψ = +120˚:+180˚) to the ψ = +90˚ transit region; secondly, going from the alpha/delta region (broadly φ = -50˚:-120˚ and ψ = -60˚:+20˚) to the ψ = +90˚ transit region; and thirdly, those somewhat similar to type II' turns, with residues (i) to (i+1) going from the ψ = -90˚ transit region to the alpha/delta region.

Figure 2.S3 A-D: Four representative examples of $\varphi \sim 0$˚ conformation residues chosen from among the 8 cases that are closest to $\varphi = 0$.

These show how different contexts can be in which this conformation is stabilized. Each panel has two parts. The first part shows a ribbon diagram of the overall protein, with the residue of interest indicated in purple and nearby waters depicted as red spheres. Chain A is shown in dark green, and, if present, chain B is light green. The second part shows a zoomed in stereo-view focused on the residue of interest. The model is shown as a ball and stick diagram, with atom coloring and green carbons for most residues and purple carbons for the $\varphi \sim 0$˚ residue. Waters are red spheres and hydrogen-bonds are shown as yellow dashed lines. Also, each panel has its own legend providing further details of the structure shown.

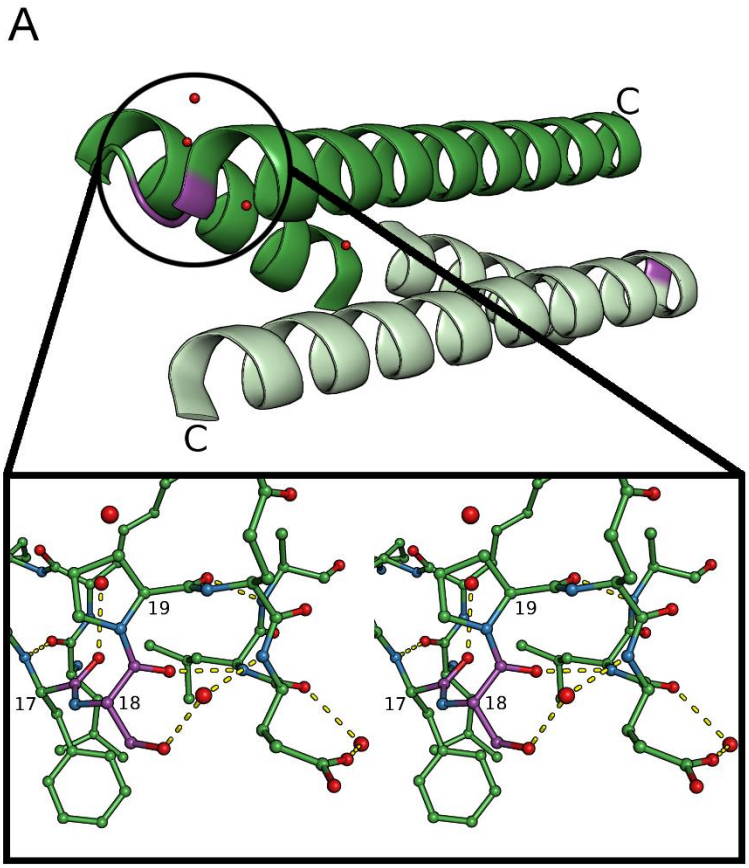

Figure 2.S3A details (see Fig. 2.S3 legend above for broader description of what is shown):

Shown is the context of residue Ser18 in chain A of PDB code 1G6U (determined at 1.5 Å resolution) with φ,ψ = 14.9º,76.3º (and φ,ψ = 17.2º,85.2º for Ser18 in Chain B). The protein is a designed domain swapped dimer[114] that is not a natural protein and has no active site. The residue of interest is highly surface exposed and exists in a tight α-α hairpin where the Ser side chain makes a water-mediated N-capping interaction with following helix. The whole protein is a 48 residue long low-complexity sequence (contains only Ala, Leu, Glu, Gln, Lys, and Ser plus one Phe and one Pro). Despite the small size of the protein and the presence of the strained conformation, the dimeric protein is highly stable, having an estimated melting temperature of 105 ˚C. The unusual with φ,ψ-angles of Ser18 were not mentioned in the original publication.

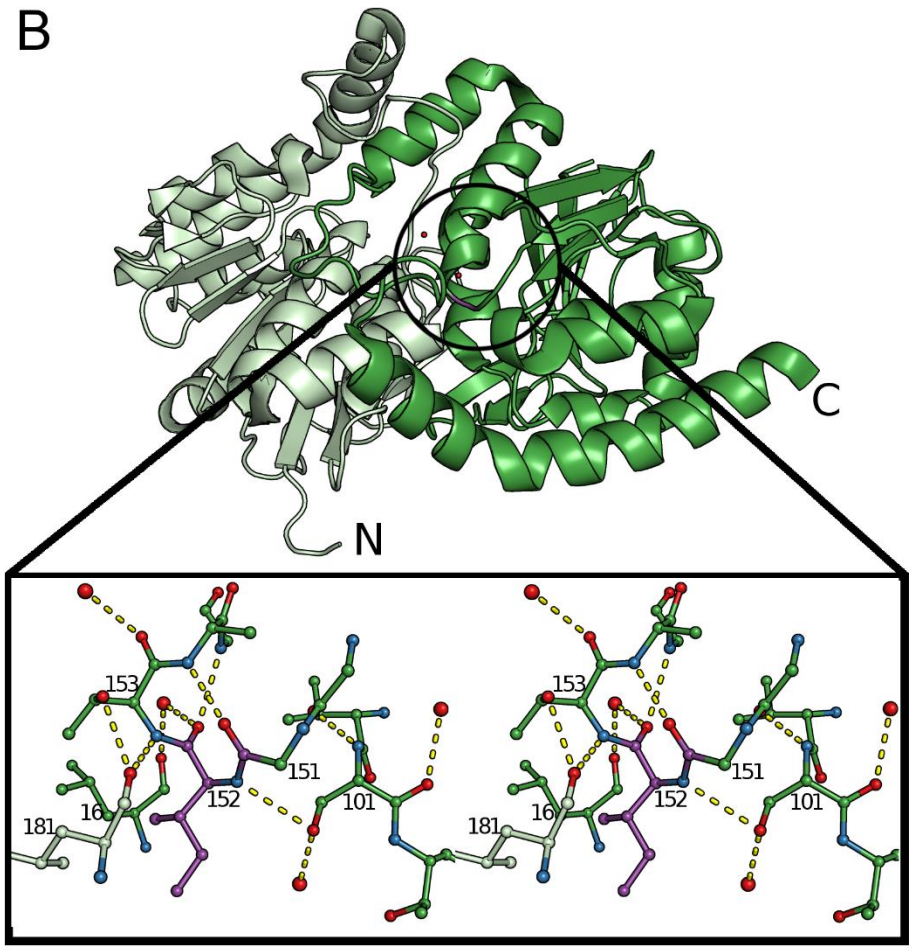

Figure 2.S3B details (see Fig. 2.S3 legend above for broader description of what is shown):

Shown is the context of residue Ile152 in chain A of PDB code 3NOQ (determined at 1.0 Å resolution) with φ,ψ = 14.4º,-83.5º (and φ,ψ = 17.2º,-85.2º for the equivalent residue in Chain B; numbered Ile154). This structure is a Cys101-to-Ser mutant of isocyanide hydratase from *Pseudomonas fluorescens*[106],and Ile152 is at the active site with its backbone amide making a hydrogen bond to the Oγ-atom of Ser101 in this structure (shown) and to the Sγ-atom of Cys101 in the predominant conformation of the wild-type enzyme (PDB entry 3NON). As described in the original report (see figure 5 of reference[106]), in the crystal structure of an inactive Cys101-to-Ala mutant (PDB entry 3NOO) this hydrogen bond is absent and residues 149-153 undergo a conformational change so that the strained φ ~ 0˚ conformation is not adopted. The authors further note that in the wild type structure this region appears to be conformationally polymorphic with the strained conformation being the major one and a C101A-like conformation being the minor one. Since the conformational change is local, this

indicates that it does not take many favorable interactions to offset the energetic cost of adoption of the outlier φ,ψ-values.

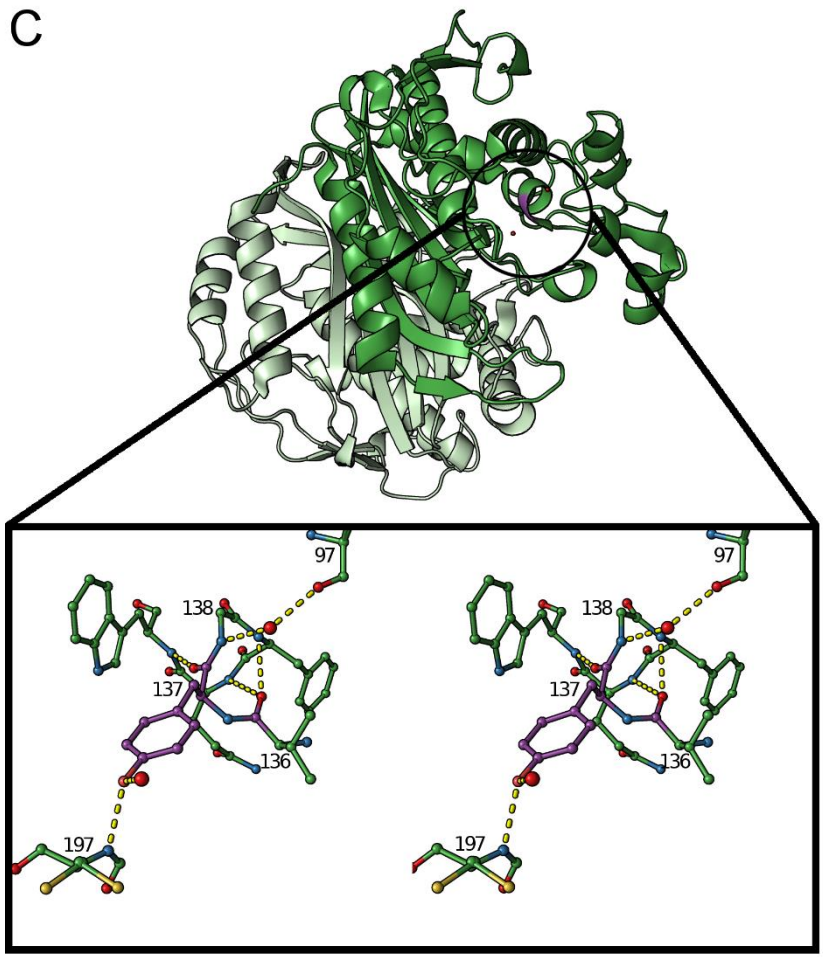

Figure 2.S3C details (see Fig. 2.S3 legend above for broader description of what is shown):

Shown is the context of residue Tyr137 in chain A of PDB code 4IQB (determined at 1.13 Å resolution) with φ,ψ = 3.9º,-87.2º (and φ,ψ = 14.8º,-93.3º for the equivalent residue in Chain B). This structure is thymidylate synthase from *Caenorhabditis elegans*, and it has not yet been described in the literature. Tyr137 is in a hydrogen-bonded turn at the start of a short α-helix, and in this unliganded complex, its side-chain is in a tight turn at the start of a short α-helix and its side chain interacts with a water molecule and the amide of residue 197. In another structure of this enzyme in complex with dUMP and the inhibitor Tomudex (PDB

entry 4IQQ; 2.9 Å resolution) solved by the same research group, the backbone adopts a similar conformation and the side chain is seen to be close enough to the active site to hydrogen bond, via the water, to the uracil of dUMP.

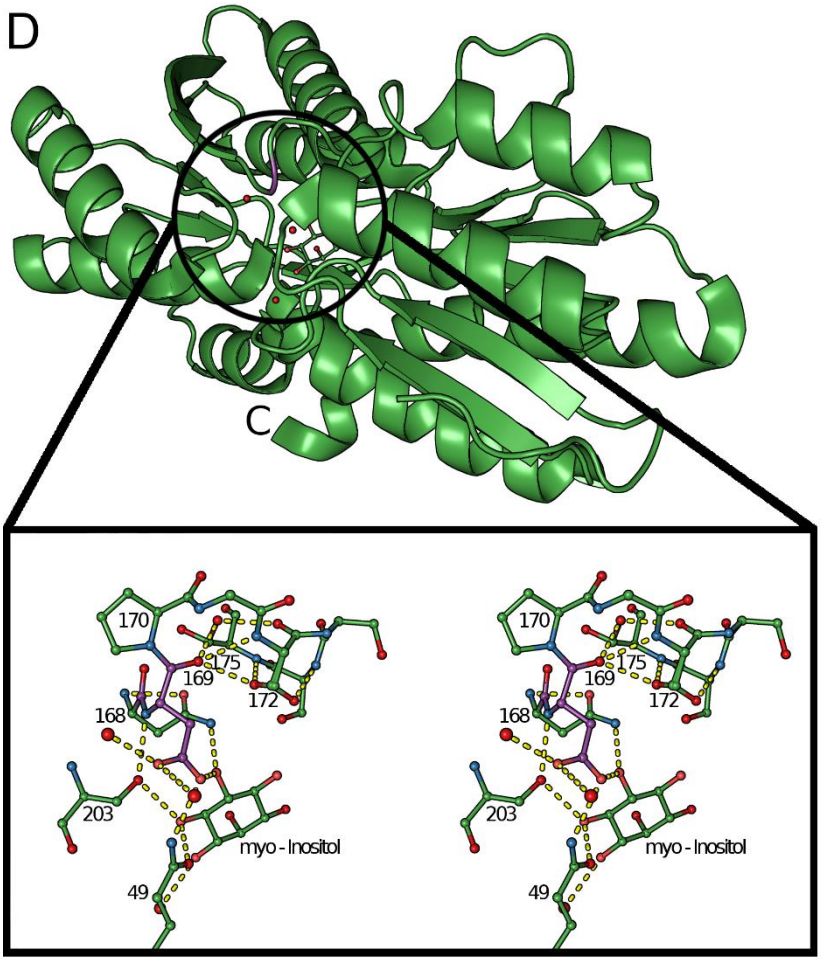

Figure 2.S3D details (see Fig. 2.S3 legend above for broader description of what is shown):

Shown is the context of residue Asp169 in chain A of PDB code 4IRX (determined at 1.45 Å resolution) with $\varphi,\psi = -2.1^{o},+88.9^{o}$ (and $\varphi,\psi = -1.8^{o},+90.0^{o}$ for the equivalent residue in Chain B). This structure is the myo-inositol binding protein from Caulobacter crescentus in complex with myo-inositol[115]. In this structure, the side chain of Asp169 is central to the ligand binding site, making hydrogen bonds with the myo-inositol as well as Gln49. The backbone conformation that also allows Asn168 to hydrogen bond to the myo-inositol is further stabilized by hydrogen bonds with two Ser side chains (172 and 175) that both adopt alternate conformations.

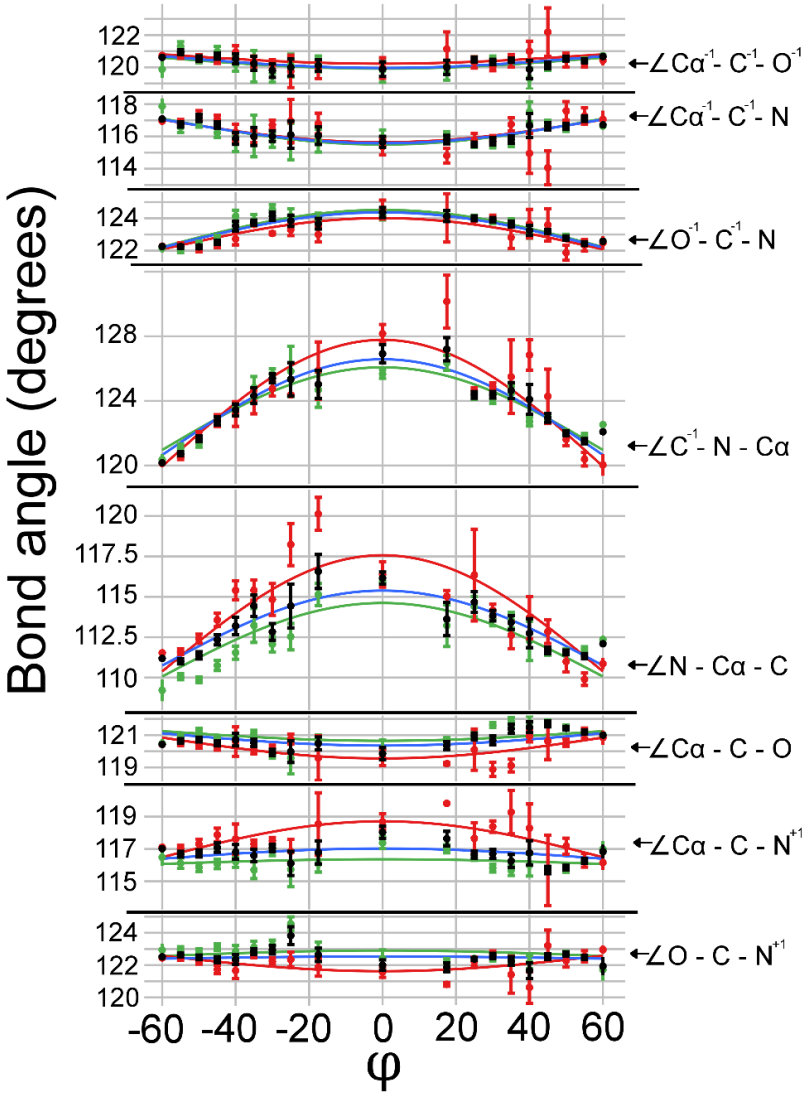

Figure 2.S4: How the average bond angle variations obtained by treating the $\psi \leq 0°$ and $\psi \geq 0°$ transitions separately compare with each other and with those based on the combined data.

Equivalent to Figure 2.3B, but with green lines and dots indicating the average values based on the $\psi \geq 0°$ observations and red lines and dots indicating the average values based on the $\psi \leq 0°$ observations. For reference, the blue lines and blacks dots show the data presented in Figure 2.3B based on combining all of the observations. Error bars depict standard errors of the mean. The smaller number of residues leads to larger uncertainties in the averages, and in our view, the most significant change is a slight expansion of ∠Cα-C-N^{+1} for the red compared with the green data. This is as would be expected as the expansion would help alleviate the Cβ…NH^{+1} close approach that would occur for non-Gly residues.

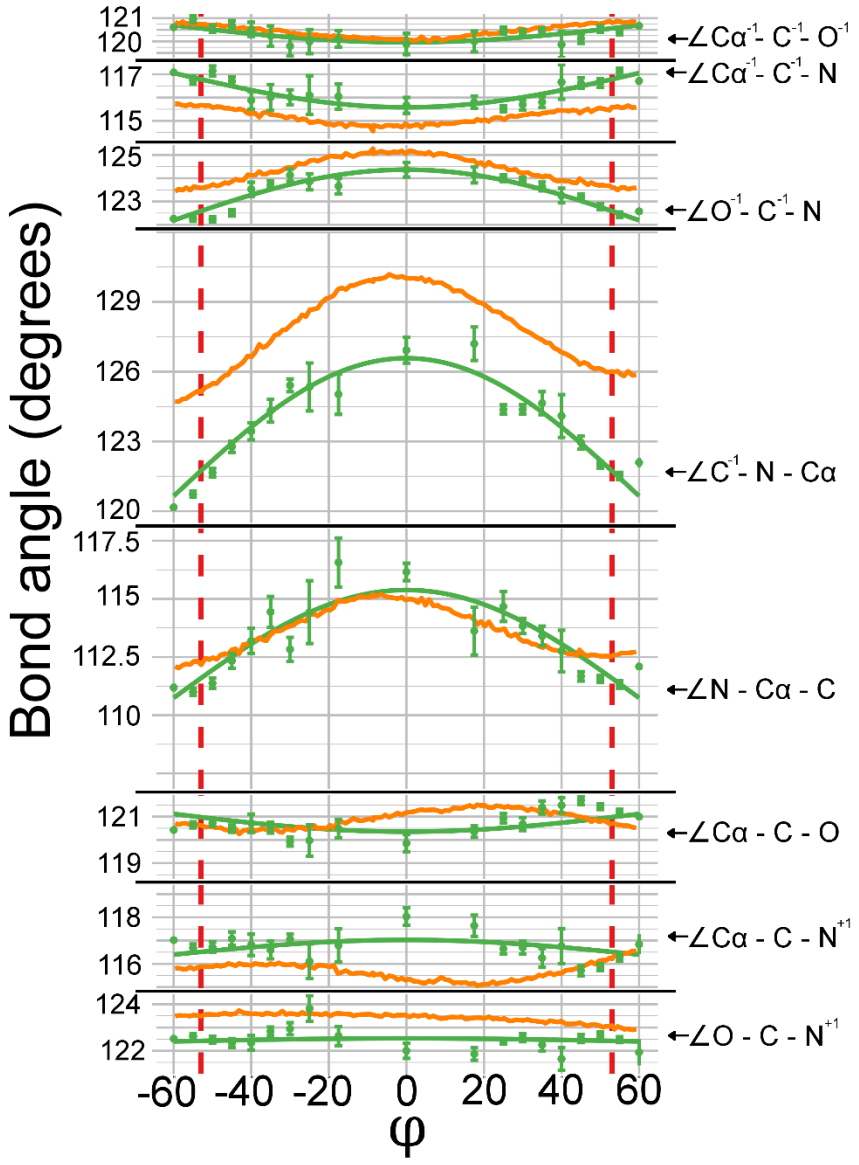

Figure 2.S5: AMBER minimizations of alanine dipeptides distort bond angles to alleviate the O^{-1} … C steric clash in $\varphi \sim 0$ conformations.

Bond angles are plotted against φ. Green lines represent calculated fits to observed data, using cosine equations to describe the systematic distortion of the bond angles; these equations can be found in Table S2. Orange lines show the calculated angles after minimization using AMBER12. Small black arrows indicate the Engh & Huber 2001 value for each angle[15]. Interestingly, the angles obtained after AMBER energy minimization in general follow the same trends observed in nature, and the greatest discrepancies between the angles from the AMBER minimizations and the empirically defined angles appear to be

largely due to standard relaxed values differing from the Engh and Huber standard values. This is seen in the parallel nature of the curves obtained for the $C\alpha^{-1}$-C-N, O^{-1}-C-N, and C^{-1}-N-Cα angles, with the last of these having the largest discrepancy, involving a roughly 4 degree displacement in the values.

Table 2.S1. Complete list of analyzed $\varphi \sim 0$ mountain pass residues.

Each residue in the range $-35° < \varphi < 35°$ which was manually checked and found to be reliably defined is included ordered by the resolution of the crystal structure. Columns provide the PDB codes for the structure containing the residue (Code), the resolution of the structure (Resol), the residue number (Res num), the chain (Chain), the residue type (Res Type), the solvent accessibility as reported by DSSP (SAS), the O^{-1}...C distance (d), the backbone torsion angles (and the backbone bond angles (φ and ψ), and the eight backbone bond angles analyzed: $\angle C\alpha^{-1}$-C^{-1}-O^{-1} (A1), $\angle C\alpha^{-1}$-C^{-1}-N (A2), $\angle O^{-1}$-C^{-1}-N (A3), $\angle C^{-1}$-N-Cα (A4), $\angle$N-Cα-C (A5), $\angle C\alpha$-C-O (A6), $\angle C\alpha$-C-N^{+1} (A7) and $\angle$O-C-N^{+1} (A8). An asterisk in the PDB Code indicates the residue is one of the observations from a protein having $\leq 25\%$ sequence identity with any other protein in the database, and was used to in calculating the variety of residue types observed that is reported in Table 2.S2.

Code	Resol (Å)	Res num	Chain	Res Type	SAS	d (Å)	φ	ψ	A1	A2	A3	A4	A5	A6	A7	A8
3O4P*	0.85	20	A	ALA	0	2.65	-34.75	125.4	119.70	116.95	123.25	125.59	111.09	121.40	114.63	123.77
4AYO*	0.85	249	A	ASP	25	2.72	-6.68	-78.5	119.28	116.64	124.06	126.84	117.86	120.44	117.39	122.08
2DDX*	0.86	157	A	SER	14	2.66	34.08	62.8	119.85	116.73	123.38	123.48	112.19	121.57	113.84	124.55
3F1L*	0.95	163	A	GLY	0	2.75	31.76	-120.3	119.17	116.18	124.65	124.65	116.51	121.43	116.04	122.43
1V0L	0.98	279	A	THR	65	2.74	34.06	61.9	119.73	117.34	122.91	125.79	112.82	121.17	117.67	121.15
3JU4*	0.98	704	A	VAL	0	2.77	33.84	66.3	121.25	113.65	125.03	126.37	111.97	124.70	111.67	123.60
3NOQ*	1	152	A	ILE	10	2.88	14.41	-83.5	120.08	114.36	125.54	131.80	115.39	119.18	119.84	120.92
4N1I*	1	261	A	HIS	5	2.77	-25.24	116.7	121.00	114.77	124.15	129.96	112.54	122.09	112.63	124.84
3QZR	1.04	111	A	SER	69	2.68	34.78	69.0	121.06	115.33	123.61	124.21	114.21	121.08	115.96	122.84
4FK9*	1.06	178	A	GLU	22	2.59	33.40	60.6	120.17	116.19	123.57	123.23	111.58	119.91	117.50	122.57
2X4K*	1.1	2	A	MET	14	2.80	28.35	62.7	119.63	117.97	122.39	124.71	119.08	120.84	117.74	121.40
4LF0*	1.1	166	A	CYS	2	2.85	34.12	-94.7	121.87	117.78	120.21	134.55	115.33	117.57	124.64	116.83
4IQB*	1.13	137	A	TYR	16	2.68	3.94	-87.2	121.63	115.25	123.11	129.59	114.49	118.78	119.69	121.36
4IQB*	1.13	137	B	TYR	16	2.68	14.77	-93.3	122.20	115.27	122.53	128.52	114.67	119.29	119.81	120.69
1QW9*	1.2	356	A	ILE	1	2.64	33.35	64.6	121.00	115.80	123.09	124.82	110.78	123.77	114.07	122.09

1QW9	1.2	356	B	ILE	1	2.65	35.71	62.9	120.49	115.33	124.01	124.56	110.23	122.74	115.37	121.88
2PGN*	1.2	77	A	CYS	0	2.77	-4.53	-89.0	118.71	115.52	125.68	127.93	117.43	118.71	119.14	122.13
2PGN	1.2	77	B	CYS	0	2.75	-7.99	-91.4	120.23	114.08	125.30	128.31	115.75	120.51	118.54	120.73
3MQD*	1.25	297	A	ALA	1	2.64	-9.16	106.2	121.84	114.75	123.41	126.30	115.64	121.44	116.38	121.98
4DD5*	1.25	355	A	ILE	0	2.72	-31.82	-70.0	119.61	117.09	123.28	125.58	114.26	119.68	117.88	122.31
4C6E*	1.26	1614	A	HIS	2	2.68	23.72	74.2	121.04	115.24	123.71	125.87	114.40	122.05	116.06	121.87
1GKP*	1.29	239	A	HIS	2	2.70	31.07	63.3	120.54	115.46	123.99	123.80	114.89	119.88	115.09	124.99
1GKP	1.29	239	B	HIS	2	2.71	36.81	57.3	121.20	115.24	123.52	122.00	116.46	119.61	116.41	123.96
1GKP	1.29	239	C	HIS	2	2.68	34.29	60.4	120.06	115.97	123.94	120.91	116.53	119.74	116.79	122.97
1GKP	1.29	239	D	HIS	2	2.73	32.81	56.1	120.33	113.37	126.07	119.80	119.38	122.13	114.09	123.76
1GKP	1.29	239	E	HIS	2	2.68	30.67	62.3	121.24	115.25	123.37	122.96	115.94	120.78	115.21	123.48
1GKP	1.29	239	F	HIS	2	2.70	31.56	61.0	121.55	112.30	126.12	120.63	116.62	122.93	117.10	119.85
2Z26*	1.29	177	A	HIS	1	2.67	31.15	66.6	121.22	114.30	124.40	124.75	113.37	121.44	116.01	122.54
2Z26	1.29	177	B	HIS	1	2.64	31.45	65.1	121.00	114.57	124.39	123.91	112.47	121.89	115.86	122.20
1R0M	1.3	221	A	GLN	0	2.79	34.04	64.1	120.40	114.53	125.06	126.25	113.84	120.88	117.34	121.76
1R0M	1.3	221	B	GLN	0	2.80	33.85	64.2	120.22	114.75	125.03	126.38	114.01	121.34	117.12	121.51
1R0M	1.3	221	C	GLN	0	2.79	36.10	62.3	121.44	114.16	124.39	125.84	114.84	120.47	117.72	121.80
1R0M	1.3	221	D	GLN	0	2.80	36.91	62.5	120.93	114.30	124.77	125.83	114.30	120.59	117.90	121.49
2GQT	1.3	115	A	ASN	0	2.65	31.48	58.8	119.72	116.51	123.76	123.52	111.61	121.62	115.64	122.73
3AJ7*	1.3	216	A	VAL	3	2.70	31.43	54.8	120.73	115.41	123.83	124.04	112.88	122.30	114.95	122.73
3BNJ	1.3	166	A	ILE	28	2.62	-27.64	117.8	120.56	115.27	124.17	125.36	110.73	121.23	115.51	123.25
3BNJ	1.3	277	A	HIS	59	2.56	21.81	71.0	120.04	115.81	124.08	123.73	113.66	120.59	117.05	122.35
3WDQ*	1.3	127	A	ASP	1	2.66	-16.99	103.0	121.33	113.99	124.63	124.75	114.28	120.74	116.40	122.85
1GWU*	1.31	288	A	THR	87	2.80	30.88	61.2	122.09	112.31	125.56	125.18	117.21	121.64	116.30	121.88
3ELF*	1.31	96	A	HIS	37	2.68	-18.85	110.1	120.48	116.47	123.05	125.15	116.92	120.76	116.72	122.45
1MKK*	1.32	26	A	CYS	9	2.66	-27.53	118.7	120.33	115.78	123.86	124.50	112.51	119.33	117.85	122.64
1MKK	1.32	26	B	CYS	9	2.71	-28.87	117.1	120.42	115.74	123.82	125.54	113.08	119.58	118.15	122.19
2EIY	1.35	213	A	SER	18	2.62	32.77	-118.0	119.96	116.61	123.42	123.09	111.25	119.43	117.84	122.73
2EIY	1.35	213	B	SER	18	2.61	33.96	-118.3	119.83	117.07	123.09	122.64	111.11	119.59	117.88	122.52
2EIY	1.35	213	C	SER	18	2.62	32.62	-117.4	119.80	116.95	123.26	123.01	111.66	119.31	117.98	122.71
2I49*	1.35	197	A	ASN	0	2.71	-20.08	-74.2	120.39	116.16	123.45	124.11	119.12	120.95	116.60	122.44
2VOV*	1.35	100	A	LYS	54	2.81	-34.81	-52.1	121.35	115.01	123.63	124.45	117.00	120.66	117.08	122.24
3IVY	1.35	311	A	THR	0	2.63	33.18	70.9	119.23	117.89	122.59	126.36	108.65	122.04	116.40	121.43
2VBA	1.36	299	A	GLY	0	2.76	15.31	87.3	120.32	116.16	123.50	127.52	118.05	120.04	116.71	123.09
2VBA	1.36	299	B	GLY	0	2.74	27.44	76.2	120.79	115.34	123.86	124.80	117.27	120.52	117.18	122.25
2VBA	1.36	299	C	GLY	0	2.77	32.79	75.1	120.44	116.22	123.34	124.95	117.63	120.67	117.14	122.17
2VBA	1.36	299	D	GLY	0	2.73	22.72	81.0	121.34	115.63	123.02	124.11	120.10	120.85	116.34	122.48
3B0X*	1.36	468	A	HIS	2	2.56	17.77	73.1	120.64	115.95	123.41	126.59	110.53	121.38	115.92	122.70
3VUR	1.36	12	A	ALA	7	2.65	31.95	-113.6	120.72	116.21	123.07	124.96	112.41	118.91	118.75	122.30

3C3Y	1.37	184	A	TRP	36	2.71	31.28	66.3	120.83	116.26	122.88	124.54	114.82	121.74	115.60	122.65
3C3Y	1.37	184	B	TRP	36	2.71	32.71	68.1	120.07	115.71	124.18	124.54	114.21	122.12	115.78	122.09
2Y24*	1.39	257	A	ASP	80	2.61	-4.33	95.5	118.41	116.68	124.78	125.00	115.82	118.51	117.32	123.97
1K3I	1.4	468	A	THR	35	2.76	34.95	47.6	120.71	115.56	123.73	124.90	113.32	120.93	117.27	121.76
1KQR*	1.4	114	A	SER	83	2.67	-31.97	122.1	120.64	116.46	122.76	123.47	114.17	120.27	117.32	122.39
1LLF*	1.4	318	A	ASP	33	2.78	-28.30	118.9	118.36	116.28	125.35	125.88	111.61	119.23	116.45	124.24
1LLF	1.4	1318	B	ASP	32	2.79	-27.73	119.2	119.78	115.20	125.02	126.71	111.41	119.49	119.10	121.38
1LLF	1.4	384	A	ASP	65	2.73	-28.52	107.0	115.29	118.37	125.97	126.32	111.87	118.13	115.82	125.83
1LLF	1.4	1384	B	ASP	63	2.74	-30.72	108.1	115.30	118.32	126.27	126.34	111.60	118.62	116.57	124.64
1O98	1.4	230	A	ASP	1	2.77	-25.30	-63.9	120.75	115.61	123.64	124.42	119.54	121.07	116.10	122.81
2C5A	1.4	203	A	ASN	39	2.81	32.63	80.2	120.19	115.94	123.81	129.07	113.59	122.66	114.37	122.91
2C5A	1.4	203	B	ASN	39	2.80	34.23	84.7	121.34	114.35	124.00	128.97	112.49	123.33	114.17	122.42
2EPL*	1.4	376	X	ASP	0	2.70	-21.91	128.9	119.33	115.89	124.62	122.96	116.65	120.47	114.78	124.63
2Q7W*	1.4	251	A	TYR	65	2.66	-31.94	-64.3	119.68	117.10	123.12	125.57	112.09	119.91	117.05	122.89
3FO3*	1.4	225	A	LEU	33	2.64	33.13	67.6	120.14	115.78	124.02	123.41	113.25	123.80	114.60	121.50
3FO3	1.4	225	B	LEU	33	2.61	31.40	69.0	119.87	115.92	124.10	124.20	111.54	123.14	115.27	121.33
3FO3	1.4	361	A	HIS	48	2.71	18.62	71.4	120.29	115.24	124.47	126.99	115.36	120.90	117.69	121.32
3FO3	1.4	361	B	HIS	48	2.69	19.21	71.7	120.57	115.92	123.47	126.16	115.98	120.66	117.23	121.91
3M6Z*	1.4	125	A	VAL	52	2.67	33.32	68.9	120.71	115.54	123.75	124.38	111.88	120.31	116.41	123.28
3MDU	1.4	136	A	GLU	57	2.64	-32.15	-60.2	120.13	116.88	122.97	123.42	113.71	120.32	117.45	122.22
4GNQ*	1.4	236	A	GLY	0	2.83	-21.18	-72.5	119.31	117.44	122.55	127.65	121.15	118.20	120.48	121.31
4IV0*	1.4	35	A	SER	1	2.78	-34.55	120.9	121.58	113.69	124.73	126.03	114.60	120.55	116.93	122.52
4IV0	1.4	35	B	SER	1	2.80	-31.97	119.9	121.60	113.70	124.70	126.27	115.36	120.92	116.60	122.49
4IZU*	1.4	145	A	CYS	8	2.85	29.70	-106.2	121.50	115.88	122.59	127.23	116.35	115.98	119.03	124.88
4MB1*	1.4	200	A	VAL	13	2.76	30.52	53.7	119.22	116.88	123.88	128.27	111.90	122.35	115.15	122.40
1X9D	1.41	463	A	ASP	21	2.70	-24.43	-67.1	118.75	118.30	122.93	124.31	116.93	120.43	117.60	121.88
1RK6*	1.43	250	A	HIS	5	2.66	33.58	62.5	121.60	114.91	123.48	124.58	112.49	121.80	115.63	122.48
3NHI	1.43	252	A	LYS	73	2.70	-33.88	-63.0	119.53	116.97	123.49	123.51	115.60	120.26	117.09	122.61
1KB0	1.44	128	A	LYS	76	2.73	34.56	60.9	120.42	114.35	125.15	122.53	115.83	122.22	114.86	122.89
4N4B*	1.44	285	A	HIS	16	2.71	-18.35	114.4	121.10	114.44	124.41	128.42	113.70	121.52	115.04	123.15
1PA2	1.45	286	A	SER	83	2.81	33.75	61.4	122.89	112.96	124.15	126.61	115.62	123.32	118.30	118.35
2CVD*	1.45	185	A	PRO	107	2.77	-33.53	-60.9	120.59	117.44	121.98	124.31	117.32	120.38	116.93	122.69
3DAQ	1.45	70	A	ASP	119	2.71	31.45	65.6	120.69	115.73	123.58	123.77	114.53	121.97	114.22	123.81
3DAQ	1.45	70	B	ASP	119	2.69	29.62	66.7	120.49	115.68	123.83	124.05	113.81	121.80	114.04	124.14
3DAQ	1.45	70	C	ASP	119	2.70	28.86	67.9	120.40	115.79	123.80	124.05	114.50	122.08	114.53	123.38
3DAQ	1.45	70	D	ASP	119	2.69	28.78	67.1	120.74	115.89	123.37	124.49	113.65	122.15	114.13	123.72
4F0B	1.45	84	A	THR	6	2.68	-33.92	-61.5	120.89	115.35	123.76	122.45	113.53	119.74	117.34	122.88
4F0B	1.45	84	B	THR	6	2.71	-34.71	-58.6	119.89	116.43	123.63	122.74	114.54	120.23	116.77	122.99
4HDR*	1.45	35	B	SER	47	2.71	28.63	-106.6	119.31	114.82	125.04	124.38	113.79	115.90	120.95	122.88

4HDR	1.45	35	D	SER	47	2.75	30.54	-105.8	118.78	116.29	124.44	126.33	113.47	117.35	119.85	122.67
4IRX	1.45	169	A	ASP	10	2.68	-2.09	88.9	119.92	115.79	124.28	125.53	116.02	120.17	117.98	121.83
4IRX	1.45	169	B	ASP	10	2.70	-1.76	90.00	119.09	116.62	124.27	125.87	116.28	120.27	117.83	121.90
4L6D*	1.45	144	A	SER	0	2.72	27.43	-124.5	120.65	115.36	123.98	124.82	113.50	118.83	118.63	122.50
4L6D	1.45	144	B	SER	0	2.72	29.71	-126.5	120.54	115.07	124.39	124.64	113.01	119.28	118.55	122.17
4L6D	1.45	144	C	SER	0	2.69	28.23	-126.0	120.50	115.16	124.33	124.68	112.54	118.84	118.40	122.72
4L6D	1.45	144	D	SER	0	2.74	32.63	-127.0	120.55	115.34	124.11	124.18	113.71	119.68	118.05	122.27
4L6D	1.45	144	E	SER	0	2.73	31.09	-126.6	120.45	115.61	123.94	123.84	114.10	119.85	117.94	122.18
4L6D	1.45	144	F	SER	0	2.71	27.86	-124.5	120.58	115.16	124.24	124.20	113.40	119.47	118.34	122.18
4L6D	1.45	144	G	SER	0	2.75	30.95	-126.1	120.51	115.43	124.05	124.31	113.99	119.56	118.29	122.13
4L6D	1.45	144	H	SER	0	2.73	30.39	-129.6	120.49	115.53	123.97	124.28	113.99	118.85	119.06	122.09
2UUQ	1.46	288	A	SER	40	2.61	31.33	63.9	119.20	116.61	124.19	122.93	111.33	120.56	117.28	122.16
4MHF*	1.46	239	A	ALA	0	2.65	-32.46	110.6	120.05	116.55	123.37	124.34	110.97	119.82	117.57	122.60
2Y8K	1.47	171	A	GLU	9	2.63	24.90	71.4	119.98	116.91	123.11	125.23	113.32	120.16	117.63	122.09
1G6U*	1.48	18	A	SER	51	2.60	14.93	76.3	117.98	117.05	124.72	127.09	109.50	120.24	117.87	121.47
1G6U	1.48	18	B	SER	51	2.62	18.66	69.8	117.85	116.48	125.55	126.42	109.45	120.91	116.63	122.33
4J7A*	1.49	230	A	CYS	1	2.68	26.33	59.2	119.49	115.78	124.71	124.47	114.01	121.05	116.74	122.21
4J7A	1.49	230	B	CYS	1	2.67	25.82	58.8	119.99	115.42	124.58	124.31	114.03	120.76	116.84	122.40
4J7A	1.49	230	C	CYS	1	2.68	26.30	59.1	119.76	115.24	125.00	124.62	114.08	120.76	116.93	122.31
4J7A	1.49	230	D	CYS	1	2.67	26.55	59.5	120.45	115.18	124.36	124.45	113.81	121.01	116.54	122.45
1BQC	1.5	128	A	GLU	21	2.57	28.26	67.5	121.55	116.54	121.90	124.98	111.84	120.71	116.70	122.58
1HT6	1.5	198	A	SER	88	2.78	32.39	59.2	119.10	116.08	124.81	123.79	117.60	122.29	116.79	120.91
1OFZ	1.5	258	A	THR	41	2.69	31.21	66.8	121.45	113.96	124.58	124.63	113.40	121.31	116.82	121.86
1OFZ	1.5	258	B	THR	41	2.72	35.35	61.8	120.05	115.52	124.42	124.81	113.26	122.24	115.70	122.05
2BKM	1.5	72	A	MET	112	2.57	31.95	54.6	118.96	117.64	123.39	122.57	109.23	122.21	114.96	122.83
2BKM	1.5	72	B	MET	112	2.73	33.12	56.1	119.77	116.45	123.74	123.99	111.46	122.71	113.84	123.40
2HP0*	1.5	422	A	PHE	208	2.74	-33.25	120.2	121.52	114.22	124.24	124.75	115.42	121.29	115.00	123.71
2JG0*	1.5	45	A	GLY	20	2.74	22.84	-118.6	119.95	116.18	123.81	124.44	119.19	121.38	116.69	121.92
2JG0	1.5	519	A	GLY	0	2.68	-31.55	-69.0	120.22	116.93	122.82	123.91	116.69	120.20	118.08	121.68
2O7I*	1.5	500	A	MET	2	2.71	34.30	62.8	119.15	117.77	123.06	123.44	112.28	120.95	116.78	122.27
2PKF*	1.5	171	A	SER	12	2.76	27.93	-114.5	118.62	117.58	123.78	125.08	115.01	119.97	116.91	123.11
2PKF	1.5	171	B	SER	12	2.76	28.15	-110.6	119.29	116.94	123.77	124.78	115.59	120.02	116.92	123.04
2WM5	1.5	313	A	SER	0	2.69	29.49	67.7	118.22	119.68	122.03	123.64	110.82	121.03	118.58	120.39
2ZBX	1.5	289	A	ALA	53	2.61	23.31	76.0	120.45	115.30	124.13	124.25	113.83	121.61	115.40	122.98
3BB7	1.5	126	A	THR	67	2.70	-25.80	114.9	120.48	115.44	123.98	125.39	113.33	119.80	115.95	124.25
3FWN	1.5	228	A	SER	36	2.63	-34.45	118.8	117.47	119.25	123.28	122.96	111.80	121.19	116.25	122.43
3GIP	1.5	248	A	HIS	3	2.57	24.08	63.1	120.95	114.99	124.05	123.52	112.74	121.41	115.77	122.82
3GIP	1.5	248	B	HIS	3	2.59	24.52	64.2	120.65	115.24	124.10	123.57	112.99	121.66	115.85	122.47
3KIZ*	1.5	262	A	SER	1	2.55	-24.86	110.8	118.03	118.87	123.10	122.40	114.00	119.12	117.33	123.55

3KIZ	1.5	262	B	SER	1	2.53	-20.67	110.2	118.85	117.92	122.94	122.17	114.14	120.75	117.52	121.62
3QOM	1.5	180	A	GLU	22	2.60	34.92	60.0	120.65	116.34	122.86	123.48	109.88	122.41	115.00	122.42
4AC7*	1.5	275	C	HIS	2	2.62	28.70	63.1	120.58	115.77	123.60	123.23	112.61	121.87	114.43	123.66
4AC7	1.5	52	B	THR	69	2.74	-29.00	126.4	121.40	114.57	124.02	126.11	113.40	120.13	115.96	123.90
4C72	1.5	312	A	GLY	0	2.67	33.76	72.4	120.15	117.49	121.97	122.22	115.84	121.20	115.74	122.94
4C72	1.5	312	B	GLY	0	2.68	32.77	74.2	120.17	115.43	123.49	121.80	116.77	120.95	115.09	123.73
4DQ6*	1.5	121	A	TYR	53	2.76	-28.95	119.5	120.88	113.73	125.36	127.91	108.67	119.84	117.05	123.06
4DQ6	1.5	121	B	TYR	53	2.68	-24.26	117.0	120.93	113.66	125.30	125.55	110.25	117.34	117.02	125.58
4ECQ*	1.5	16	A	CYS	16	2.54	25.39	63.2	121.02	115.38	123.59	122.78	112.13	120.89	116.52	122.59
4IJ5	1.5	169	A	PHE	0	2.83	-34.22	-53.0	120.18	116.15	123.62	126.43	116.41	119.40	118.26	122.33
4IJ5	1.5	169	B	PHE	0	2.82	-32.04	-53.2	121.29	115.57	123.14	125.32	117.43	120.82	117.08	122.09
4LR2*	1.5	339	A	ASP	35	2.73	-31.59	123.9	120.91	114.68	124.41	124.91	111.40	121.10	115.65	123.24
2BKL	1.5	557	A	VAL	21	2.67	28.84	60.7	121.80	113.92	124.27	125.30	112.43	120.70	117.35	121.94
2BKL	1.5	557	B	VAL	21	2.68	27.87	61.4	121.61	114.14	124.23	125.12	113.51	121.25	115.75	122.99

Table 2.S2. Frequency of amino acid types in the mountain pass transition region. The frequencies of amino acids in the range of $-25° < \varphi < 25°$ for a subset of residues from representative proteins with < 25% sequence identity.

Residues counted to generate these frequencies are indicated with an asterisk in Table 2.S1.

Residue type	Number of Occurrences
ARG	0
ALA	3
ASN	1
ASP	3
CYS	2
GLN	1
GLU	1
GLY	4
HIS	3
ILE	3
LEU	2
LYS	2
MET	2
SER	9
THR	2
TRP	1
TYR	1
VAL	2
PRO	0
PHE	0

Table 2.S3. Equations governing φ-dependent changes in geometry during transition through the mountain pass.

The ψ dependencies on φ are modeled as a line, and for each backbone bond angle, the dependence on φ is modeled as a cosine function. The equations are only valid for describing the changes taking place from $-60° < \varphi < 60°$.

Angle	**Equation [valid for $-60 < \varphi < 60$]**
ψ (for $\psi > 0$ pass)	$-0.91 * (\varphi) + 92.42$
ψ (for $\psi < 0$ pass)	$-0.93 * (\varphi) - 89.53$
Ca(i-1) - C(i-1) - O(i-1)	$-1.47 * \cos((\varphi) * \pi / 120) + 117.06$
Ca(i-1) - C(i-1) - N	$-0.73 * \cos((\varphi) * \pi / 120) + 120.70$
O(i-1) - C(i-1) - N	$2.18 * \cos((\varphi) * \pi / 120) + 122.19$
C(i-1) - N - Ca	$5.93 * \cos((\varphi) * \pi / 120) + 120.65$
N - Ca - C	$2.38 * \cos((\varphi) * \pi / 60) + 113.81$
Ca - C - O	$-0.76 * \cos((\varphi) * \pi / 120) + 121.12$
Ca - C - N(i+1)	$0.63 * \cos((\varphi) * \pi / 120) + 116.40$
O - C -N(i+1)	$0.13 * \cos((\varphi) * \pi / 120) + 122.40$

Table 2.S4. Further details of data plotted in Figure 3 including the ranges for and numbers of observations in each φ bin and the average distances and angles.

Columns report the range of each φ bin used in the analysis, the number of observations (N) contributing to the averages in each bin, the averages and standard errors [in parentheses for the most significant figures; e.g. 120.61±1.03 is written as 120.61(103)] of the $O^{-1}...C$ distance (d) and the backbone bond angles A1 – A8 as defined in Table 2.S1. Most φ bins are 5º wide, but they are wider in the less populated central region in order to have more observations in the bin.

φ bin	N	d (Å)	A1	A2	A3	A4	A5	A6	A7	A8
-60±2.5	100	2.89 (01)	120.61 (14)	117.08 (14)	122.26 (11)	120.17 (16)	111.19 (17)	120.43 (11)	117.02 (13)	122.52 (01)
-55±2.5	100	2.82 (00)	120.95 (15)	116.75 (17)	122.25 (10)	120.73 (13)	111.00 (16)	120.65 (13)	116.68 (18)	122.63 (11)
-50±2.5	97	2.78 (01)	120.55 (16)	117.14 (22)	122.23 (11)	121.66 (19)	111.37 (23)	120.72 (15)	116.73 (22)	122.48 (15)
-45±2.5	77	2.77 (01)	120.70 (19)	116.70 (19)	122.52 (14)	122.77 (24)	112.34 (32)	120.48 (15)	117.09 (29)	122.33 (19)
-40±2.5	22	2.74 (01)	120.52 (30)	115.88 (37)	123.53 (30)	123.44 (37)	113.20 (54)	120.72 (39)	116.82 (46)	122.35 (31)
-35±2.5	10	2.73 (02)	120.23 (44)	116.00 (56)	123.74 (16)	124.32 (05)	114.44 (66)	120.52 (23)	116.59 (38)	122.83 (19)
-30±2.5	18	2.72 (01)	119.80 (43)	116.01 (32)	124.13 (26)	125.41 (28)	112.83 (52)	119.92 (19)	117.06 (23)	122.95 (26)
-25±2.5	6	2.70 (03)	119.99 (52)	116.11 (83)	123.85 (35)	125.34 (103)	114.43 (135)	119.97 (67)	116.11 (75)	123.82 (55)
-17.5±5	7	2.69 (03)	120.11 (36)	116.05 (54)	123.66 (33)	125.03 (87)	116.57 (106)	120.49 (04)	116.79 (72)	122.64 (41)
0±12.5	8	2.69 (02)	119.89 (45)	115.67 (34)	124.36 (31)	126.92 (55)	116.16 (38)	119.85 (37)	118.03 (38)	122.00 (33)
17.5±5	9	2.67 (03)	120.00 (45)	115.80 (26)	124.14 (34)	127.20 (72)	113.62 (103)	120.35 (25)	117.64 (46)	121.87 (27)
25±2.5	14	2.66 (02)	120.47 (15)	115.51 (13)	124.00 (15)	124.37 (02)	114.67 (64)	120.92 (21)	116.65 (22)	122.38 (08)
30±2.5	38	2.70 (01)	120.34 (16)	115.70 (23)	123.91 (14)	124.38 (02)	113.84 (33)	120.68 (28)	116.70 (28)	122.56 (17)
35±2.5	31	2.71 (01)	120.44 (14)	115.81 (23)	123.65 (19)	124.64 (49)	113.41 (43)	121.40 (28)	116.25 (04)	122.26 (26)
40±2.5	6	2.73 (02)	119.87 (57)	116.67 (74)	123.26 (32)	124.09 (92)	112.75 (09)	121.49 (33)	116.76 (75)	121.66 (49)
45±2.5	55	2.76 (01)	120.12 (21)	116.63 (23)	123.19 (12)	122.95 (27)	111.67 (02)	121.69 (18)	115.70 (23)	122.52 (17)

50±2.5	100	2.79 (01)	120.54 (15)	116.62 (19)	122.78 (13)	122.00 (14)	111.55 (15)	121.42 (12)	115.85 (14)	122.69 (13)
55±2.5	99	2.85 (01)	120.40 (12)	117.10 (17)	122.43 (12)	121.51 (16)	111.32 (17)	121.19 (13)	116.27 (16)	122.48 (13)
60±2.5	96	2.94 (01)	120.67 (12)	116.71 (18)	122.58 (13)	122.09 (02)	112.09 (15)	121.00 (13)	116.84 (43)	121.93 (56)

Chapter 3

On the Reliability of Peptide Non-Planarity Seen in Ultra-High Resolution Crystal Structures.

Andrew E. Brereton and P. Andrew Karplus

Published in *Protein Science* (2016), 25, pp 926-932,

Abstract

Ultra-high resolution protein crystal structures have been considered as relatively reliable sources for defining details of protein geometry, such as the extent to which the peptide unit deviates from planarity. Chellapa and Rose[29] recently called this into question, reporting that for a dozen representative protein structures determined at ~1 Å resolution, the diffraction data could be equally well fit with models restrained to have highly planar peptides, i.e. having a standard deviation of the ω torsion angles of only ~1˚ instead of the typically observed value of ~6˚. Here, we document both conceptual and practical shortcomings of that study and show that the more tightly restrained models are demonstrably incorrect and do not fit the diffraction data equally well. We emphasize the importance of inspecting electron density maps when investigating the agreement between a model and its experimental data. Overall, this report reinforces that modern standard refinement protocols have been well-conceived and that ultra-high resolution protein crystal structures, when evaluated carefully and used with an awareness of their levels of coordinate uncertainty, are powerful sources of information for providing reliable information about the details of protein geometry.

Introduction

Since the mid-1990s, there has been a large increase in the number of protein structures solved at resolutions of near 1 Å and better, primarily due to the greater availability of intense synchrotron X-ray sources and techniques for rapid-cooling of protein crystals to cryogenic temperatures for data collection[116]. At such resolutions the diffraction data are sufficiently extensive that the geometric restraints that define the expected bond lengths, angles, and planarity become less important and less influential[117,118], and the coordinate uncertainties for well-ordered parts of the protein drop down into the range of 0.01 – 0.05 Å[118–120]. For these reasons, such ultra-high resolution protein crystal structures have been considered reliable sources for gaining insight into features of protein geometry that differ from the standard expected molecular geometries. Examples include the geometric distortions of ligands and

protein groups in enzyme active sites[121–125]; the systematic variation in peptide backbone bond angles as a function of both: the backbone conformation in commonly observed regions of ϕ,ψ-space[23], and in rarely observed high-energy transition conformations[126]; and a level of non-planarity of the peptide unit in proteins much greater than expected based on data from lower resolution protein crystal structures that were strongly influenced by peptide planarity restraints[30].

In the case of peptide non-planarity, our group[30] built on the work of others[108,120,127] to show that in protein structures determined at 1 Å resolution and better, the ω torsion angles (defined by the $C\alpha_{i-1}$-C_{i-1}-N_i-$C\alpha_i$ atoms and equal to 180˚ for a perfectly planar *trans* peptide unit) are rather broadly distributed. The standard deviation was ~ 6.3° for *trans* peptide bonds with about 12% and 0.5% of residues deviating more than 10˚and 20˚, respectively, from planarity. As not all parts of a crystal structure have the same level of reliability, we examined the electron density of every peptide with $\omega \geq 20^{\circ}$ from planar to assess which were reliable; then using the reliable examples, we showed that these highly non-planar peptides are not just in active sites, but represent a mundane aspect of protein structure that simply reflects the 'frustration'[128–130] that occurs as proteins fold into their tertiary structures. We (see Figure 3D of Berkholz et al[30]) and others[127,131] made very clear that such levels of deviation from planarity are fully consistent with the estimate originally defined by Pauling that deviations of ~10˚ from planarity would come at a cost of about 1 kcal/mol[12].

Recently, Chellapa and Rose (CR) challenged these conclusions, calling into question whether such deviations from planarity are "a necessary implication from the available data"[29] and also incorrectly claiming that the reports of such deviations from planarity are "raising doubts about Pauling's consequential inference that distortions from planarity come at significant energetic cost." To support their claims of such deviations not being reliably determined, they reported that re-refinements of twelve ultra-high resolution protein structures using tighter restraints on ω yielded alternative protein models in which the ω angles were much closer to 180˚, "without consequent reduction in reported evaluation metrics (e.g., R-values)"[29]. They also claimed that even for ultra-high resolution structures different refinement packages

led to "distinctly different ω-angle signatures", indicating that one cannot rely on ultra-high resolution crystal structures to obtain unbiased and accurate geometric details.

Here, we report that there are multiple shortcomings of the study by CR and that their conclusions are not valid. It is certainly true that refinements as done by CR using much tighter ω restraints do yield structures that have lower deviations from planarity, but it is far from true that these alternative models provide equally good fits to the diffraction data. Two key shortcomings in the CR study were a failure to carry out proper control refinements, and, even more importantly, a primary reliance on global statistics to evaluate model quality rather than the inspection of difference electron density maps and 2Fo-Fc maps to assess details of how well each model agrees with its data.

Refinement Protocols

In order to reinvestigate the effects of tightened ω restraints on model quality, we used the same 12 structures studied by CR (Table 3.1), and re-refined them just as they described[29], with a specific seed used for all refinements to ensure consistency and repeatability. Briefly, starting from each PDB entry, an initial round of five cycles of refinement was done, using 'phenix.refine' with the options 'strategy=individual_sites', 'wc=0', and 'main.random_seed=2772306'. The resulting model was then used as the input for two subsequent parallel refinements: one used the same tight ω restraint that was applied by CR; the other was done to serve as a control refinement and used a standard ω restraint. These refinements were each run for five cycles using the following options: 'strategy=individual_sites+individual_sites_real_space+individual_adp+occupancies', 'wxc_scale=0.4', 'optimize_xyz_weight=True', 'optimize_adp_weight=True', 'optimize_mask=True', 'wc=1', and 'main.random_seed=2772306'. The "tight" refinements differed from the standard refinements only by setting 'omega_esd_override_value = 0.5'. In the following sections, these two refinements will be called "standard" and "tight".

Table 3.1 The application of tight ω-restraints significantly increases overall R-values over the 12 test cases.

PDB ID	Software	Res[a]	Deposited[b]		R_{work}			R_{free}		
			R_{work}	R_{free}	Std.[c]	Tight[c]	Δ[c]	Std.[c]	Tight[c]	Δ[c]
2CWS	SHELXL	1.00	10.8	13.8	10.9	11.7	0.8	12.8	13.8	1.0
2GUD	REFMAC	0.94	14.5	15.5	13.6	14.1	0.5	15.2	16.1	0.9
2OV0	SHELXL	0.75	12.8	13.9	13.1	13.7	0.6	14.1	14.7	0.6
2P5K	REFMAC	1.00	13.5	15.9	13.0	13.2	0.2	15.4	15.9	0.5
2PNE	REFMAC	0.98	14.1	17.0	13.5	14.5	1.0	16.4	17.5	1.1
2QSK	REFMAC	1.00	13.9	16.1	13.7	14.2	0.5	15.9	16.5	0.6
3D1P	REFMAC	0.98	12.3	13.4	12.2	12.7	0.5	13.5	14.1	0.6
3F7L	SHELXL	0.99	11.9	14.2	11.7	11.8	0.1	14.0	14.3	0.3
3QL9	PHENIX	0.93	12.8	13.6	11.9	11.9	0.0	12.8	13.1	0.3
4AQO	REFMAC	0.99	12.8	16.4	13.0	13.7	0.7	16.6	16.9	0.3
4JP6	REFMAC	1.00	15.8	18.4	16.5	16.6	0.1	18.9	19.1	0.2
4MTU	SHELXL	0.97	14.1	15.7	14.3	14.8	0.5	15.4	15.8	0.4
	Average	0.96					0.5			0.6
	σ	0.07					0.3			0.3
	p-value[d]					3.0E-04			3.8E-05	

[a] Resolution

[b] R_{work} and R_{free} values recalculated by *Phenix* for each model as deposited in the pdb (as was done by CR[29]).

[c] Reports R-values for models from refinement using standard ω-restraints (Std) or tight ω–restraints (Tight), and the "Tight – Std" difference (Δ).

[d] A paired, two-tailed t-test was used to obtain p-values.

R-values are consistently worse with the tight ω restraints

The R-values of our parallel re-refinements of each ultra-high resolution structure are reported in Table 3.1, with the R-values calculated by Phenix of the deposited models also provided for comparison. The models produced by the tight ω restraints had R_{work} values ranging from 0.0 to 1.0% higher and R_{free} values ranging from 0.2 to 1.1% higher (Table 3.1). The statistical significance of this consistent decrease in the overall agreement of the model to the data seen over the dozen structures was assessed using a paired two-tailed t-test. The p-values of 0.00003 and 0.0000038 for the changes in R_{work} and R_{free} respectively (Table 3.1), show that even just based on these overall R-values, the standard ω restraint produces models that are significantly better than the tight ω restraint.

CR did not see such a consistent difference in the overall R-values because they compared the R-values of the structures refined using tight ω restraints to the R-values calculated for the structures as they were deposited. The shortcoming of that comparison is that it is not properly controlled for all changes but the tightness of the ω restraint. For instance, for PDB entry 3QL9, the re-refinement using Phenix with tight ω restraints lowered R_{work} and R_{free} by 0.9 and 0.5 % respectively compared with the R-values of the deposited coordinates (Table 3.1). While one might be tempted to conclude that tightening the ω restraints led to a better model, it is crucial to note that one has also changed the refinement program from an older version of Phenix (the coordinates were deposited in 2011) to a newer version. Indeed, as shown by our control refinement using the current version of Phenix with the standard ω restraint, R_{work} decreased by the same amount, and R_{free} decreased by 0.3% more (Table 3.1). Thus, the decrease in R-values seen was due to something about the newer version of Phenix, rather than due to tightening the ω restraint.

Also, while changes in R-values on the order of 0.5 - 1% may seem small, such changes are actually quite notable. In this regard, we note that crystallographers often expend much effort toward the end of a refinement trying to get structural details just right, and at this stage a drop in R_{free} in the 0.5-1.0% range is seen as a strong validation that the changes made were worthwhile. Also, when the R-values

are as low as the ones for the structures analyzed here, a drop from, say, 14% to 13% is a substantial fractional improvement, especially given that overall R-values are global indicators that are not very sensitive to small changes in the positions of a small subset of atoms in a large protein structure[117].

Electron density maps show that models from tight ω restraints are not correct

As a global statistic, R-values are fairly insensitive to the incorrect positioning of a few atoms, and are unsuitable for using to evaluate whether any particular part of a model is correct. For this reason, the decisions about how to improve a structure during crystallographic refinement are made based on inspection of electron density maps, and as noted in a recent review about how to avoid pitfalls during structure determination: "a model must always be thoroughly scrutinized visually against electron density maps before accepting it as final."[132] At the ultra-high resolutions of the structures studied here, well-ordered parts of the protein have atoms fully resolved, making interpretation relatively easy.

For all 12 test cases, inspection of the electron density map reveals that in the tightly restrained structures some peptide units are very clearly incorrectly fit.† As a dramatic example we consider the peptide between residues 189 and 190 in PDB entry 2CWS, for which the standard refinement led to a peptide with ω ~33˚ from planarity. For this peptide, the tight restraints produced a structure that placed atom 190-N well outside of the strong 2Fo-Fc density for that atom (Figure 3.1A), and for which the difference map has a very large pair of positive and negative peaks further making clear that the 190-N atom needs to be shifted in order to agree with the data (Figure 3.1B). In difference electron density maps for mostly well-fit models, the root-mean-square electron density of the map (ρ_{rms}; also commonly called σ) is taken as an upper limit of its noise level (since many peaks are due to signal). In general, peaks that are smaller than $\pm 3 * \rho_{rms}$ are considered not reliably distinguishable from

†Each of the structures has additional strong difference map peaks that are associated with other shortcomings of the models (a commonplace occurrence in deposited models[133]), but these peaks show up in both the standard and tight refinements and so are not relevant to this study.

noise and the larger the peak the more significantly it indicates something that is wrong with the model[117]. These ~13*ρ_{rms} pair of peaks (Figure 3.1B) are a glaring indicator of an atom in the wrong position. In contrast, the Fo-Fc difference map calculated using the model refined by standard restraints does not show any peaks at this location indicating a need for atoms to shift (data not shown). Among the other 11 test-case structures refined with 'tight' restraints, for two the largest difference peak associated with a peptide plane also has a peak height of greater than 10*ρ_{rms}. For the other 9 structures, the strongest peptide-plane-associated difference peaks are between 5 and 10 ρ_{rms}, heights that are smaller but still very significant.

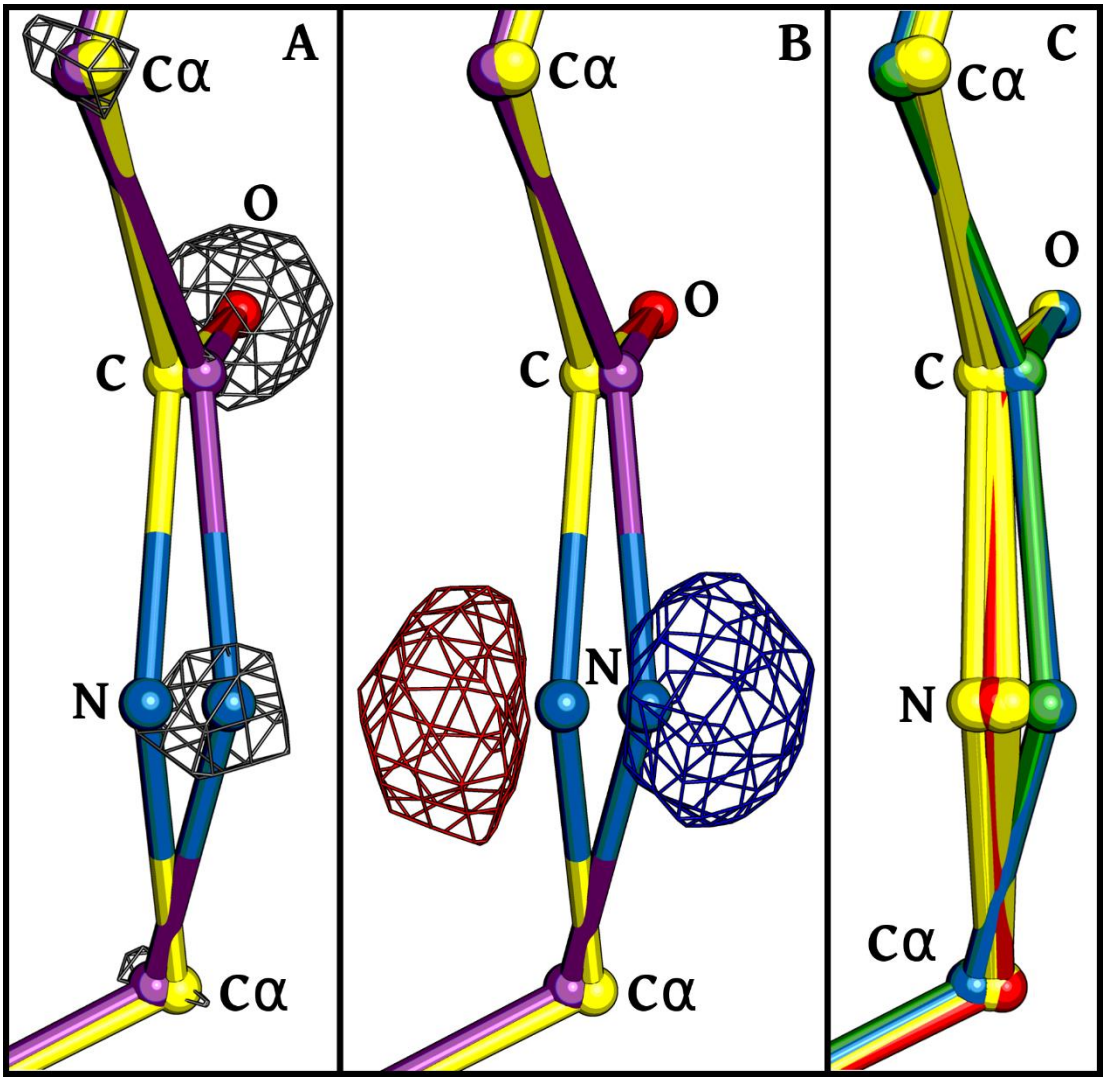

Figure 3.1 Evidence from electron density that tight ω-restraints lead to incorrect models.

Backbone atoms (labeled C, N, O and Cα) of the peptide unit between residues 189 and 190 of PDB-ID 2CWS are shown for multiple models, using atom coloring (oxygen - red; nitrogen – blue; carbon – as defined for each model). **(A)** The model resulting from refinement using tight ω-restraints (yellow carbons; ω=175.6°) is shown along with its 2Fo-Fc electron density (grey mesh; contoured at 7.5*ρ_{rms}), revealing that the N-atom is clearly placed incorrectly. Additionally shown is the model resulting from refinement using standard ω-restraints (purple carbons; ω=147.2°) which does fit the electron density well. **(B)** The same structures are shown with the Fo-Fc difference electron density also calculated using the tightly-restrained model. Negative (red mesh) and positive (blue mesh) density are contoured at ±7.5*ρ_{rms}. **(C)** Shown are the models from 10 pairs of independent refinements done either using tight ω-restraints (yellow) or standard ω-restraints (blue). The deposited model (green) and the model refined by CR[29] (red) are also shown.

To test whether a single pair of refinements carried out for each structure can be taken as a reliable representative of the structures that could result from each refinement protocol, we carried out a set of independent refinements from different starting models and see how much spread would occur in the atom positions[23,134,135]. For PDB entry 2CWS, we created ten different starting models using the 'shake' algorithm of Phenix with the setting 'modify.sites.shake=0.2'; this randomly moves the model atoms so that the mean shift of all atoms is 0.2 Å. Then each of these 10 models was re-refined using the "tight" and "standard" protocols and the same experimental data. The results were that the rms spread of backbone atoms within the standard and tight sets of 10 structures was just 0.011 Å and 0.017 Å, respectively. Visually, for the 189-190 peptide in structure 2CWS, the 10 structures created using each refinement protocol are very tightly clustered so that the atomic shift that occurred between protocols for the 190-N atom is quite reliably defined (Figure 3.1C). Furthermore, the 2CWS coordinate set generated by CR using tight restraints is situated in the midst of the 10 tightly restrained structures we generated, and the original PDB entry is situated in the midst of the 10 "standard" structures (Figure

3.1C). We conclude from these observations, that at these resolutions, a single refinement provides representative coordinates with a precision of <0.02 Å for well-ordered parts of the structure. Interestingly, this "shake" experiment also validates that using tight ω restraints does somewhat degrade the overall R-values: for the 10 tight refinements, the mean R_{work} / R_{free} were 11.5 ± 0.5 (SD) / 13.8 ± 0.5, and for the 10 standard refinements, the mean R_{work} / R_{free} were 11.2 ± 0.3 / 13.2 ± 0.4. Comparing R-values from the 10 tight refinements with those from the 10 standard refinements using a paired t-test yields p-values of 0.1 and 0.02 for R_{work} and R_{free}, respectively. Given that all of the tightly restrained models are incorrect for the peptide before residue 190 (Fig. 3.1C), the lack of significance for the change in R_{work} using a $P<0.05$ criterion does not indicate the models are equally valid, but underscores the unsuitability of this global statistic for assessing details of model quality.

Tightly restraining ω causes shifts in many atoms in excess of their positional uncertainty

For a more global analysis of how the uncertainty in atomic positions compares with the atomic shifts caused by the tight ω restraints, we calculated for each peptide a 'peptide shift ratio' by dividing the shift in positions that occurred between the tight and standard refinements by an estimate of the positional uncertainty of the atoms (as defined in the Figure 3.2 legend). This ratio increases fairly linearly as a function of how much ω deviated from 180° in the standard refinement (Figure 3.2). When |ω-180°| is ~5°, nearly all peptides exhibit a shift that is greater than the estimated coordinate uncertainty, and for |ω-180°| ~10°, nearly all peptides have been shifted more than triple their experimental uncertainty. This shows that across the dozen test structures the non-planarity of peptides having ω over ~5° away from 180° can be considered to be defined reliably enough that the structure restrained to be planar is not an equally valid alternate interpretation, but in fact is a model that is not consistent with the diffraction data.

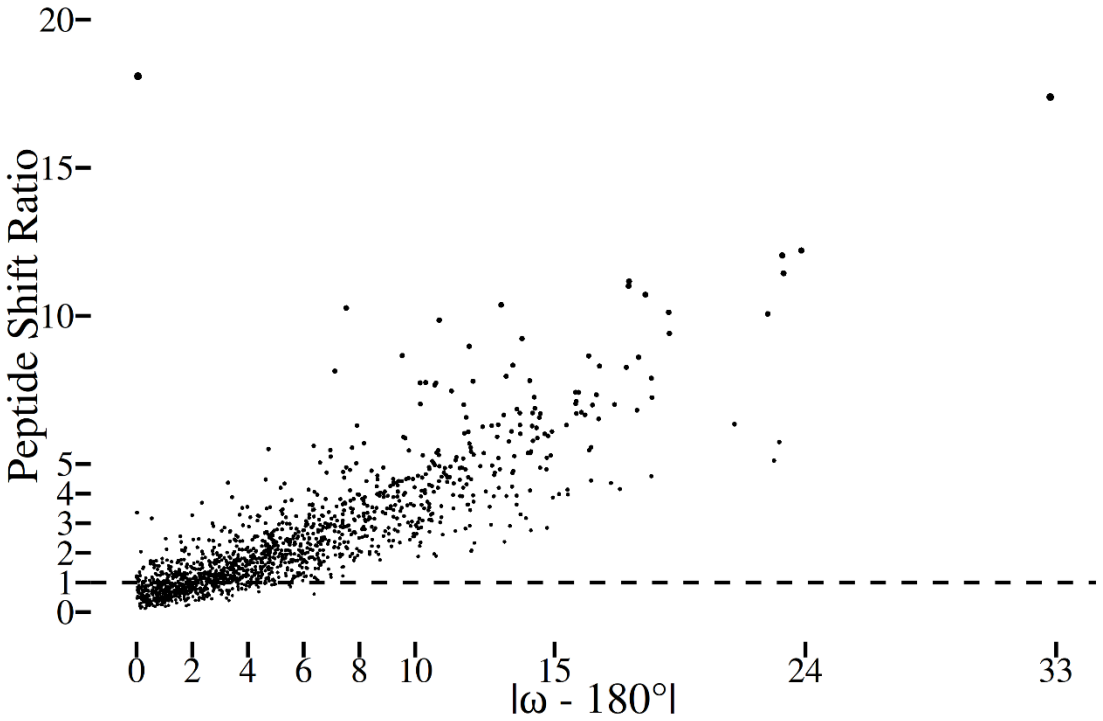

Figure 3.2 Significant atomic shifts are caused by tight ω restraints.

Plotted for each peptide in the dozen test structures is the peptide shift ratio (defined below) as a function of the degree to which ω deviates from 180° in the standard refinement. For each atom in a structure refined using standard restraints, the standard uncertainty in the position of each atom[119] was estimated using the Online_DPI webserver[136]. Also, the shift for each atom between the standard vs. tight **ω** refined structures was calculated. Noting that the tight restraints often lead mostly to shifts in the central C, O and N atoms of a peptide (e.g. Figure 3.1), we defined a 'peptide shift ratio' for each peptide as the rms of the shifts of the three central atoms in the peptide (i.e. O_{i-1}, C_{i-1}, and N_i) divided by the rms of the estimated standard uncertainties of the same three atoms. A value of 1 means that the rms shift in the atom positions is equal to the uncertainty in the positions of those atoms. The most non-planar residue in the dataset is the 2CWS 189-190 peptide shown in Figure 3.1. The one outlier in the plot is the 179-180 peptide from PDB entry 3QL9 with an ω angle in the standard refinement that is only 0.1˚ from planar but for which the backbone oxygen shifts ~0.4 Å to yield a peptide shift ratio of ~18. This can be rationalized in that this peptide oxygen has high anisotropy and that the method used to estimate the positional uncertainty does not take the anisotropy into account[136].

At these resolutions, ω angle distributions do not depend on refinement software

An additional argument provided by CR[29], that the tight restraints led to valid alternative models, was that Phenix, Refmac, and SHELX led to "distinctly different" distributions of ω angles, with the same mean but "significantly different tails" (see Figure 5 of that paper). Using the same set of structures that they used, we repeated the analysis and found that in fact the ω-distributions are remarkably similar, closely matching with regard to the median, the standard deviation, the 25th and 75th percentiles and even the 0.1st and 99.9th percentiles (Figure 3.3A). With regard to assessing the upper and lower limits of the extreme outliers, as was pointed out by Berkholz *et al.*[30], it is crucial to inspect the electron density maps to remove examples that are not reliable. Doing this leaves no outliers over 40˚ from planar in any of the sets (Figure 3.3A), and leaves the Refmac and SHELX distributions having slightly more extreme outliers, which is reasonable as three to four times as many residues are included in those distributions. Also, that the detailed values of the extreme outliers differs for each distribution makes sense, because the sets contain different proteins. The remarkable similarity of these distributions is consistent with the expectation that at resolutions near 1 Å, the diffraction data are sufficiently extensive that differences in restraints used by different refinement packages should have little influence on the resulting structure[23,117]. Further support for this conclusion is that when a comparison is made of the ω distributions for the same set of structures refined by different programs, even the outliers match within a few degrees (compare plots D and S in the Figure 3.3B left hand panel).

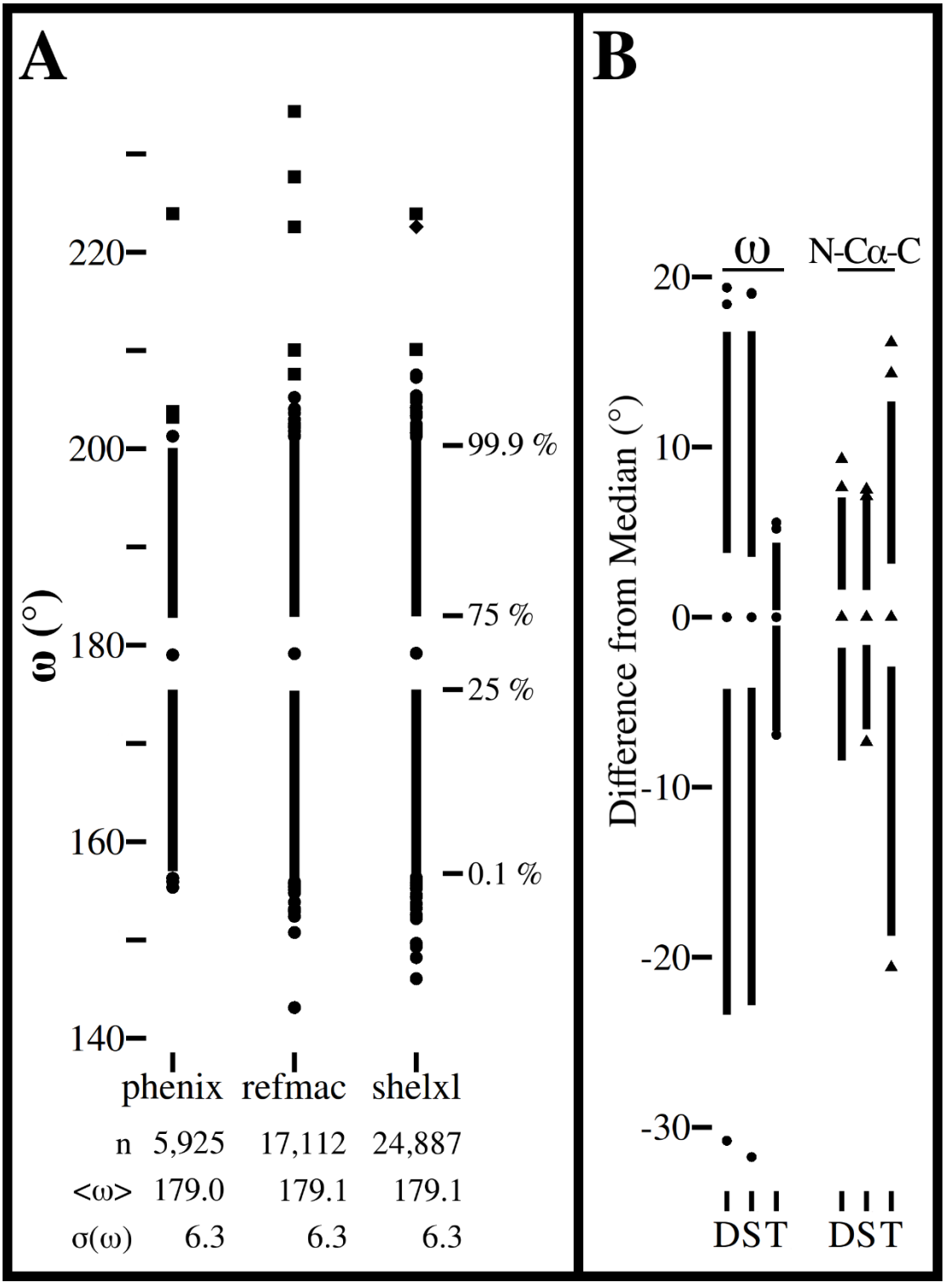
A
220
200
180
160
140
ω (°)
99.9 %
75 %
25 %
0.1 %
phenix
refmac
shelxl
n 5,925 17,112 24,887
<ω> 179.0 179.1 179.1
σ(ω) 6.3 6.3 6.3
B
ω
N-Cα-C
20
10
0
-10
-20
-30
Difference from Median (°)
DST
DST

Figure 3.3 ω-angle distributions from three refinement programs and distortion of the N-Cα-C angle caused by the tight ω restraints.

(A) ω-angle distributions for 40, 92 and 137 structures identified by CR[29] as having been refined at ≤ 1 Å resolution by Phenix, Refmac and SHELX, respectively. For each refinement package (identified by name), the number of refined residues, the mean and standard deviation of ω, and a Tufte boxplot[137] are shown. In each boxplot, the central dot marks the median, the upper line extends from the 75th percentile to the 99.9th percentile, the lower line extends from the 25th to the 0.1st percentile and the most extreme 0.1% of observations at each end are shown as individual circles, squares or diamonds. For each distribution, observations were manually checked for the quality of the fit to their electron density starting with the furthest outlier and continuing until a reliably modeled example was found. Observations were categorized as incorrect (squares), unable to be assessed due to unavailable data (diamonds), or reliable (circles). (B). Tufte boxplots[137] showing the distributions of the ω (circles) and N-Cα-C (triangles) angles relative to their median values for residues in the dozen test structures. The left plots show ω for the structures as deposited (D), and re-refined using standard (S) or tight (T) ω-restraints, and the right plots similarly show the N-Cα-C angles. Standard deviations for the ω distributions are: 6.5°, 6.5°, and 1.0°. For the N-Cα-C angle distributions the standard deviations are 2.5°, 2.3°, and 4.6°.

Tight ω restraints cause unreasonable secondary distortions

An interesting observation pointed out by CR but not further discussed was that the application of tight ω restraints during refinement also results in a wider spread of N-Cα-C angles. We have confirmed this observation (Figure 3.3B), and find it easy to explain. As we have noted (see Figure 3.2 legend), forcing non-planar peptides to become planar tends to involve shifts of the peptide C, O and N atoms, and any movement of the backbone C and/or N atoms will change the N-Cα-C angle. As an example, for the 2CWS 189-190 peptide (Figure 3.1) ω changes from 147.2° to 175.6° between the standard and tight refinements, with the N-Cα-C angle of residue 189 changing minimally from 111.6° to 112.8°, but that of residue 190 changing from 105.7° to 92.1°. As we have made clear above, this "alternative" model does not

agree with the diffraction data (e.g. Figure 3.1), and such levels of distortion of an N-Cα-C angle are unprecedented even in ultra-high resolution structures.

Synthesis

We hope that the above analyses lay to rest any concerns that may have been raised by CR[29], and reinforce what has been known for decades in the protein crystallographic community: that ultra-high resolution structure determinations of proteins have unique value for defining accurate protein structures[118,120]. We also hope that we have made clear the limited value of global metrics, and the importance of inspecting electron density maps to assess the reliability of any given feature in a crystal structure. In this example, by showing that peptide non-planarity in ultra-high resolution protein crystal structures is unequivocally supported by experimental evidence, and that alternative models in which peptide units are forced to be highly planar are clearly not compatible with the data. Using highly increased ω restraints during refinements leads to notably worse R-values, shifts in the positions of atoms that are well beyond the experimental uncertainty, a worsening of other aspects of protein geometry, and most importantly, clear regions of disagreement between the model and the electron density maps.

One additional point we would like to clarify relates to the broad claim of CR[29] that “a refined protein crystal structure is essentially an under-determined model.” They supported this point by quoting the abstract of a paper by protein crystallographer Tom Blundell and coworkers[138] who wrote that "disregarding structural heterogeneity introduces degeneracy into the structure determination process, as many single, isotropic models exist that explain the diffraction data equally well. The large differences among these models imply that the accuracy of crystallographic structures has been widely overestimated. Further, it suggests that analyses that depend on small differences in the relative positions of atoms may be flawed." What apparently was not recognized by CR is that in this statement the authors were only referring to low and medium resolution structures, where

"heterogeneity is difficult to identify and model, and are therefore approximated by a single, average conformation with isotropic variance."[138]

In contrast, for structures determined at near 1 Å resolution heterogeneity is relatively easy to identify and model, and anisotropic motions of atoms are also accounted for. Indeed, Blundell and coworkers explicitly point this out in their introduction: "Modeling anisotropic motion and structural heterogeneity has been limited to proteins that diffract to atomic resolution, due to the necessity for a high parameter-to-observation ratio[139,140]." So the comment that "a refined protein crystal structure is essentially an under-determined model", is only accurate for structures solved at medium to low resolution, and explains why the more planar ω angles seen in those structures cannot be taken as reliable[30,127]. CR's comment does not apply to the structures solved at resolutions near 1 Å and better, as these are sufficiently over-determined that they are able to provide insights into the true details of protein geometry.

Finally, we want to repeat in closing, that the levels of deviation from planarity seen in the ultra-high resolution protein structures and in small molecule peptides are not at all in conflict with Pauling's ideas about the planarity of the peptide bond, but are strikingly consistent with them[131]. We fully recognize that for protein structure prediction, it has been and still can be very helpful to assume that peptide planarity is absolute, but this must be recognized as a simplification that in itself is not consistent with Pauling's thinking about the peptide unit. That accounting accurately for such geometric details is of practical value was recently provided by a study showing that for protein prediction with Rosetta can be enhanced by allowing ω torsion angles to deviate from planarity at the levels seen in ultra-high resolution structures[85].

Acknowledgements

We thank George Chellapa and George Rose for answers to questions about their study and also for providing the coordinate sets they generated by their

refinements, the exact parameters of their refinements, and the list of PDB codes they used for their Figure 5. This work was supported in part by NIH grant R01-GM083136 to PAK. The authors do not report any conflict of interests.

Chapter 4

Ensemblator v3: Robust Atom-level Comparative Analyses and Classification of Protein Structure Ensembles

Andrew E. Brereton and P. Andrew Karplus

In the Press in *Protein Science*, July 2017

Abstract

Ensembles of protein structures are increasingly used to represent the conformational variation of a protein as determined by experiment and/or by molecular simulations, as well as uncertainties that may be associated with structure determinations or predictions. Making the best use of such information requires the ability to quantitatively compare entire ensembles. For this reason, we recently introduced the Ensemblator[141], a novel approach to compare user-defined groups of models, in residue level detail. Here we describe Ensemblator v3, an open-source program that employs the same basic ensemble comparison strategy but includes major advances that make it more robust, powerful, and user-friendly. Ensemblator v3 carries out multiple sequence alignments to facilitate the generation of ensembles from non-identical input structures, automatically optimizes the key global overlay parameter, optionally performs "ensemble clustering" to classify the models into subgroups, and calculates a novel "discrimination index" that quantifies similarities and differences, at residue or atom level, between each pair of subgroups. The clustering and automatic options mean that no pre-knowledge about an ensemble is required for its analysis. After describing the novel features of Ensemblator v3, we demonstrate its utility using three case studies that illustrate the ease with which complex analyses are accomplished, and the kinds of insights derived from clustering into subgroups and from the detailed information that locates significant differences. The Ensemblator v3 enhances the structural biology toolbox by greatly expanding the kinds of problems to which this ensemble comparison strategy can be applied.

Introduction

Proteins are dynamic biological macromolecules that sample many different conformations depending on their intrinsic structural properties and their environment. Even for natively folded proteins, the *true* native state cannot be perfectly represented by a single model, meaning that an ensemble of structures is needed to capture the breadth of the native state[142–144]. Also, ensembles can be used to capture the uncertainty associated with a structure determination or prediction

approach. For both reasons, the use of ensembles to describe protein structure is a critical component of especially NMR, but also cryo-EM, X-ray crystallography, molecular dynamics simulations, and structure predictions. In addition, even though protein crystal structures are still typically modeled as a single conformation, the gathering of structures from multiple independent structure determinations into an "X-ray ensemble" provides a more complete view of the range of conformations associated with the native state, as has been underscored by the creators of the Conformational Diversity of the Native State (CoDNaS) database[71].

To make the best use of such information, it is critical to be able to quantitatively compare and analyze ensembles of protein structures without losing the ensemble information; yet few methods exist for direct quantitative comparisons of ensembles. For instance, authors of one recent report concluded that to account for the role of conformational diversity in assessing protein predictions "would necessarily require new improvements and novel methodologies of model evaluation."[145] To address this need, in 2015, we introduced the Ensemblator 1.0[141] as a conceptually simple tool for global and local comparisons of ensembles of structures that reveals residue- (and even atom-) level details about systematic differences between ensembles. Also in 2015, the ENCORE[146] toolkit was released to address the "need for algorithms and software that can be used to compare structural ensembles in the same way as the root-mean-square-deviation is often used to compare static structures." ENCORE very effectively enables the comparison of large sets (10,000s) of protein structures, however, ENCORE does not provide residue-level details of where differences occur. As noted by the authors, "it is difficult to provide a simple geometric interpretation of the scores, [and] we suggest they are currently best interpreted in a relative fashion (e.g. ensemble A is more similar to B than to C)."

The power of Ensemblator 1.0 for providing geometrically meaningful details of structural differences was demonstrated in our original paper[141] through analyses of an RNAse Sa[147] NMR ensemble and its comparison to a 2-member X-ray ensemble, as well as comparisons for a different protein of an 8-member X-ray ensemble to three NMR-derived ensembles generated by different refinement

approaches. In addition to illustrating the value of Ensemblator 1.0 analyses, however, we acknowledged a serious limitation related to the need to have the compared proteins be identical in sequence and have input files contain identical atoms in the same order.

In addressing this limitation, we have developed Ensemblator v3, a substantially more robust, versatile and user-friendly tool. It has been entirely developed in Python, and extensive new features have been added such as multiple sequence alignments using MUSCLE[148] to handle proteins of diverse sequence, "ensemble clustering" of the models into subgroups, and the calculation of a novel "discrimination index" to quantify the levels of similarity/difference between any pair of compared subgroups, per atom or per residue. Recently, we applied the Ensemblator v3 to readily locate subtle differences between an NMR-based structure of the HIV reverse transcriptase thumb domain and the same domain as seen in the 28 highest resolution reverse transcriptase crystal structures[149]. Here, we describe the novel features of Ensemblator v3, along with three case studies which briefly showcase its utility for generating useful information and facilitating insight.

Description of the Ensemblator v3

Strategy

The essential comparison strategies implemented in Ensemblator v3 are identical to those of Ensemblator 1.0 and involve: (1) carrying out a complete set of pair-wise comparisons to define a set of global core atoms with consistent positions in all structures being compared (see Figure 2A of Clark *et. al.*[141]) and using those to guide a global overlay from which atom-level global comparisons can be made, (2) carrying out a complete set of pair-wise comparisons using the locally-overlaid dipeptide residual (LODR) as a measure of residue-level local backbone similarity (see Figure 3 of Clark *et. al.*[141]), and finally (3), for both the global and local comparisons of the two subgroups of structures for which comparison was sought, calculate four quantities: the two intra-subgroup variations, the inter-subgroup variation, and the closest approach of any member of subgroup 1 with a member of

subgroup 2 (see Figure 1A of Clark *et. al.*[141]). However, everything else about Ensemblator v3 is different as a result of the complete recoding from scratch in Python. The Ensemblator v3 has just two stages: "prepare" and "analyze". In the prepare stage, input structure files are processed to build a single PDB-formatted "ensemble file" for analysis. In the analyze stage, the comparisons noted above are carried out, and then either user-defined subgroups are compared as noted in step "(3)" above, or automated clustering is performed and the resulting subgroups are compared with each other. For each pair of sub-groups compared, the Ensemblator reports three key metrics. A "global" output file gives the RMSDs, global discrimination index (DI), and core-status for each atom based on a final "best" overlay; a "local" output file, which reports the LODR scores and local DI for each residue; and a discrimination index file which reports for each residue the global, local and unified DI values. The following sections provide a basic description of key steps, and readers are referred to the Ensemblator documentation (on GitHub) for further details[150] including a detailed description of these and other output files produced by the Ensemblator.

Preparation of an "ensemble file"

To prepare an "ensemble file" that contains the atoms common to all input structures, the user must provide a set of input structures in either PDB or mmCIF format. The Ensemblator will first convert each input file into a set of separate files for each chain (or model), and each alternate conformation present. Then either immediately, or after a sequence alignment is done, these files are assembled residue-by-residue into a set of models in which any atoms not present in every included structure are removed (e.g. truncating aligned Ser and Tyr sidechains both to Cβ, and for an aligned Lys and Met removing Cδ and Sδ but retaining Cε). If atoms in regions of interest are lost in this process, a user can identify and leave the causative input file(s) out of the analysis. To facilitate this, a maximum number of allowed chain breaks per model can be specified. A benefit of this process is that by limiting the residues present in any single input file one can trivially create an ensemble file that only has a specific region of interest.

If all PDB files to be included have the same residue numbering throughout, they can be combined without carrying out a sequence alignment; otherwise, an alignment using MUSCLE[148] (which must be installed separately) may be done as part of the preparation. In this case, the sequences in the split files are aligned and the residues are renumbered per the multiple sequence alignment. To filter out nonhomologous chains from the final ensemble, a series of alignments are carried out with increasing stringency on model similarity (using a BLOSUM62[43] based similarity score) until the cutoff reaches a user defined value. In the aligned sequences, the residue numbers in all output files will reflect the multiple sequence alignment numbering rather than the original values.

Determination of the common-core atoms and global overlay

The global overlay of the structures is the standard least-squares best overlay calculated using a set of 'common-core" atoms that are selected using the process described by Clark *et. al.*[141] In Ensemblator 1.0, the common-core calculation was carried out for a wide range of pre-specified cutoff-distances (d_{cut}) and then the user had to decide which result to use. In the Ensemblator v3, the user can define d_{cut}, but the recommended option is to have it identified automatically by the Ensemblator. What d_{cut} value is "best" is subjective and will depend on the ensemble and the goal of the analysis, but our experience with a variety of proteins has led us to conclude that a good place to start is with a common core including 20-40% of the atoms. So in the automatic mode, a systematic process is followed to obtain a d_{cut} value producing a common core in that range.

Clustering of structures using evidence accumulation and ensemble clustering

A completely new feature of the Ensemblator v3 is automatic clustering. In the process of defining the common-core (above), the program accumulates for every pair of structures the fraction of atoms (p) in the core of that pair (i.e. aligning closer than d_{cut}) and their RMSD ($RMSD_c$) as well as the RMSD of the non-core atoms ($RMSD_{nc}$). The distance score used for clustering is defined as: *distance score* = $RMSD_c^{p} * RMSD_{nc}^{1-p}$, which is essentially a weighted geometric mean[151] of the

$RMSD_c$ and $RMSD_{nc}$. This novel distance metric has a few useful qualities. First, its two extreme values are simply $RMSD_c$ (if all the atoms are in the core) or $RMSD_{nc}$ (if no atoms are in the core). Second, because $RMSD_c$ will always be smaller than $RMSD_{nc}$, it will be more heavily weighted, due to the fact that a geometric mean is always smaller than an arithmetic mean when the terms are not all equal and all the terms are positive[152]. This is advantageous as we are more interested in the similarity of the core atoms than we are in the difference in the non-core atoms (but we still want to utilize information present in the non-core atoms). Third, the favoring of $RMSD_c$ also makes the values more resistant to extreme outliers. Using a more traditional distance metric, such as the arithmetic mean or the total RMSD, outliers would be far away from the other points in the N-dimensional space (for N-models), increasing the overall sparsity and worsening the quality of the subsequent clustering experiments.

The clustering is done by "ensemble clustering", that combines the results from multiple independent clustering approaches and is known to be more robust and insensitive to noise[153]. First, affinity propagation[154,155] is carried out, perhaps a few thousand times, varying the "preference" value from a low number that results in a single cluster, increasing by 1% each run until every point is its own cluster. Next, k-means clustering[156] is performed 10*(N-2) times, increasing the specified number of clusters, K, from 2 to N-1, and running ten iterations for each K value with different initial conditions. Each of these independent clustering results are used fill a co-occurrence matrix, a form of evidence accumulation[157], which records how many times each model is clustered with each other model. Finally, agglomerative hierarchical clustering is performed on this co-occurrence matrix as a "finishing technique"[153], and provides both the final clusters used for comparisons, and a dendrogram that indicates the relationships between the models and clusters. The final number of clusters, between 2 and a user-specified maximum number of clusters, will be the solution that provides the best average silhouette index[158] (a metric that is partly designed to detect misclassifications in clustering experiments). Finally, the Ensemblator performs t-SNE dimensionality reduction[155] on the original

distance matrix to provide an independent visual interpretation of the distribution. This dimensionality reduction uses a set of default parameters and may not perfectly show groupings in two-dimensional space that are analogous to the results of the clustering experiment, but all the distance information that is used in the clustering and dimensionality reduction is output to a file, allowing for more detailed analyses by the user.

The local overlay strategy and LODR score

As described by Clark *et al*[141], the locally overlaid dipeptide residual (LODR) is a simple distance-based quantity that assesses the similarity between any pair of backbone conformations. Briefly, to calculate it, the equivalent dipeptides from two models are overlaid based on the Cα, C, O, N, and Cα atoms of the peptide unit preceding the residue, and then the LODR-score is defined as the RMSD between the C, O, N and Cα atoms in the subsequent peptide unit (see Figure 3A of Clarke *et al.*[141]). Given this definition, LODR values cannot be calculated for the first and last residues in a protein or for residues bordering chain-breaks as there are not complete peptide units on both sides of these residues. LODR values range from 0 Å for identical conformations to ~5 Å for residues differing by 180˚in their phi values (see Figure 3B of Clarke *et al.*[141]).

Calculation of the discrimination index (DI)

Our "discrimination index" combines local and global information into a single metric that indicates how similar or different a given residue or atom is between two sets of structures. It is based on the mathematics used for calculating silhouette scores[158]. Considering two groups of structures (M and N), a discrimination score assessing the significance of differences can be calculated for each atom in each group, as the mean of the pairwise distances between the groups minus the mean of the pairwise distances within the group, divided by the higher of the two values:

$$discrimination\ score = (\text{mean}(d_{inter}) - \text{mean}(d_{intra})) / \max(\text{mean}(d_{inter}), \text{mean}(d_{intra}))$$

Because this measure differs depending on which group is taken as the reference group, values are calculated for group M and for group N and averaged to create the global discrimination index (DI) for each atom. To create a residue-level global DI, the global DI values for the N, Cα, C, and O atoms of each residue are averaged. A local DI for the backbone conformation is similarly calculated for each residue based on the LODR values. Each of these scores is saved and output in a table containing all the global or local information about each atom or residue, respectively.

A unified DI for each residue is then defined as the average of the residue-based global and local DI values. This measure goes from near 0 to near 1 as the groups go from indistinguishable to systematically distinct. Whereas the individual local and global DI values have additional information, the value of the unified DI is that it provides a single plot that allows facile identification of the most significant regions of backbone difference between the two groups being compared.

Program Details

The Ensemblator v3 is written in Python, and is currently maintained and distributed from a GitHub repository[150], where the source code is freely available. It exists as three python scripts: a core script which does all the computation, and two handler scripts which use either a command line or graphical user interface (GUI) to pass options and input files to the core script. As output, the Ensemblator provides all the data produced during analysis, as well as automatically generated plots for all the key metrics. The Ensemblator v3 GUI was written using the *tkinter* Python library, which should ensure compatibility with a wide range of systems. Furthermore, the Ensemblator is capable of running on multiple processors in parallel to speed larger comparisons. Issues and bugs are reported and tracked on GitHub. As they are developed, other useful, related scripts will also available in this repository (*e.g.* currently available is a script to choose a representative model from each subgroup of a larger ensemble).

Case Studies

Case Study 1: Basic tests using the NMR solution structure of RNase Sa

Since Ensemblator v3 was a rewrite from scratch, we sought first to document that the basic algorithms are correctly coded by showing that it delivers the same results for previous test cases. We chose the analysis of an RNase Sa NMR ensemble[147] for which we had identified a peptide flip between residues 82 and 83 relative to the crystal structure (PDB code: 1RGG), and also that residues 31-33 adopted a conformation in models 19 and 20 of the NMR ensemble that were unusual enough to be considered implausible[141]. The reanalysis of the RNase Sa ensemble with Ensemblator v3 not only reproduced our earlier results (data not shown), but it additionally illustrated the value of automatic clustering to lead to further insight. The 20 RNase Sa NMR models cluster into three subgroups, and consistent with previous results, residues 30-33 are highlighted by their high unified DI as a region of difference between groups (Figure 4.1A). Notable is that residues 45-53 have a DI even higher than residues 30-33, and are thus a region of even greater significant difference.

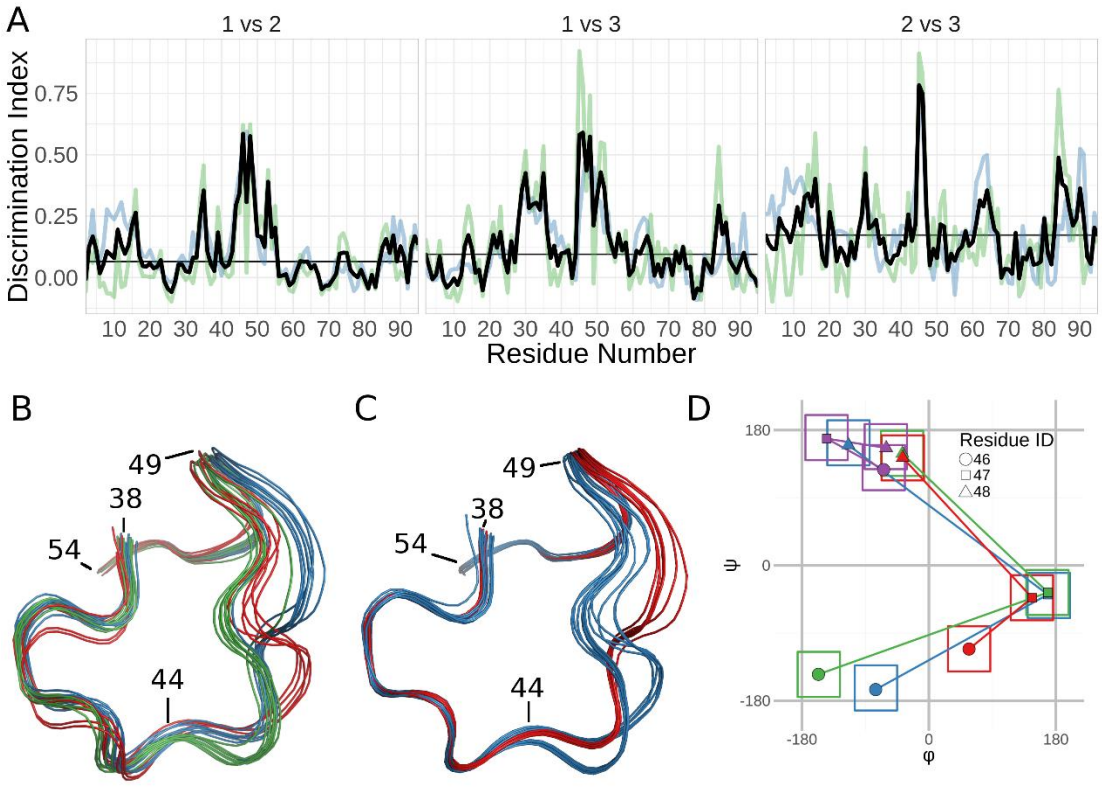

Figure 4.1 Analysis of the solution structure of RNase Sa.

(A) Discrimination Index (DI) plots for the pairwise comparisons of the three groups identified by the Ensemblator. The residue-based global DI (blue) and the local DI (green) are averaged to create the unified DI (red). The median unified DI is also indicated (black line). **(B)** Wire-diagram tracing of the backbone path in the region of largest inter-group difference (residues 44-49): Group 1 (blue; models 1,2,7,8,10,13-15); Group 2 (green; models 3-6,9,11,12); Group 3 (red; models 16-20). **(C)** Wire-diagram as in (B), for groups identified by analysis of only residues 38-58: Group 1 (blue; models 3-7,9,12,16-20); Group 2 (red; models 1,2,8,10,11,13-15). The tighter backbone spread results from the more local overlay. **(D)** φ,ψ values for residues 46 (circles), 47 (squares), and 48 (triangles) representative of the three groups shown in panel (B) (blue, green, red) and the X-ray structures (purple). The ±30° boxes indicate the areas used in Protein Geometry Database[28] searches for tripeptides present in structures solved at 1.5 Å resolution or better that have no more than 25% sequence identity to one another. The tripeptide conformation in all the X-ray models was found 467 times (0.34% of all tripeptides), while zero occurrences were found for the NMR conformations.

Inspection reveals three distinct conformations for residues 45-53 that mostly but not perfectly match the clusters (Figure 4.1B). Such local mismatches can occur if the outlier models are more similar to their respective groups elsewhere, because the clustering is based on global similarity rather than the similarity of this particular region. This illustrates that the DI values, by taking all models into account, is much more useful for discovering significant differences compared with our previously recommended strategy of looking for regions where the closest approach distance is greater than the within group variation; the latter criteria shows nothing abnormal at this region. A quick rerun of the Ensemblator on only residues 38-58 results in a precise separation into two groups (Figure 4.1C), with one of the groups having two slightly different conformations.

To determine the relative plausibility of the three backbone paths, we looked at the φ and ψ angles of the three-residue segment with the highest deviations (*i.e.* residues 46, 47, 48) in the 20 NMR models as well as in a set of RNase SA crystal structures (Figure 4.1D). Surprisingly, each of the three NMR paths through φ,ψ space differ substantially from the conformations in all of the crystal structures. Even more notable, Protein Geometry Database[28] searches showed that none of the conformations adopted by residues 46-48 in the NMR models has ever been seen in a large set of deposited high resolution crystal structures, whereas the conformation observed in the crystal structures is observed 467 times (~0.4% of tripeptides) (Figure 4.1D). This suggests that just like the conformations seen for residues 31-33 in models 19 and 20[141], all of the conformations of residues 46-48 in all of the NMR models are dubious.

Case Study 2: Clustering of a mixed-source ensemble using the FK506 binding protein (FKBP).

Recently, Tyka *et al.*[159] showed, using a set of FKBP models produced from X-ray crystallography, NMR, and Rosetta, that the models are all similar, with the Rosetta-produced template based models (based on an FKBP crystal structure) having less variability than the NMR models, but more than the crystal structures (see Figure 6 of that paper). We requested these models to test the extent to which the

Ensemblator could guide the discovery of systematic differences among them. The ensemble we received included 34, 30 and 25 models designated as "X-ray", "NMR" (from two studies), and Rosetta, respectively. Ensemblator analysis with automatic clustering readily divided the set into three subsets that as visualized by the t-SNE dimensionality reduction plot (Figure 4.2A) can be seen to largely, but not perfectly, correspond to their original labels. Importantly, the exceptions all identified structures that had misleading designations: the ten models of one NMR ensemble (PDB entry 1F40) that clustered with the X-ray structures were from a study[160] in which the ligand placement into FKBP was based on NMR observations, but the protein coordinates were taken unchanged from a crystal structure (PDB entry 1FKG); and the two models designated as "X-ray" (entries 1FKS and 1FKT) that grouped with the NMR-derived models in PDB entry 1FKR[161], were actually not crystal structures, but were a 21st member of the NMR ensemble and an average structure based on the other 21 models. Based on a consultation with the Baker Lab, it seems that the inclusion of these models in the set we received stemmed from a difficulty in retrieving archived data, and as these mislabeled models brought no unique information, we removed them from further analyses.

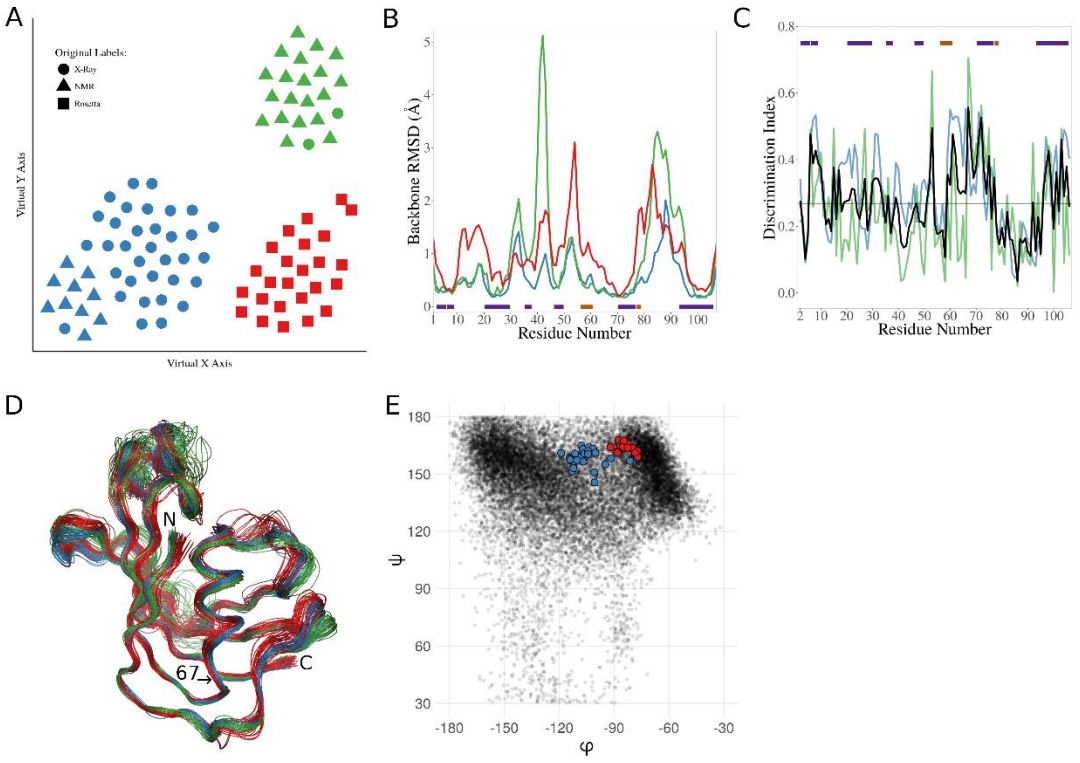

Figure 4.2 Analysis of a mixed-source ensemble of the FK506 binding protein (FKBP)

(A) t-SNE dimensionality reduction results showing a 2D visualization of the relationships between the models in the N-dimensional space used to cluster them. Per the key, the shape of each point represents the original label for a given model, and the clusters are differentiated by color (1 – blue, 2 – green, 3 – red). **(B)** Backbone RMSDs along the chain for the final set of X-ray (blue), NMR (green), or Rosetta (red) produced models. The bars indicate positions of β-strands (purple), and α/3-10 helices (orange). **(C)** Discrimination Index (DI) plots for the Rosetta models vs the X-ray models. Residue-based global (blue), local (green), and unified (black) DI are shown, along with the median unified DI (horizontal black line). Secondary structure indicated as in (B). **(D)** Wire-diagram tracing the backbone for the X-ray (blue), the NMR (green) and the Rosetta (red) models. The N- and C-terminal are indicated, as well as the position of residue 67, at the base of an α-helix. **(E)** The φ,ψ-angles for serine 67 in the Rosetta (red) and the X-ray structures (blue) are shown. As context, the φ,ψ-values of all serine residues in crystal structures at 1.5 Å resolution or better with ≤ 25% sequence identity to one another are indicated (black dots).

With the mislabeled models removed, Ensemblator clustering perfectly separated the X-ray, NMR, and Rosetta models into subgroups, implying that systematic differences do exist between them despite their similar appearances. Consistent with the findings by Tyka *et al.*[159], the Rosetta models include more overall variation than the X-ray models, and the NMR models even more so (Figure 4.2B). However, the Ensemblator analysis yields the additional information that the higher variation in the NMR models is not at all uniform, but the NMR models have much more variation in two loops, even while they have less variation than Rosetta in two other loops. Examination of the unified DI plots reveals that while the NMR ensemble appears to be roughly equally distinct from the Rosetta and the X-ray models (Supplemental Figure 4.S1), the Rosetta and X-ray models are rather similar, but have a handful of high DI peaks (Figure 4.2C). It is outside the scope of this paper to analyze all the differences, but as an example we consider here the highest peak, near residue 67. Inspection of the models reveals that the absolute difference between the Rosetta models and the crystal structures at this position is quite small (Figure 4.2D), but it is significant because the variation in each subgroup is even smaller. The difference originates in the φ,ψ angles of Ser67, with the Rosetta models having values shifted toward the more densely populated P_{II}-region compared to most of the X-ray structures (including 2PPP, the structure that was used as the template for the Rosetta models) (Figure 4.2E). This shift could plausibly be caused by the Rosetta knowledge-based φ,ψ-potential[162], which would favor the more populated conformation.

Case Study 3: Domain and hinge residue identification using calmodulin (CaM) crystal structures

In Clark *et al*[141], it was noted but not demonstrated that the Ensemblator is designed for the analysis of single domains (or multidomain proteins that do not undergo domain movements), but that the local LODR comparisons done by the Ensemblator could be useful for identifying flexible hinge regions. This would in turn allow Ensemblator analysis of the separate domains. To illustrate this application we used calmodulin, which has homologous N- and C-terminal EF-hand domains, and

undergoes a large conformational change upon binding peptide ligands with what has been described as "no significant conformational change within each domain (residues 4 to 74 and 82 to 146)."[163] Using the CoDNaS database[71], we collected the set of all calmodulin crystal structures solved at 1.8 Å resolution or better. This X-ray ensemble contains 16 models from ten crystal structures that all have bound calcium and represent 6 different crystal forms; the five crystal structures without a peptide ligand are from the same crystal form.

Ensemblator clustering splits these 16 models into two groups, corresponding to the ligand-free dumbbell conformation and ligand-bound globular conformation with calmodulin wrapped around the peptide ligand (Figure 4.3A). As seen in the global RMSD plot, the first domains overlay well, making the second domains very distant (Figure 4.3B, middle panel). Based on this plot, it is impossible to learn about intra-domain global differences in the C-terminal domains because the domain shift dominates the plot. In contrast to the global analysis, the local analysis (Figure 4.3B, lower panel), shows clearly that within each domain the conformations are highly similar (low LODR scores), and readily identifiable is that residues 74-80 are linker residues that not only change conformation upon peptide-binding but also are somewhat variable among the ligand-bound models. With this information in hand, it is trivial to then build separate ensemble files after truncating one PDB input file to either contain only the N- or the C-terminal domain, to be able to perform a separate Ensemblator analysis for each domain. These runs then yield meaningful global results for both domains which combined with the unchanged local results leads to a more informative unified DI (Figure 4.3C and D).

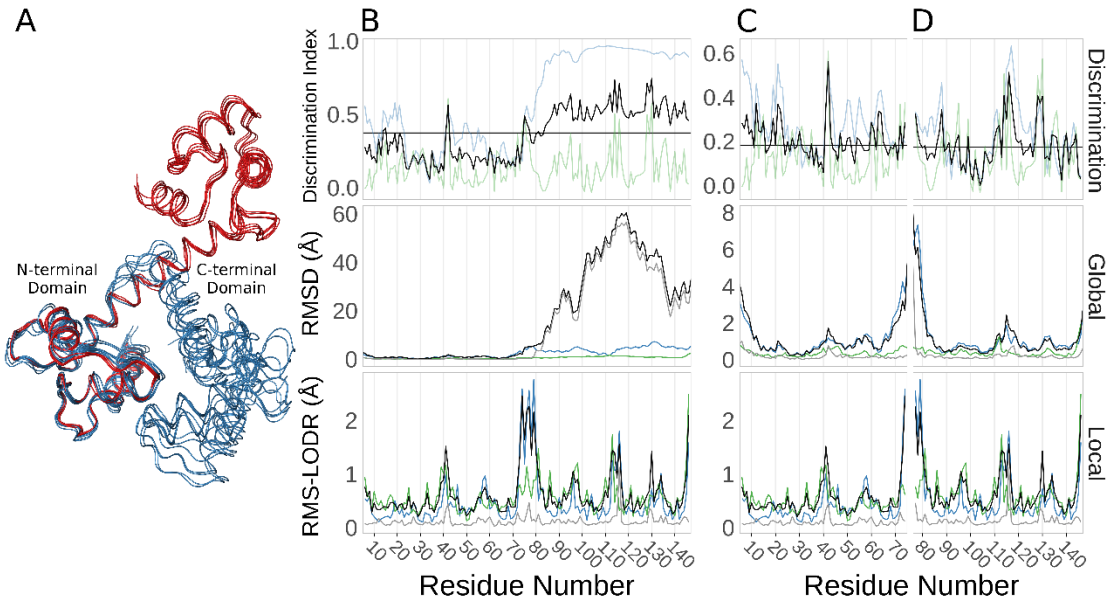

Figure 4.3 Ensemblator analysis of calmodulin (CaM) crystal structures.

(A) Wire-diagram backbone tracing for the ligand-bound models (blue), and the ligand-free models (red), as overlayed by the Ensemblator. **(B)** Discrimination indices (top panel; global (blue), local (green), unified (black), and median unified (horizontal black line)), and RMSDs from the global (middle panel) and local (bottom panel) comparisons for the entire CaM protein. In the global and local comparisons, the within group variation is shown for the ligand-bound (green) and ligand-free (blue) conformations. Also indicated is the inter-group variation (black) and the closest approach distances (grey). **(C)** As in (B), except the analysis only included the N-terminal domain. **(D)** As in (B), except the analysis only included the C-terminal domain.

Interestingly, the DI plots for the separate domains each contain a dominant peak occurring at residues 41 and 114 (Figure 4.3C and D, upper panel). These residues are at equivalent positions in the two EF-hand domains, in a loop between the E and F helices. Whereas both have been noted before as residues that commonly interact with the bound peptides[164,165], we have not found any mention in the extensive calmodulin literature that upon ligand binding these residues tend to undergo a similar conformational change from the beta-region to the P_{II}-region of the φ,ψ-plot (Supplemental Figure 4.S2). This backbone conformational change does not occur in every ligand bound conformation, but may be of interest for further analysis.

Discussion

Ensemblator 1.0[141] introduced a tool that allowed the direct comparisons of ensembles of protein structures, rather than requiring the ensembles to be represented by a single exemplar or average structure. It also provided detailed information, for both global and local comparisons, that allowed an unprecedented residue-level pinpointing of significant differences between the sets of structures. Our original goal in improving on Ensemblator 1.0 was to make the program much more widely applicable, by making it robust to: differences in the input coordinate files such as missing atoms and changes in residue numbering or atom order, and minor differences in sequence such that point mutants and homologs could be included in comparisons. That we have done this is well documented though Case Study #3 in which a diverse set of PDB entries obtained from the CoDNaS database for calmodulin are quickly combined for analysis, and when a separate analysis of the N- and C-terminal domains is targeted as a follow-up study, these files are also easily prepared. Three additional minor program enhancements are an algorithm for finding a suitable d_{cut} value for carrying out the global overlay, a GUI interface, and the ability to run on multiple processors to increase speed and scalability. The most time-intensive part of the Ensemblator is the pairwise comparisons, and running on eight cores, its runtime is about two hours for an ensemble of ~1000 200-residue structures.

In addition to these important technical improvements, the Ensemblator v3 also includes two innovations that greatly enhance the information it can provide. These are a clustering option that automatically finds conformational subgroups, and the reporting of a novel "discrimination index" as a useful metric for identifying regions of significant difference or similarity. In Ensemblator 1.0, the user had to define which models belong to each group of structures being compared – such as comparing two NMR-ensembles to each other, an X-ray ensemble to an NMR ensemble, or a set of liganded structures to unliganded structures – but this user-driven approach is much less powerful than allowing features common to groups of structures to be automatically recognized through clustering. Whereas there is no universal best-approach to clustering, we have implemented a type of "ensemble

clustering" that has been documented as being broadly effective, especially on biological data[153]. Each of the three case studies illustrates utility of the clustering for discovering interesting subgroups among ensembles analyzed. Especially noteworthy in our view is Case Study #2 in which the clustering makes it absolutely clear that FKBP models derived from crystal structures, NMR analyses, or from Rosetta modeling are not simply versions of the same average structures with differing amounts of uncertainty or spread; instead they are readily distinguishable as being different from each other, despite that not being obvious by visual examination.

Each of the case studies also nicely illustrates the value of the novel discrimination index (DI) as a major improvement over our previous suggestion[141] that the most significant differences between subgroups of structures will be the places at which the closest approach of any member of the two subgroups was larger than the spread of the ensembles. The latter metric completely misses cases in which two sets of models are widely different, but happen to have at least one member that is similar. The unified DI, in contrast, takes the full ensemble information into account as well as giving weight to both the global comparison and to the local comparison. For RNase Sa, this DI strongly identifies residues 45-50 as a segment of major difference between subgroups (Figure 4.1A) even though each subgroup has a member that is like the other subgroup (Figure 4.1B). When comparing the X-ray FKBP structures to the Rosetta produced structures, the top DI peak identified a small but significant difference would otherwise be entirely non-obvious from visual inspection of the ensembles (Figure 4.2D), but that could be a clue to how to improve the Rosetta force field (*e.g.* Song *et al,* 2011[87]). For calmodulin, in addition to making the hinge region readily identifiable, the discrimination index enabled the identification of a small conformational change within each domain that seems to strongly correlate with ligand-binding and that, as far as we found, had not been noticed before despite extensive studies that have been done on calmodulin. The DI metric is simple to understand and use in practice, and further examples of its utility can be seen in our recently published identification of regions of significant difference between a solution NMR structure of HIV reverse transcriptase thumb domain and the same domain as seen in crystal structures of reverse transcriptase [149].

The effective analysis of ensembles of protein structures requires many tools, and the Ensemblator fills a gap by being able to compare ensembles of on the order of hundreds of structures and provide exquisitely detailed information about atom- and residue-level differences in conformation between groups of models. This purpose is quite different than that of ENCORE which enables the comparison of very large sets (10,000s) of protein structures, but does not provide residue-level details[146]. We suggest that the programs could effectively be used in concert with each other, for instance by using ENCORE to group very large sets of structures and then using the Ensemblator to analyze representatives of each of the ENCORE groups, to identify the nature of the most notable conformational differences between them.

The Ensemblator v3 takes the conceptual advances of the original Ensemblator[141], and makes them easily applicable to a much wider set of protein models. Furthermore, it extends the general methodology such that the only strictly required user-input is a set of protein structures to analyze; the user no longer needs to have any preconceived knowledge about the structures (*e.g.* subgroups to compare or the d_{cut} value that would identify the ideal core). While the Ensemblator provides the greatest amount of information when applied to single domains, its application to multi-domain proteins allows the identification of domains that have consistent internal folding as well as variable linker regions that may connect them, and as is seen in Case Study #3, it can be serially applied to the whole protein and then to identified domains to maximize the information gained. Also, even for a single domain, as seen in Case Study #1, it can be effectively applied to any substructure of interest to ensure that the global overlay and clustering provide the greatest information about that region (Figure 4.1C). The kinds of insights generated here in the three well-studied proteins used as case studies, along with our recent analysis of an NMR-derived structures of the HIV thumb domain[149] illustrate how Ensemblator comparisons add a unique and useful tool to the structural biology toolbox.

Supplementary Material

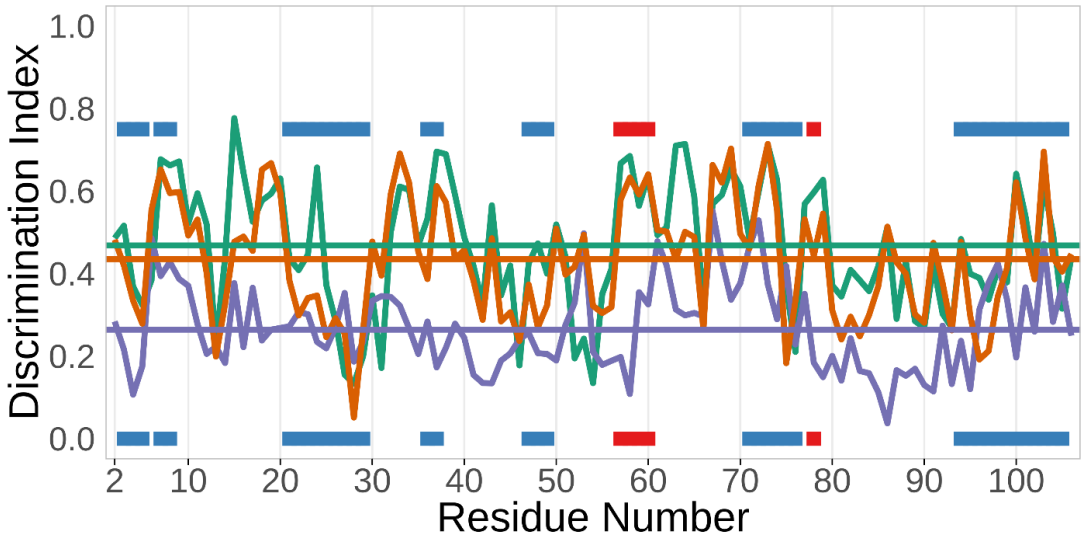

Figure 4.S1 Unified Discrimination Index values for each pair of groups of a mixed-source FKBP ensemble.

The unified DI curves are shown for the 'NMR vs Rosetta' (orange) 'NMR vs X-ray' (green), and 'X-ray vs Rosetta' (purple) comparisons, with the median unified DI of each comparison indicated by a horizontal line. Colored bars indicate the positions of β-strands (blue) and α/3-10 helices (red). The much lower values of the purple trace indicate that the X-ray structures and the Rosetta models are much more similar to each other than either group is to the NMR models.

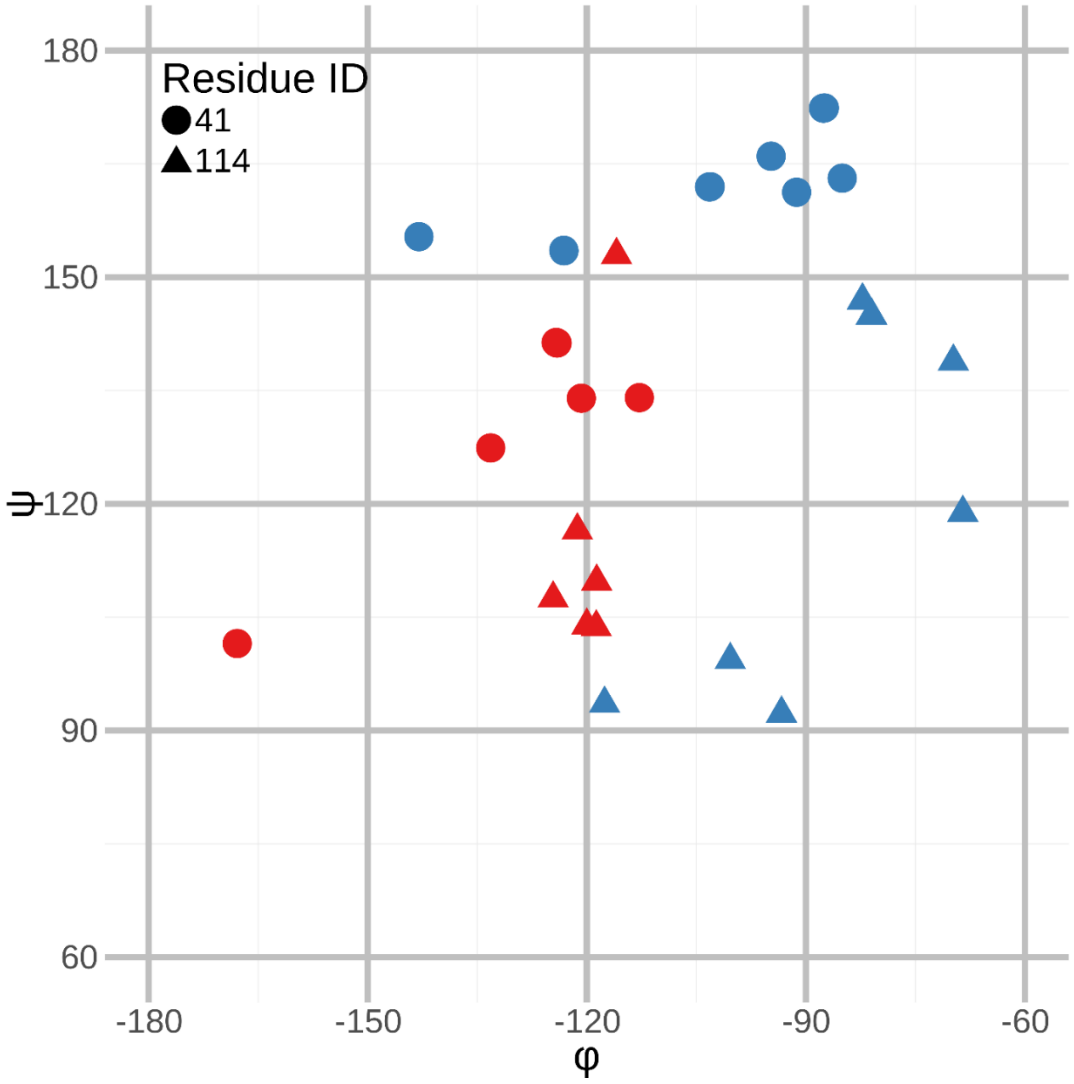

Figure 4.S2 Commonly observed conformational shift of residues 41 and 114 upon ligand binding in calmodulin crystal structures.

Residue 41 (circles) and residue 114 (triangles) are equivalent residues in the N- and C-terminal domains of calmodulin, and frequently (5/7 of residue 41, 4/7 of residue 114) undergo a conformational shift upon ligand binding, moving from in the general β-region to the P_{II}-region. Each point represents the φ,ψ value observed in a given model, from the ligand-bound (blue) or ligand-free (red) crystal structures. The ligand bound structures used in this analysis were: 1MXE, 1QS7, 1QTX, 2X51, and 3L9I. The ligand-free structures were: 1CLM, 1EXR, 1OSA, 3CLN, and 4CLN.

Acknowledgements

This work was supported in part by NIH grant R01GM083136 to PAK. We thank Hahnbeom Park, Mike Tyka and David Baker from David Baker's research group for providing their FKBP models for our test case. We thank the developers of Biopython, scikit-learn, and SciPy; without these fantastic tools, it would not have been possible to develop the Ensemblator v3. We also thank Nathan Jespersen for helpful discussions and extensive beta testing.

Chapter 5

Conclusion

Impacts and Highlights of Reported Work

In this section, I will discuss some of the major findings from each chapter, and how the protein science community has received the work. Then, in the remaining sections of this concluding chapter, I will outline possible future work, and wrap up with some conclusions on Boltzmann's principle, the protein folding problem, and the nature of research into protein structure.

In chapter 2, we used ultra-high resolution crystal structures to describe the details of a high-energy transition conformation that occurs during protein folding and conformational switching. These structures had stabilized individual residues in conformations that were analogous to conformations from partway along the transition pathway. The significant findings of this work are fourfold: (1) for the first time, experimentally determined information about this high-energy transition was obtainable; (2) despite the energetic cost of stabilizing this conformation, folded proteins (even very stable ones) can do so in a huge variety of contexts and environments; (3) the methodology used to obtain this information could be generally applied to other systems if enough data are available; and (4), contemporary molecular dynamics (MD) forcefields are not able to accurately reproduce the details of this transition conformation. Thus far, it is these last two findings that appear to have had the largest impact, with researchers using the same methodology to describe RNA folding pathways[166], or as partial justification for improvements made to the AMBER forcefield for MD[167]. The work in chapter 2 stands as an example of the way that atomistic research can contribute to more holistic research, as details from ultra-high resolution crystal structures are used to improve MD forcefields.

In chapter 3, we investigated claims that observed instances of peptide non-planarity seen in crystal structures represented errors in model building rather than the reality of widespread "non-ideal" geometry in protein structure. We were able to clearly show that model refinement is improved by using standard restraints on planarity, as compared to tightened restraints that more strictly enforce planarity. We also concluded that the distribution describing the planarity of the peptide unit is relatively unchanged across different subsets of structures and different refinement

methods, indicating that the observed non-planarity in ultra-high resolution structures in the PDB is not an artifact of refinement, but a real feature present in the experimental data and thus the underlying structures. As part of this analysis, we showed that using very tight restraints on planarity introduces local errors into protein models. The impact of this research is primarily that it helps to set the record straight about non-planarity in proteins. It is now widely recognized that non-planarity is a feature of real protein structures, and that structures cannot be accurately represented by models that strictly enforce planarity[168].

Lastly, in chapter 4, we describe and demonstrate the use of the Ensemblator V3, a software tool that quantifies similarities and differences between ensembles of protein structures, at residue or atom level. There exists a gap in the field of structural biology, where there is an increasing need for tools that compare more realistic representations of protein native state, such as ensembles, as opposed to single models. The features offered by the Ensemblator, especially the ability to fully automate analyses, and the use of the discrimination index to locate significant regions of similarity or difference, represent a great step forward in filling this gap. Already, the Ensemblator has been used in collaborations to compare structures of HIV reverse transcriptase thumb domain[149], and to compare methods for producing NMR ensembles of alpha-lytic protease (manuscript under review). Though the software is only now becoming available in an easy-to-use and practically useful form, I am hopeful that the Ensemblator will have a substantial impact on the field, as it has valuable new capabilities that enable researchers to compare large sets of structures without losing the immense amount of information an ensemble contains beyond what is contained in a single model. I also think that the software has the potential to be a useful tool for moving research forward in the field, aiding researchers in investigating how to effectively describe the scope of structural variation that is encompassed in the native state of a folded protein.

Directions for Future Research

When it comes to the atomistic research like that in chapters 2 and 3, which uses ultra-high resolution crystal structures to define more accurate details of protein

structure, the possibilities for similar research to be done in the future are almost limitless. That said, there are a few specific projects that would be especially interesting to carry out. In the case of the transition between the two "allowed" sides of the Ramachandran plot, we investigated the conformation of the transition that occurs near $\varphi = 0°$. However, this is not the only possible pathway for transition; as discussed in chapter 2, a residue could also pass through the $\varphi \sim +160°$ region. At the time that the study was completed, it was impossible to describe the conformation of that transition using the same methods described in chapter 2. First, there are simply not enough residues trapped in those conformations to get reliable averages for measurements of the details of the transition conformations in say 20˚ steps in φ. Furthermore, more total data would be needed than in the $\varphi \sim 0°$ transition; since the "least disallowed" region on this side of the plot is a broad valley rather than a narrow pathway, it need not be the case that every transitioning residue would follow the same path. Overcoming these difficulties involve the same simple solution: with more data, it might be possible to describe this alternate transition.

Regarding the work in chapter 3, there is an immediately achievable project that follows the same template, but with a different feature of protein structure. In a different report, Dr. George Rose again suggested that what we see in deposited protein crystal structures is incorrect, and needs to be fixed by including tight restraints (in this case on hydrogen bonding geometry)[169]. Panasik *et al*[169] investigated β-turns in protein crystal structures, and found that a surprising number of them had poor geometry for the *i* to *i-3* hydrogen bond that defines these types of turns. Their conclusion was that this was due to errors in model building, and that rearranging the chain to improve the hydrogen bonding would thus improve the quality of the models. Contrary to this, when we carried out preliminary research, we observed that almost all these "non-ideal" turns are in fact very well defined, and almost all have a very plausible explanation for their unusual geometry (*e.g.* many are so-called "water-bridged" turns, with a water molecule that hydrogen bonds to both residue *i* and *i-3*). This is another case where ultra-high resolution crystal structures can inform about the details of "non-ideal" or non-canonical geometry, and represents a very achievable research project worth doing in the near future.

Lastly, work on the Ensemblator is still ongoing. Aside from general bug fixing and maintenance, there are a few new features that could be substantive improvements. Just as the LODR enables local, context-free analysis of differences in backbone conformation, a new metric could be developed that would compare sidechain conformations in a similar way. Even before such a metric is implemented, it would still be possible to modify the Ensemblator to provide a graph that would reveal information about side chains; results from the global analysis already contain details about sidechain atoms. Another possible feature that would be interesting to research would be shifting to a more continuous definition for the common core. Currently, the common core is made up of only the atoms that are core atoms in every pair of models, and only common core atoms are used to overlay all the models for the global analysis. As a different algorithm, it would be possible to assign each atom a variable "core" value, which would range from 0 to 1, and capture the portion of pairs of models in which it was a core atom. At the very least, this would be interesting information to have about an ensemble. At best, it could also be used to generate a weighting scheme to overlay the models using all atoms. Careful testing would need to be done to determine if and how this improved the quality of the overlays.

Concluding Statements

In chapter 1, I discussed the relationship between Boltzmann's principle, and our understanding of protein structure. As one chooses, it is possible to label research as atomistic or holistic; in this case chapters 2 and 3 represent the atomistic work I completed, and chapter 4 represents a more holistic project. I have demonstrated some examples of atomistic or holistic work, and I have discussed the relationship between the two modes of description, but important questions remain: Why describe things in these terms at all? What is gained by looking at the study of protein structure in this way? Before answering these important questions, first let me describe what we *do not* gain from thinking in these terms. In imagining a large holistic understanding of protein structure, rich in atomistic details, one might be tempted to believe it is possible to obtain a "complete" understanding of protein structure (*i.e.* to

know everything there is to know about protein structure). The point that I want to get across is that this goal is not only unrealistic, but it is fundamentally unobtainable.

Mathematician Norbert Wiener once said "The best material model of a cat is another, or preferably the same, cat"[170]. Wiener was discussing the extreme limit of fidelity for a model. If a model were to become maximally faithful to the thing it described, it would simply *be* the thing that it described. As discussed previously, when atomistic details are used to create a holistic model, there must be a loss of information. This is like the difference in entropy between the macrostate and microstate descriptions of a system; it is built into the definition. The difference here however is that we are not discussing the details of a closed, isolated, and theoretically enumerable system (*e.g.* a noble gas in a container). Instead, the entirety of "protein structure" is a nebulous concept, with fuzzy borders. As we include into any model more details required to understand how protein structure works, we would also need to include models for many other, non-protein systems (*e.g.* when the ribosome, an RNA machine, "stalls" during translation, it has an impact on the folding of the protein[171], and thus an effect on the final conformation). So, any holistic model within the umbrella of "protein structure" would also bear relationships to other holistic models existing at the same "level" (*i.e.* with similar amounts of informational entropy). Though it is impossible to define this limit for certain, a "complete" description of protein structure would likely require so much external information, that it would be impossible to ever reach 100% fidelity. Even if this were not the case, the it remains true that the most accurate model of the system would simply the system itself (*i.e.* modeling FKBP folding by actually expressing and folding FKBP), and it is clear that this is not the goal that researchers typically have in mind when they investigate the nature of protein structure. This logical extreme in fidelity is not a target worth pursuing, instead emphasis should be placed on achieving a description of protein structure that has high utility, rather than one with high fidelity.

Now we can begin to answer our original question: what do we gain from thinking of protein structure research as atomistic or holistic? Statistician George Box coined a famous aphorism that is highly relevant here: "All models are wrong, but

some are useful"[172]. By explicitly considering the scale and context at which we define our models (*i.e.* roughly estimating the amount of entropy, by considering the discarded information that went into their creation), we save ourselves from the trap of trying to value models based on their "accuracy" or how "true" they are. The best metric for the value of any model is its utility. For example, the consensus in the protein structure research community is that a better description of the native state of a protein is needed, and that an ensemble of conformations is a better model for this than a single conformation. An overlooked fact however, is that while the ensemble description is more detailed and accurate, both models exist on the same level of entropy. If one had an ensemble of every conformation a given protein adopts in the complete native state, it *would not* be true that each conformation would represent a microstate, and the ensemble the macrostate. Instead, *the total ensemble of conformations* would represent *the* microstate, and there would be no macrostate. If everything has been enumerated, then there is no loss, there is no holistic model, there is no macrostate. What utility would such an ensemble have? To use it to answer any question about that protein's structure, other than "What are all the exact possible positons for all the atoms in this protein?", would require a tremendous amount of work. It is important to consider this, because to some extent this is the stated goal of attempts to solve the protein folding problem. Most current methods for predicting structure attempt to produce such ensembles of all "real" conformations as faithfully as possible. Not only is this goal so difficult as to be essentially impossible to achieve, even if it were achieved, it would be impossible to test. If we could test it, we wouldn't need to predict the conformations in the first place, we would simply solve them all experimentally. To predict these total ensembles is a Sisyphean task, one that even if achieved would have little practical value to researchers who are interested in actually answering questions about protein structure. But what is the alternative?

The value of thinking of models as atomistic or holistic is that it solves exactly this problem. In this dissertation, I have called the Ensemblator a holistic work. This is not because it deals with ensembles of protein structures rather than single models, it is because it explicitly and irreversibly *discards* information about protein structure

ensembles. The Ensemblator allows researchers to input a very atomistic, exhaustive description of a protein's structure (*i.e.* an ensemble of many conformations), and it outputs very holistic, simple results (*e.g.* a discrimination index that reveals exactly where the most significant conformational differences are). The utility of the Ensemblator arises from the processes it uses to discard information during the analysis stage. This process is truly holistic, and macro-scale: information was discarded, and it cannot be reversed (the fact that typically that information is still saved on the computer is irrelevant). The Ensemblator generates entropy, in very specific ways, and that is what makes it useful. Likewise, the works in chapter 2 and 3 are similar. I call them atomistic because each project ended when everything was enumerated, but one value of both works is that a generalized form of the results could be incorporated into forcefields or other models of protein structure to improve their utility; that is a holistic form of the results from chapter 2 and 3.

How can we understand protein structure? The short answer is: we can't. We can only model it with differing levels of fidelity and entropy. It should be possible to solve the protein folding problem, but only by defining the "answer" as something that is obtainable. What is the best way to describe the native state of a protein? By enumerating all of the possible conformations, exhaustively? There might be more utility in having a model of the native state that is more holistic. Such a description might have information about stability, allostery, charge, dynamics, and other such macro-scale descriptors; instead of an exhaustive list of every position of every atom. The utility of the model should be what determines what information is included or discarded. The difference in entropy between the sequence of a protein, and every single conformation it adopts is extremely large; *de novo* structure prediction with this as its goal might never be achieved. The difference in entropy between sequence and a more holistic model of native state might be much smaller, making it likely much easier to predict.

As protein structure research goes forward, it will always be critical to consider the exact nature of the questions being asked, and the scale of the goals one is striving toward. By thinking about protein structure with this Boltzmann guided framework in mind, one can describe the trees *and* the forest (or at least *a* forest), as

utility demands. It will always be possible to solve problems and answer questions about protein structure, as I have done within this dissertation, however, it will never be possible to "complete" our understanding of protein structure. There will always be new problems, new mysteries, and new models to build.

References

1. Freitas Jr. RA Table 3-2. In: Nanomedicine. Landes Bioscience; 1999.

2. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838.

3. Pace CN, Shirley BA, McNutt M, Gajiwala K (1996) Forces contributing to the conformational stability of proteins. FASEB J 10:75–83.

4. Callaway DJ (1994) Solvent-induced organization: a physical model of folding myoglobin. Proteins 20:124–138.

5. Tsai C-J, Maizel JV, Nussinov R (2000) Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. Proc Natl Acad Sci U S A 97:12038–12043.

6. Dill KA, MacCallum JL (2012) The Protein-Folding Problem, 50 Years On. Science 338:1042–1046.

7. Stefani M, Dobson CM (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. J Mol Med 81:678–699.

8. Kim PS, Baldwin RL (1990) Intermediates in the folding reactions of small proteins. Ann. Rev. Biochem. 59:631–660.

9. Sippl MJ (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Computer-Aided Mol Des 7:473–501.

10. Bull HB, Breese K (1974) Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. Arch Biochem Biophys 161:665–670.

11. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc Nat Acad Sci 37:205–211.

12. Corey RB, Pauling L (1953) Fundamental Dimensions of Polypeptide Chains. Proceedings of the Royal Society of London B: Biological Sciences 141:10–20.

13. Pauling L (1993) How my interest in proteins developed. Protein Science 2:1060–1063.

14. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallographica Section A Foundations of Crystallography 47:392–400.

15. Engh RA, Huber R International tables for crystallography. In: Rossmann MG, Arnold E, editors. International tables for crystallography. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2001. pp. 382–392.

16. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7:95–99.

17. Ramachandran GN, Sasisekharan V (1968) Conformation of polypeptides and proteins. Adv. Protein Chem. 23:283–438.

18. Schäfer L, Cao M (1995) Predictions of protein backbone bond distances and angles from first principles. J Mol Struc: THEOCHEM 333:201–208.

19. Karplus PA (1996) Experimentally observed conformation-dependent geometry and hidden strain in proteins. Pro Sci 5:1406–1420.

20. Evans PR (2007) An introduction to stereochemical restraints. Acta Crystallographica Section D: Biological Crystallography 63:58–61.

21. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Op Struc Bio 19:145–155.

22. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Meth Enz 383:66–93.

23. Berkholz DS, Shapovalov MV, Dunbrack Jr. RL, Karplus PA (2009) Conformation Dependence of Backbone Geometry in Proteins. Structure 17:1316–1325.

24. Moriarty NW, Tronrud DE, Adams PD, Karplus PA (2014) Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement. FEBS J 281:4061–4071.

25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucl. Acids Res. 28:235–242.

26. Kendrew JC, Bodo G, Dintzis HM, Parrish R, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662–666.

27. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In: Protein Crystallography. Methods in Molecular Biology. Humana Press, New York, NY; 2017. pp. 627–641. Available from: https://link.springer.com/protocol/10.1007/978-1-4939-7000-1_26

28. Berkholz DS, Krenesky PB, Davidson JR, Karplus PA (2009) Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucl. Acids Res.:gkp1013.

29. Chellapa GD, Rose GD (2015) On interpretation of protein X-ray structures: Planarity of the peptide unit. Proteins 83:1687–1692.

30. Berkholz DS, Driggers CM, Shapovalov MV, Dunbrack RL, Karplus PA (2012) Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. PNAS 109:449–453.

31. Kleywegt GJ, Jones TA (1996) Phi/Psi-chology: Ramachandran revisited. Structure 4:1395–1400.

32. Zhou AQ, O'Hern CS, Regan L (2011) Revisiting the Ramachandran plot from a new angle. Protein Science 20:1166–1171.

33. Ho BK, Thomas A, Brasseur R (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. Protein Science 12:2508–2522.

34. Hollingsworth SA, Karplus PA (2010) A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. BioMolecular Concepts 1:271–283.

35. Oliveira CAF de, Hamelberg D, McCammon JA (2007) Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study. The Journal of Chemical Physics 127:175105.

36. Hollingsworth SA, Lewis MC, Berkholz DS, Wong W-K, Karplus PA (2012) (ϕ,ψ)2 Motifs: A Purely Conformation-Based Fine-Grained Enumeration of Protein Parts at the Two-Residue Level. Journal of Molecular Biology 416:78–93.

37. Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, et al. (2010) A short survey on protein blocks. Biophys Rev 2:137–145.

38. de Brevern A g., Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41:271–287.

39. Harris JI, Sanger F, Naughton MA (1956) Species differences in insulin. Archives of Biochemistry and Biophysics 65:427–438.

40. Aitken A (1976) Protein evolution in cyanobacteria. Nature 263:793–796.

41. Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science 185:862–864.

42. Thorne JL (2000) Models of protein sequence evolution and their applications. Current Opinion in Genetics & Development 10:602–605.

43. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919.

44. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL (2000) The Pfam Protein Families Database. Nucleic Acids Res 28:263–266.

45. Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) HOMSTRAD: A database of protein structure alignments for homologous families. Protein Science 7:2469–2471.

46. Holm L, Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 25:231–234.

47. Holm L, Sander C (1999) Protein folds and families: sequence and structure alignments. Nucleic Acids Res 27:244–247.

48. Bork P, Sander C, Valencia A (1993) Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Science 2:31–40.

49. Lupas AN, Ponting CP, Russell RB (2001) On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? Journal of Structural Biology 134:191–203.

50. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function11Edited by A. R. Fersht. Journal of Molecular Biology 291:177–196.

51. Sippl MJ, Weitckus S, Flöckner H In Search of Protein Folds. In: Jr KMM, Grand SML, editors. The Protein Folding Problem and Tertiary Structure Prediction. Birkhäuser Boston; 1994. pp. 353–407. Available from: http://link.springer.com/chapter/10.1007/978-1-4684-6831-1_12

52. Wodnak SJ, Rooman MJ (1993) Generating and testing protein folds. Current Opinion in Structural Biology 3:247–259.

53. Wolf YI, Grishin NV, Koonin EV (2000) Estimating the number of protein folds and families from complete genome data11Edited by J. Thornton. Journal of Molecular Biology 299:897–905.

54. Lella M, Mahalakshmi R (2017) Metamorphic Proteins: Emergence of Dual Protein Folds from One Primary Sequence. Biochemistry 56:2971–2984.

55. Murzin AG (2008) Metamorphic Proteins. Science 320:1725–1726.

56. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF (2008) Interconversion between two unrelated protein folds in the lymphotactin native state. PNAS 105:5057–5062.

57. Rose GD (1979) Hierarchic organization of domains in globular proteins. Journal of Molecular Biology 134:447–470.

58. Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.

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

60. Poduslo JF, Howell KG (2015) Unique molecular signatures of Alzheimer's disease amyloid β peptide mutations and deletion during aggregate/oligomer/fibril formation. Journal of Neuroscience Research 93:410–423.

61. Fraser JS, Bedem H van den, Samelson AJ, Lang PT, Holton JM, Echols N, Alber T (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. PNAS 108:16247–16252.

62. Keedy DA, Kenner LR, Warkentin M, Woldeyes RA, Hopkins JB, Thompson MC, Brewster AS, Van Benschoten AH, Baxter EL, Uervirojnangkoorn M, et al. (2015) Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. Elife 4.

63. Russi S, González A, Kenner LR, Keedy DA, Fraser JS, van den Bedem H (2017) Conformational variation of proteins at room temperature is not dominated by radiation damage. J Synchrotron Radiat 24:73–82.

64. van den Bedem H, Fraser JS (2015) Integrative, dynamic structural biology at atomic resolution--it's about time. Nat. Methods 12:307–318.

65. Maguid S, Fernández-Alberti S, Parisi G, Echave J (2006) Evolutionary Conservation of Protein Backbone Flexibility. J Mol Evol 63:448–457.

66. Pandini A, Mauri G, Bordogna A, Bonati L (2007) Detecting similarities among distant homologous proteins by comparison of domain flexibilities. Protein Eng Des Sel 20:285–299.

67. Davis IW, Arendall III WB, Richardson DC, Richardson JS (2006) The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. Structure 14:265–274.

68. Friedland GD, Lakomek N-A, Griesinger C, Meiler J, Kortemme T (2009) A Correspondence Between Solution-State Dynamics of an Individual Protein and the Sequence and Conformational Diversity of its Family. PLoS Comput Biol 5:e1000393.

69. Marsh JA, Teichmann SA (2014) Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. BioEssays 36:209–218.

70. Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M (2006) Relation between native ensembles and experimental structures of proteins. PNAS 103:10901–10906.

71. Monzon AM, Rohr CO, Fornasari MS, Parisi G (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database (Oxford) 2016.

72. Lange OF, Lakomek N-A, Farès C, Schröder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, Groot BL de (2008) Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. Science 320:1471–1475.

73. F. Ángyán A, Gáspári Z (2013) Ensemble-Based Interpretations of NMR Structural Data to Describe Protein Internal Dynamics. Molecules 18:10548–10567.

74. Lindorff-Larsen K, Ferkinghoff-Borg J (2009) Similarity Measures for Protein Ensembles. PLOS ONE 4:e4203.

75. DiMaio F (2013) Advances in Rosetta structure prediction for difficult molecular-replacement problems. Acta Cryst D 69:2202–2208.

76. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2016) Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. Proteins. Prot Struc Func Bioinf 84:4-14.

77. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP) — round x. Proteins 82:1–6.

78. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, et al. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 77:89–99.

79. Xu D, Zhang J, Roy A, Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79:147–160.

80. Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69:108–117.

81. Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. Proteins 77:100–113.

82. Ovchinnikov S, Kim DE, Wang RY-R, Liu Y, DiMaio F, Baker D (2016) Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins 84:67–75.

83. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. Elife 4:e09248.

84. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. 23:47–55.

85. Stein A, Kortemme T (2013) Improvements to Robotics-Inspired Conformational Sampling in Rosetta. PLoS ONE 8:e63090.

86. Miao Y, Sinko W, Pierce L, Bucher D, Walker RC, McCammon JA (2014) Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. J. Chem. Theory Comput. 10:2677–2689.

87. Song Y, Tyka M, Leaver-Fay A, Thompson J, Baker D (2011) Structure-guided forcefield optimization. Proteins 79:1898–1909.

88. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227.

89. Huang P-S, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D (2014) High thermodynamic stability of parametrically designed helical bundles. Science 346:481–485.

90. Huang P-S, Boyken SE, Baker D (2016) The coming of age of de novo protein design. Nature 537:320–327.

91. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, et al. (2013) Computational Design of Ligand Binding Proteins with High Affinity and Selectivity. Nature 501:212–216.

92. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science 302:1364–1368.

93. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, et al. (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. Science 357:168–175.

94. Nelson KJ, Perkins A, Van Swearingen AED, Hartman S, Brereton AE, Parsonage D, Salsbury FR, Karplus PA, Poole LB (2017) Experimentally Dissecting the Origins of Peroxiredoxin Catalysis. Antioxidants & Redox Signaling [Internet]. Available from: http://online.liebertpub.com/doi/abs/10.1089/ars.2016.6922

95. Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, et al. (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 43:D345-356.

96. Oldfield CJ, Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Annu. Rev. Biochem. 83:553–584.

97. Faller CE, Reilly KA, Hills RD, Guvench O (2013) Peptide Backbone Sampling Convergence with the Adaptive Biasing Force Algorithm. J. Phys. Chem. B 117:518–526.

98. Gunasekaran K, Ramakrishnan C, Balaram P (1996) Disallowed Ramachandran Conformations of Amino Acid Residues in Protein Structures. Journal of Molecular Biology 264:191–198.

99. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. 66:12–21.

100. Grdadolnik J, Mohacek-Grosev V, Baldwin RL, Avbelj F (2011) Populations of the three major backbone conformations in 19 amino acid dipeptides. PNAS 108:1794–1798.

101. Schmidpeter PAM, Koch JR, Schmid FX (2015) Control of protein function by prolyl isomerization. Biochimica et Biophysica Acta (BBA) - General Subjects 1850:1973–1982.

102. Guvench O, Qu C-K, MacKerell AD (2007) Tyr66 acts as a conformational switch in the closed-to-open transition of the SHP-2 N-SH2-domain phosphotyrosine-peptide binding cleft. BMC Structural Biology 7:14.

103. Jamison II FW, Foster TJ, Barker JA, Hills Jr RD, Guvench O (2011) Mechanism of Binding Site Conformational Switching in the CD44–Hyaluronan Protein–Carbohydrate Binding Interaction. J Mol Biol 406:631–647.

104. Kalmankar NV, Ramakrishnan C, Balaram P (2014) Sparsely populated residue conformations in protein structures: Revisiting “experimental” Ramachandran maps. Proteins 82:1101–1112.

105. Richardson JS, Richardson DC, Keedy DA The Plot thickens: more data, more dimensions, more uses. In: Biomolecular Forms and Functions: A Celebration of. ; 2013. pp. 46–61.

106. Lakshminarasimhan M, Madzelan P, Nan R, Milkovic NM, Wilson MA (2010) Evolution of New Enzymatic Function by Structural Modulation of Cysteine Reactivity in Pseudomonas fluorescens Isocyanide Hydratase. J. Biol. Chem. 285:29651–29661.

107. Esposito L, Balasco N, De Simone A, Berisio R, Vitagliano L, Esposito L, Balasco N, De Simone A, Berisio R, Vitagliano L (2013) Interplay between Peptide Bond Geometrical Parameters in Nonglobular Structural Contexts, Interplay between Peptide Bond Geometrical Parameters in Nonglobular Structural Contexts. BioMed Research International, BioMed Research International 2013, 2013:e326914.

108. Improta R, Vitagliano L, Esposito L (2011) Peptide Bond Distortions from Planarity: New Insights from Quantum Mechanical Calculations and Peptide/Protein Crystal Structures. PLoS ONE 6:e24533.

109. Cerutti DS, Freddolino PL, Duke RE, Case DA (2010) Simulations of a Protein Crystal with a High Resolution X-ray Structure: Evaluation of Force Fields and Water Models. J. Phys. Chem. B 114:12811–12824.

110. Muybridge E (1878) The science of the horse's motions. Sci Am 39:241.

111. Tien MZ, Sydykova DK, Meyer AG, Wilke CO (2013) PeptideBuilder: A simple Python library to generate model peptides. PeerJ 1:e80.

112. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696.

113. Case D, Darden T, Cheatham T, Simmerling C, Wang J, Duke R, Luo R, Walker R, Zhang W, Merz K, et al. (2012) AMBER 12. Available from: http://ambermd.org/

114. Ogihara NL, Ghirlanda G, Bryson JW, Gingery M, DeGrado WF, Eisenberg D (2001) Design of three-dimensional domain-swapped dimers and fibrous oligomers. PNAS 98:1404–1409.

115. Herrou J, Crosson S (2013) myo-inositol and D-ribose ligand discrimination in an ABC periplasmic binding protein. J. Bacteriol. 195:2379–2388.

116. Dauter Z, Lamzin VS, Wilson KS (1995) Proteins at atomic resolution. Current Opinion in Structural Biology 5:784–790.

117. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J. 275:1–21.

118. Dauter Z, Lamzin VS, Wilson KS (1997) The benefits of atomic resolution. Current Opinion in Structural Biology 7:681–688.

119. Cruickshank DWJ (1999) Remarks about protein structure precision. Acta Crystallographica Section D Biological Crystallography 55:583–601.

120. Wilson K, Butterworth S, Dauter Z, Lamzin V, Walsh M, Wodak S, Pontius J, Richelle J, Vaguine A, Sander C, et al. (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. Journal of molecular biology 276:417–436.

121. Fuhrmann CN, Kelch BA, Ota N, Agard DA (2004) The 0.83 Å resolution crystal structure of α-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. Journal of molecular biology 338:999–1013.

122. Lüdtke S, Neumann P, Erixon KM, Leeper F, Kluger R, Ficner R, Tittmann K (2013) Sub-ångström-resolution crystallography reveals physical distortions that enhance reactivity of a covalent enzymatic intermediate. Nature chemistry 5:762–767.

123. Berkholz DS, Faber HR, Savvides SN, Karplus PA (2008) Catalytic cycle of human glutathione reductase near 1 \AA resolution. Journal of molecular biology 382:371–384.

124. Guérin DM, Lascombe M-B, Costabel M, Souchon H, Lamzin V, Béguin P, Alzari PM (2002) Atomic (0.94 Å) resolution structure of an inverting glycosidase in complex with substrate. Journal of molecular biology 316:1061–1069.

125. Getzoff ED, Gutwin KN, Genick UK (2003) Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation. Nature Structural & Molecular Biology 10:663–668.

126. Brereton AE, Karplus PA (2015) Native proteins trap high-energy transit conformations. Science Advances 1.

127. MacArthur MW, Thornton JM (1996) Deviations from Planarity of the Peptide Bond in Peptides and Proteins. Journal of Molecular Biology 264:1180–1195.

128. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG (2007) Localizing frustration in native proteins and protein assemblies. PNAS 104:19819–19824.

129. Gianni S, Camilloni C, Giri R, Toto A, Bonetti D, Morrone A, Sormanni P, Brunori M, Vendruscolo M (2014) Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein. PNAS 111:14141–14146.

130. Sutto L, Lätzer J, Hegler JA, Ferreiro DU, Wolynes PG (2007) Consequences of localized frustration for the folding mechanism of the IM7 protein. PNAS 104:19825–19830.

131. Edison AS (2001) Linus Pauling and the planar peptide bond. Nat Struct Mol Biol 8:201–202.

132. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2013) Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. FEBS J 280:5705–5736.

133. Dauter Z, Wlodawer A, Minor W, Jaskolski M, Rupp B (2014) Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. IUCrJ 1:179–193.

134. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr 68:352–367.

135. Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D Biol. Crystallogr. 58:1948–1954.

136. Kumar KSD, Gurusaran M, Satheesh SN, Radha P, Pavithra S, Thulaa Tharshan KPS, Helliwell JR, Sekar K (2015) Online_DPI: a web server to calculate the diffraction precision index for a protein structure. Journal of Applied Crystallography 48:939–942.

137. Tufte ER, Graves-Morris P The visual display of quantitative information. Graphics press Cheshire, CT; 1983.

138. DePristo MA, de Bakker PIW, Blundell TL (2004) Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. Structure 12:831–838.

139. Ringe D, Petsko GA [19]Study of protein dynamics by X-ray diffraction. In: Enzymology B-M in, editor. Vol. 131. Enzyme Structure Part L. Academic Press; 1986. pp. 389–433. Available from: http://www.sciencedirect.com/science/article/pii/0076687986310504

140. Wilson MA, Brunger AT (2000) The 1.0 Å crystal structure of Ca2+-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity1. Journal of Molecular Biology 301:1237–1256.

141. Clark SA, Tronrud DE, Andrew Karplus P (2015) Residue-level global and local ensemble-ensemble comparisons of protein domains. Protein Science 24:1528–1542.

142. Elber R, Karplus M (1987) Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. Science 235:318–321.

143. Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Is one solution good enough? Nat Struct Mol Biol 13:184–185.

144. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132.

145. Palopoli N, Monzon AM, Parisi G, Fornasari MS (2016) Addressing the Role of Conformational Diversity in Protein Structure Prediction. PLOS ONE 11:e0154923.

146. Tiberti M, Papaleo E, Bengtsen T, Boomsma W, Lindorff-Larsen K (2015) ENCORE: Software for Quantitative Ensemble Comparison. PLoS Comput. Biol. 11:e1004415.

147. Laurents D, Pérez-Cañadillas JM, Santoro J, Rico M, Schell D, Pace CN, Bruix M (2001) Solution structure and dynamics of ribonuclease Sa. Proteins 44:200–211.

148. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

149. Sharaf NG, Brereton AE, Byeon I-JL, Andrew Karplus P, Gronenborn AM (2016) NMR structure of the HIV-1 reverse transcriptase thumb subdomain. J. Biomol. NMR.

150. Brereton AE (2016) atomoton/ensemblator. GitHub [Internet]. Available from: https://github.com/atomoton/ensemblator

151. Spizman L, Weinstein MA (2008) A note on utilizing the geometric mean: when, why and how the forensic economist should employ the geometric mean. J. Legal Econ. 15:43.

152. Cauchy A-L (1821) Cours d'analyse de l'École royale polytechnique. Premiere partie. Analyse algébrique. Gallica-Math, ØEuvres completes sér 2.

153. Ronan T, Qi Z, Naegle KM (2016) Avoiding common pitfalls when clustering biological data. Sci. Signal. 9:re6-re6.

154. Frey BJ, Dueck D (2007) Clustering by Passing Messages Between Data Points. Science 315:972–976.

155. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830.

156. Jones E, Oliphant T, Peterson P, others SciPy: Open source scientific tools for Python. 2001. Available from: http://www.scipy.org/

157. Fred A, Jain AK Evidence Accumulation Clustering Based on the K-Means Algorithm. In: Caelli T, Amin A, Duin RPW, Ridder D de, Kamel M, editors. Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2002. pp. 442–451. Available from: http://link.springer.com/chapter/10.1007/3-540-70659-3_46

158. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65.

159. Tyka MD, Keedy DA, André I, DiMaio F, Song Y, Richardson DC, Richardson JS, Baker D (2011) Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. Journal of Molecular Biology 405:607–618.

160. Sich C, Improta S, Cowley DJ, Guenet C, Merly JP, Teufel M, Saudek V (2000) Solution structure of a neurotrophic ligand bound to FKBP12 and its effects on protein dynamics. Eur. J. Biochem. 267:5342–5355.

161. Michnick SW, Rosen MK, Wandless TJ, Karplus M, Schreiber SL (1991) Solution structure of FKBP, a rotamase enzyme and receptor for FK506 and rapamycin. Science 252:836–839.

162. Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with rosetta. Proteins 55:656–677.

163. Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB, Bax A (1992) Solution structure of a calmodulin-target peptide complex by multidimensional NMR. Science 256:632–638.

164. Xu B, Chelikani P, Bhullar RP (2012) Characterization and Functional Analysis of the Calmodulin-Binding Domain of Rac1 GTPase. PLOS ONE 7:e42975.

165. Song J-G, Kostan J, Drepper F, Knapp B, de Almeida Ribeiro E, Konarev PV, Grishkovskaya I, Wiche G, Gregor M, Svergun DI, et al. (2015) Structural insights into Ca2+-calmodulin regulation of Plectin 1a-integrin β4 interaction in hemidesmosomes. Structure 23:558–570.

166. Bottaro S, Gil-Ley A, Bussi G (2016) RNA folding pathways in stop motion. Nucleic Acids Res 44:5883–5891.

167. Gil-Ley A, Bottaro S, Bussi G (2016) Empirical Corrections to the Amber RNA Force Field with Target Metadynamics. J. Chem. Theory Comput. 12:2790–2798.

168. Matthews BW (2016) How planar are planar peptide bonds? Protein Science 25:776–777.

169. Panasik N, Fleming PJ, Rose GD (2005) Hydrogen-bonded turns in proteins: The case for a recount. Protein Science 14:2910–2914.

170. Rosenblueth A, Wiener N (1945) The role of models in science. Philosophy of science 12:316–321.

171. Rodnina MV (2016) The ribosome in action: Tuning of translational efficiency and protein folding. Protein Science 25:1390–1406.

172. Box GE (1979) Robustness in the strategy of scientific model building. Robustness in statistics 1:201–236.