NEXT-GENERATION ANTIBODY MODELING

by

Brian D. Weitzner

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2015

ABSTRACT

Antibodies are important immunological molecules that can bind a diverse array of foreign molecules. The genetic mechanism that gives rise to antibodies and many antibody sequences is known, but only by studying three-dimensional structures of antibodies and antibody–antigen complexes can we reveal immunological mechanisms and provide a starting point for developing rationally designed antibodies. With the advent of high-throughput sequencing technologies, the gap between the number of sequences and structures is widening, demanding accurate antibody modeling methods. Our previously developed method, RosettaAntibody, served as a starting point for antibody structure prediction. In this dissertation, I detail my work assessing the predictive power of RosettaAntibody, and the development and testing of new methods to address its weaknesses. First, I describe an effort to assess the accuracy of RosettaAntibody on a set of unpublished crystal structures. This challenge enabled us to combine manual and automated methods for selecting models and compare RosettaAntibody to other antibody modeling methods. The most challenging aspect of structure prediction in this assessment proved to be

modeling the third complementarity determining region loop on the heavy chain (CDR H3). Next I detail my work in studying CDR H3 loops to uncover why a vast majority of them contain a kink at the loop's C-terminus. Part of this work involved searching the Protein Data Bank (PDB) for structures with a similar geometry of the amino acid residues at the base of the loop, leading to a set of CDR H3-like loops from non-antibody proteins. With a clearer understanding of CDR H3 loop structures and the most detailed description of the kink to date, I developed a new loop modeling routine that utilizes this information to restrict the geometry of the loop to be kinked, resulting in an improvement in the weakest aspect of antibody structure prediction. In summary, the structure prediction methods I have developed and structural analyses I have performed provide a means to begin to address the widening sequence–structure gap. Additionally, these methods can be used to perform structural analysis in the development of rationally designed antibodies.

Advisor: Prof. Jeffrey J. Gray (Chemical & Biomolecular Engineering)

Reader: Prof. Marc D. Donohue (Chemical & Biomolecular Engineering)

Reader: Prof. Roland L. Dunbrack, Jr (Institute for Cancer Research, Fox Chase Cancer Center)

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Jeff Gray. I first met Jeff in 2007 while I was still an undergraduate and was struggling to perform a few simple simulations. He invited me to come to Baltimore for a week in the summer to spend some time learning from him and his students. While I was interacting with his students, I saw the environment Jeff had cultivated and was impressed with his students' confidence, curiosity and knowledge. Near the end of my week in Baltimore, I asked Jeff why some seemingly important task was impossible to perform in Rosetta. I remember his response very clearly: "Why don't you come here for graduate school and make it possible?"

When I ended up at Johns Hopkins in 2009 and joined Jeff's lab, I asked him what project I would be working on. He said, "it's your thesis—what do you think is interesting?" This kind of freedom is extremely rare and it inspired me to think critically very early on in my doctoral studies. Jeff has struck a balance between critical but nurturing, focused but free and successful both at work and at home. I am truly honored to have been a part of his group. Thanks, Jeff.

Any acknowledgements I could make would be incomplete without promi-

nently featuring Roland Dunbrack. Roland invited me to join his lab while I was in high school as part of a student scientist program at Fox Chase Cancer Center. He was the first person to introduce me to computational research, programming and protein structure science. He was also interested in my opinions on current events including sociopolitical topics and was comfortable engaging me in conservations over tea. Roland continued to mentor me throughout college and he made sure that my work in his group would result in a publication. When I started at Johns Hopkins, I realized there was some overlap in some of the problems we were trying to solve and I felt conflicted about using the ideas that I had gotten while working in his lab. I discussed this with him and he told me that it is quite difficult to "unknow" something and we ended up collaborating. Our collaboration resulted in countless hours spent at Ultimo discussing work and generally having a good time. Like Jeff, Roland has become much more than a mentor to me and I count him among my close friends.

I would like to thank Professors Dilipkumar Asthagiri, Rachel Karchin, Rebecca Schulman, Marc Ostermeier, Marc Donohue, Ingo Ruczinski, Michael Bevan and Elijah Roberts for being on my thesis committee. Not only did their feedback help me understand my own research better, they kept me focused on the task at hand.

My work would have been impossible if not the support from the National Institutes of Health. In particular, I acknowledge grants R01 GM84453 to Roland

Dunbrack and R01 GM078221 to Jeffrey Gray for supporting my work.

I am grateful to the members of the Gray lab, past and present, who have helped make a productive, fun and respectful workplace. Aroop Sircar's work on the original RosettaAntibody and the development of camelid antibody modeling code in Rosetta 3.0 is greatly appreciated as it served as a foundation for my work. I'd like to thank Daisuke Kuroda for sharing his knowledge of antibody structures and literature. Jason Labonte's ability to jump into computational research and his desire to develop high-quality code is truly inspirational. Beyond that, he serves as the Dungeon Master for our lab-wide Dungeons & Dragons campaign.

Soon after I joined the lab, four students graduated, leaving Krishna Kilambi and myself as the senior lab members. Krishna was hard at work and helped to show me that we were going to pull through and develop a mastery of our methods. In the same vein, the members of the Rosetta Commons have collectively invested countless hours of support and I would like to especially thank Andrew Leaver-Fay, Matthew O'Meara and Samuel DeLuca not only for their help and expertise, but for their friendship over the years.

Sergey Lyskov and I have had an extremely productive working relationship. Simple musings about user interfaces or help messages in PyRosetta have turned into complete ideas that have enabled others to learn to use Rosetta. Sergey is always excited to discuss philosophy, technology or to recommend a good book.

My friends have helped me keep sanity and I am thankful for all of them. I

am particularly thankful for “emergency” homebrew club meetings on Saturdays. When I decided to train for a marathon, Daniel Beltrán became my coach/spirit guide and Michael Pacella quickly became my running buddy. Without them, I would have collapsed after 10 miles.

I have had the great privilege of counting Gregg Duncan among my friends. Whether it was driving to DC on a Wednesday to catch a show at Black Cat, brewing an elaborate beer, finishing an enormous batch of chili or just needing to hang out, Gregg was not only there, but ready to have a good time and laugh hard.

My parents, Bruce and Deborah Weitzner, have always supported and encouraged me. I am continually amazed when I consider how much time, effort and love they have given to me over the years and that I am but one of their five children. Their dedication as parents is unrivaled and their example of selflessness is a challenge I can only hope to meet.

To my parents, Bruce and Deborah, who made me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Introduction to antibodies

Immunoglobulin G (IgG) proteins, commonly referred to as antibodies, are the major molecule of the vertebrate adaptive immune system. Antibodies bind to specific regions of nearly any foreign molecule and, once bound, promote other cells to degranulate or phagocytose the pathogen. In addition to their natural proclivities, antibodies have proven to be a robust model for biotechnological and pharmaceutical applications. One of the most successful antibody therapeutics, trastuzumab, targets HER2-positive breast tumors and has been shown to increase overall survival and disease free survival rates.[1,2] Nearly 100,000 women receive treatment every year, and, as a testament to its success, trastuzumab generated a revenue of $6.8 billion in FY2013.[3] Structure-based design of novel antibodies has been used to develop molecules that can serve as biosensors,[4] and custom antibodies have even become a part of routine biochemical assays (ELISA).[5,6]

In order to have the capability to bind to nearly any infectious molecule, a diverse population of antibodies is required. This diversity arises from a number of complex processes beginning in the bone marrow and ending in lymphatic tissue.[7] In the bone marrow,

hematopoietic stem cells differentiate into one of three primary blood cell types: red blood cells, white blood cells and platelets. White blood cells (leukocytes) further differentiate primarily into T cells and B cells, the latter of which ultimately express antibodies. During differentiation into a B cell, a genetic shuffling event called V(D)J recombination occurs wherein one of several Variable, Diversity and Joining gene segments are joined to form a new immunoglobulin gene. The gene segments are selected randomly and junctional diversification occurs, in which nucleotide additions and deletions are made at the interface of the segments. There are several checkpoints throughout this process to ensure that the newly formed gene produces a protein that folds properly and that the resulting antibody does not bind strongly to "self" proteins. A similar process occurs in T cells to form T cell receptors (TCRs), which are the T cell analogues of immunoglobulins. After successful V(D)J recombination, the protein is expressed on the surface of the B cell as immunoglobulin M (IgM), and the fully differentiated B cell, called a naïve B cell, leaves the bone marrow. The IgM population on naïve B cells is referred to as the naïve antibody repertoire. The size of the theoretical naïve human antibody repertoire is estimated to be $> 10^{14}$. For reference, there are roughly 10^{12} B cells in a human.[8]

The other major source of antibody diversity occurs after the host organism is exposed to a pathogen in a process called affinity maturation. Major histocompatibility complex II (MHC II) molecules form complexes with linear pieces of foreign molecules (peptide epitopes) after they have been processed by antigen presenting cells (APCs), which go on to activate T cells. Separately, a B cell receptor interacts with internalizes and process the same antigen. The B cell presents peptide epitopes on its surface and can then be

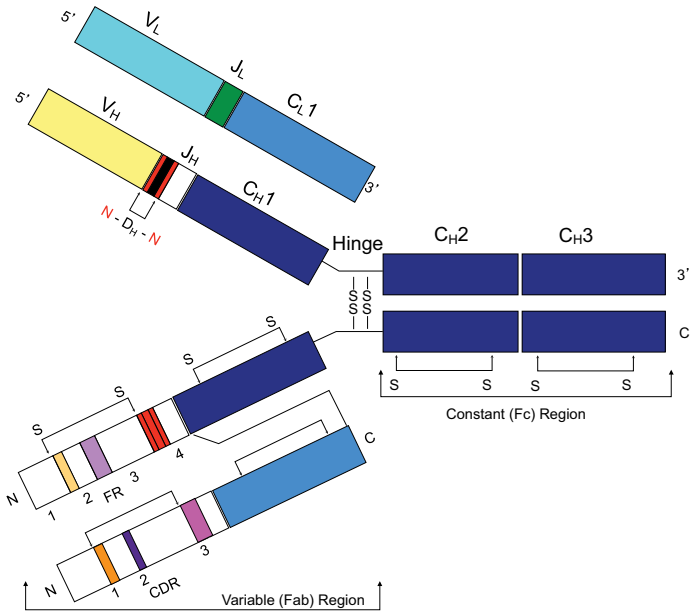

Figure 1.1: Schematic of IgG structure. In the top chains, domains encoded from germline V, D, J and C segments are indicated. Nontemplated N-nucleotides are shown in red. These top chains delineate the 5′ to 3′ genetic composition of the antibody. In the bottom chains, framework (FR) and complementarity-determining regions (CDRs) are indicated. These bottom chains delineate the N-terminal to C-terminal protein sequence. Dashed lines denote disulfide bonds. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology Georgiou *et al.* "The promise and challenge of high-throughput sequencing of the antibody repertoire",[9] copyright 2014

stimulated by a T cell activated by the same epitope. Once this occurs, the B cell is activated and it undergoes class switching to change the immunoglobulin molecule it is producing from IgM to IgG and then begins a process called somatic hypermutation in which the B cell divides rapidly with a high mutation rate in its complementarity determining regions (CDRs). As this process repeats, only the daughter cells that improve affinity survive, leading to clonal selection of stronger, more specific binders. At this point the B cell can form either a plasma B cell, which produces a large amount of antibodies to combat the infection, or a memory B cell, which can be reactivated in the event of a secondary infection.

The foundation of antibodies' utility lies in their three-dimensional structure. As

shown in Figure 1.1, antibodies consist of two sets of heavy and light chains arranged into a "Y" shape, with the four polypeptide chains joined by disulfide linkages. The heavy chain contains four domains, three adjacent constant domains (C_H1, C_H2, C_H3) and one variable domain (V_H), and the light chain consists of a single constant domain (C_L) and a variable domain (V_L). The C_H1 and V_H domains interact with the C_L and V_L domains to form the antigen-binding fragment (F_{ab}) to form the "arms" of the Y. Within the F_{ab}, both variable domains are directed away from the remaining heavy chain constant domains and make up the variable fragment (F_V). At the tip of the F_V are three complementarity determining region (CDR) loops on each chain (CDR L1–3 and CDR H1–3) that form the region of the antibody, called the paratope, that recognizes its target.[10,11] Figure 1.2 depicts the structure of an F_V with the CDR loops shown in different colors. The F_V's sequence, including the CDR loops, is determined by the genetic recombination events and affinity maturation. Thus, the primary functional difference among antibodies is the conformation and chemical identity of the CDR loops.[12]

Next-generation sequencing techniques have recently been developed to enable rapid determination of large numbers of antibody sequences.[9,13,14] Coupling these techniques with library-based screening or isolation of IgGs from whole blood samples can yield methods to identify high-affinity binders to a desired antigen. However, no information about the targeted region of the antigen (the epitope) can be gleaned from these processes. This level of detail is required in order to design therapeutic antibodies or to design vaccines that are mimetics of extremely infectious antigens. In order to consider specific antibody–antigen interactions, structural models are required.

1.2 Overview of antibody homology modeling

Structural models of proteins can be elucidated using data from one or more experimental techniques. The most commonly used methods are single-crystal X-ray diffraction, nuclear magnetic resonance (NMR) spectroscopy, and neutron diffraction. X-ray diffraction can yield high-resolution structural models of molecules of any size. However, obtaining high-quality crystals of proteins is often challenging. Producing crystals of sufficient quality is only the first of many challenges, as interpreting diffraction data can be a time consuming process and may even require the use of a synchrotron to tune the wavelength of the diffracting X-rays. NMR spectroscopy experiments are performed on proteins in solution, bypassing the challenges associated with crystallization. In order to produce a usable signal, some heavy atoms in the protein need to be isotopically labelled (typically ^{13}C and ^{15}N). The measured signals are then converted into various real-space restraints, which are then represented as energies and used in a molecular mechanics modeling package to produce a set of structures that collectively explain the restraints. However, the measured signals degrade as the size of the protein being studied increases, which currently makes NMR unsuitable for generating high-resolution models of large proteins. Protein structures can also be elucidated via neutron diffraction experiments, which can be thought of as direct observation of nuclear coordinates because neutrons are diffracted by atomic nuclei and not electrons. Because nuclei are small compared to electron density, neutron diffraction experiments require much larger crystals than those for X-ray diffraction experiments. Additionally, a nuclear reactor is needed as a source of neutrons. Single-crystal X-ray

diffraction, NMR spectroscopy, and neutron diffraction have proven extremely useful and have been used to generate the more than 100,000 protein structural models in the Protein Data Bank (PDB). However not all protein structures can be determined with these methods, and limited resources make it impossible to determine the structures of all of the sequences identified in high-throughput sequencing experiments.

To bridge the widening sequence–structure gap, one must turn to computational structure prediction methods. Homology modeling uses parts of structural models of related proteins that are predicted to be conserved in the target protein. Because the non-CDR loop regions of antibodies (framework regions) are structurally conserved, antibodies are highly amenable to homology modeling methods. There are nearly 2,000 antibody structures in the PDB[15] that can be used as templates for homology modeling. Analysis of these structures has revealed that five of the six CDR loops (CDR L1–3, H1, H2) adopt a limited number of distinct structures, referred to as canonical loop conformations.[16–19] The canonical conformation of a particular CDR loop can be readily identified from its length and sequence. Antibody homology modeling methods must identify the best templates for the heavy and light chain framework regions and canonical CDR loops and assemble them to form a reasonable structural model.

The remaining piece is modeling the CDR H3 loop. Because it is the only one of the six CDR loops that does not adopt canonical conformations, CDR H3 must be modeled *de novo*. Unsurprisingly, CDR H3 modeling is the most challenging aspect of antibody structure prediction.

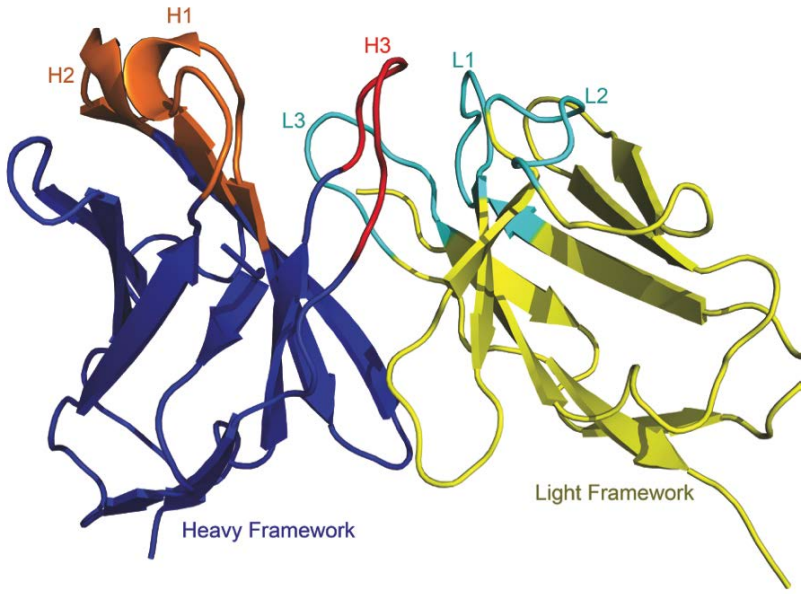

Figure 1.2: Variable (F_V) region of the anti-lysozyme antibody (1BQL[20]). The heavy chain variable region (V_H) comprises of the heavy framework (blue), the canonical complementarity determining regions (CDR) H1 (orange), H2 (orange) and the hypervariable CDR H3 (red). The light chain variable region (V_L) comprises of the light framework (yellow) and canonical CDRs L1, L2 and L3 (cyan). From Sircar A "Computational antibody structure prediction and antibody–antigen docking".[21] Reprinted with permission from Dr. Aroop Sircar.

1.3 CDR H3

Because the CDR H3 loop is the most structurally diverse region of antibodies, it is the most challenging to predict. For example, models generated by RosettaAntibody have median RMSDs of < 1.0 Å for the five non-H3 CDR loops, the median RMSD of H3 loops ranges from 1.6 Å for very short loops (4–6 residues) to 6.0 Å for very long loops (17–22).[22] In 2011, the first antibody modeling assessment (AMA) blindly tested several antibody structure prediction methods on the same set of F_V sequences and found that the other methods tell a similar story.[23]

Understanding the structure of CDR H3 has been the focus of several studies,

many of which focus on identifying structural features that are conserved across a diverse set of CDR H3 structures. Although CDR H3 loops do not adopt canonical conformations, roughly 80% do share a common structural feature: a C-terminal kink.[19,24–29] The ability to predict the presence of this kink allows restriction of some of the loop's degrees of freedom, providing a better starting point for *de novo* structure prediction. There has been progress in developing a set of rules to predict the kink from the loop sequence,[24–26] but these rules have broken down as more structures have been determined by experimental methods.[19] No replacements have been postulated thus far.

The inability to reliably generate accurate CDR H3 models is problematic because CDR H3 often plays a critical role in antigen binding.[10,30] Structural models of antibody–antigen complexes can reveal immunological mechanisms and empower protein engineers with the requisite information to propose rational mutations for a specific application.[4] However, predicting antibody–antigen complexes requires accurate input models.[31] Thus, improving CDR H3 structure prediction is necessary for the continued advancement of computational antibody structure prediction and design applications.

1.4 Molecular modeling with Rosetta

The Rosetta software suite[32] is a robust biomolecule structure prediction and design toolset. There are two main challenges in structure prediction: scoring and sampling. Scores, which are typically thought of as energies (*i.e.* lower is better), are computed from trial conformations in order to rank them. In Rosetta, a score function is a combination of physics-based and statistical potentials with a specific weights assigned to each term. The

weights and the terms themselves can vary throughout the simulation to enable novel score functions to focus on key interactions.

Sampling is the process by which trial conformations are generated. As articulated by the Levinthal Paradox,[33] the conformational space of even a small protein is too large to search exhaustively; sampling strategies must be tailored to rapidly search large regions of space and focus on areas that appear promising. To accomplish this, Rosetta simulations use both multi-scale modeling and Monte Carlo-plus-minimization for optimization and search.[34]

Multi-scale modeling in Rosetta has two stages: (1) a low-resolution mode that represents side chains with a single pseudo-atom positioned at the centroid of the side-chain atomic coordinates; and (2) an all-atom representation. Large conformational changes are made using the low-resolution representation. and the all-atom representation is used for refinement and more detailed scoring. The combination of these representations enables Rosetta algorithms to spend a larger fraction of their time evaluating favorable conformations.

Monte Carlo-plus-minimization optimization combines discrete jumps in conformation-space using library-based approaches or by explicitly sampling specific degrees of freedom with continuous gradient-based minimization. After a candidate conformation is sampled and minimized, its score is used to determine if the new conformation is an improvement. If the new conformation scores better than the previous conformation, it is accepted and the simulation continues with it. If the score is worse, it is accepted by the Metropolis criterion.[35] By combining minimization with Monte Carlo sampling, Rosetta

only needs to compare structures that are in local minima, which further focuses sampling towards favorable conformations.

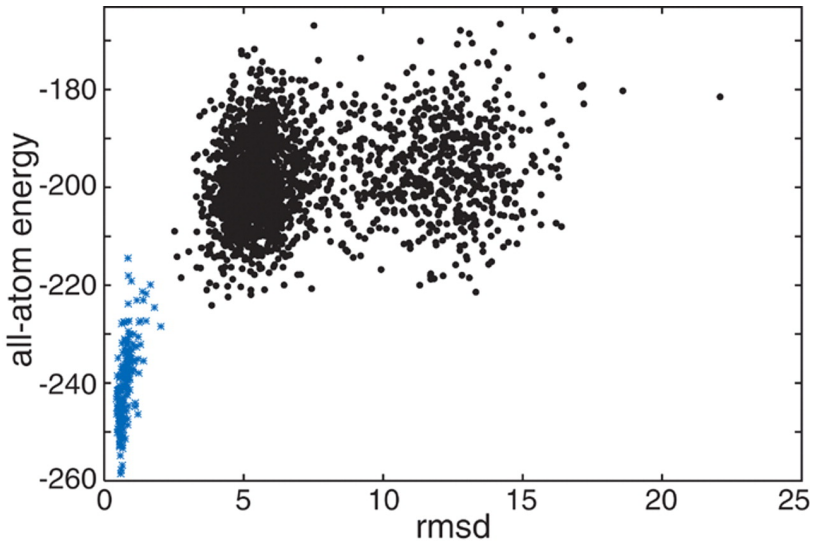

Figure 1.3: Free-energy landscape for the small protein barstar (PDB code 1A19[36]). Rosetta all-atom energy (y axis) is plotted against C_α-RMSD (x axis) for models generated by simulations starting from the native structure (refined natives, blue points) or from an extended chain (*de novo* models, black points). The free-energy function includes the entropic contribution to the solvation free energy but not the configurational entropy. From Bradley *et al.* "Toward High-Resolution de Novo Structure Prediction for Small Proteins", Science.[37] Reprinted with permission from AAAS.

In order to assess their performance, many Rosetta-based methods are tested on a set of proteins of known structure. For successful predictions, a plot of Rosetta score vs. root-mean-squared-deviation (RMSD) of the coordinates of the model from those of the native structure will show a steep drop in score as the RMSD becomes small. Figure 1.3 shows an example of a successful simulation. These plots are referred to as "funnel plots" as one can imagine a funnel-like shape in the highly dimensional conformational search space that drives the structure to the global energy minimum.

1.5 Organization of dissertation

Carl Sagan once said, "If you wish to make an apple pie from scratch, you must first invent the universe."[38] In this case, the "apple pie" is a new tool for antibody structure prediction and the "universe" that needs to be invented consists of the software tools and the skills required to design and implement highly performant code capable of addressing scientific inquiries. In Chapter 2 I summarize several projects that I have pursued, including re-architecting the libraries that constitute Rosetta, developing new frameworks to enable modeling new classes of proteins or to expressively control which degrees of freedom are accessible by a protocol, designing a modern implementation of a key Rosetta application, and creating a new visualization method. These projects have given me the necessary skills to make that apple pie.

The remainder of the dissertation focuses on recent efforts to improve Rosetta-based antibody structure prediction methods. Chapter 3 (previously published[39]) presents the results of my participation in antibody modeling assessment II (AMA II), in which 11 antibody F_Vs were predicted from sequence. All of the antibodies in AMA II had unpublished high quality crystal structures that were used to measure the accuracy of the structure prediction method. A total of seven groups participated in this challenge and a summary of the results can be found in Almagro *et al.*'s paper.[40] By comparing the results of all of the participants, the strengths and weaknesses that are present across all of the methods become apparent. In this case, accurately modeling CDR H3 stood out as a remaining challenge in the field.

Chapter 4 (previously published[41]) reports the results of a structural analysis of CDR H3 loops conducted using Rosetta. In this chapter, CDR H3 loops are compared to loops from non-antibody proteins in order to assess the frequency of "H3-like" loops and to gain insight into what factors contribute to adopting these conformations. A new hypothesis of the structural underpinnings of the observed diversity of CDR H3 is ultimately developed.

In Chapter 5, the findings from Chapters 3 and 4 are used to guide the development of a new CDR H3 structure prediction method. The performance of this method is assessed by generating models of CDR H3 loops both on the crystal frameworks and homology models of known structures.

Finally, in Chapter 6, I summarize my contributions to the Rosetta project and comment on what developments are needed for continued progress toward accurately predicting antibody structures in atomic detail.

CHAPTER II

THE DEVELOPMENT OF COMPUTATIONAL TOOLS

2.1 Introduction

In computational research, if there is not a good tool for a particular task, one can build it. While this is advantageous, it introduces an additional variable that needs to be considered: the design and engineering of the tools themselves. Design decisions can impact the stability and robustness of the tool, the ease of development of new tools in the future, and even code compilation times. In this chapter I detail my involvement in several projects focused on developing computational tools.

2.2 XRW: the eXtreme Rosetta Workshop

The Rosetta source code contains nearly 2 million lines of code and is changing constantly—there are over 1,000 active branches and over 100 revisions made every week. Unsurprisingly, compiling Rosetta can take a very long time, which can make developing new tools or fixing bugs tedious. A large part of the compile time issue could be attributed to the compiler processing dependencies introduced by `#include` directives throughout the

source code.

`#include` directives are used to instruct the compiler's preprocessor to insert the contents of another file into the current file during compilation, thus enabling complicated programs to split into many smaller files. Problems can arise, however, when the total number of inclusions is large for each file or if there are cycles in the dependency graph.

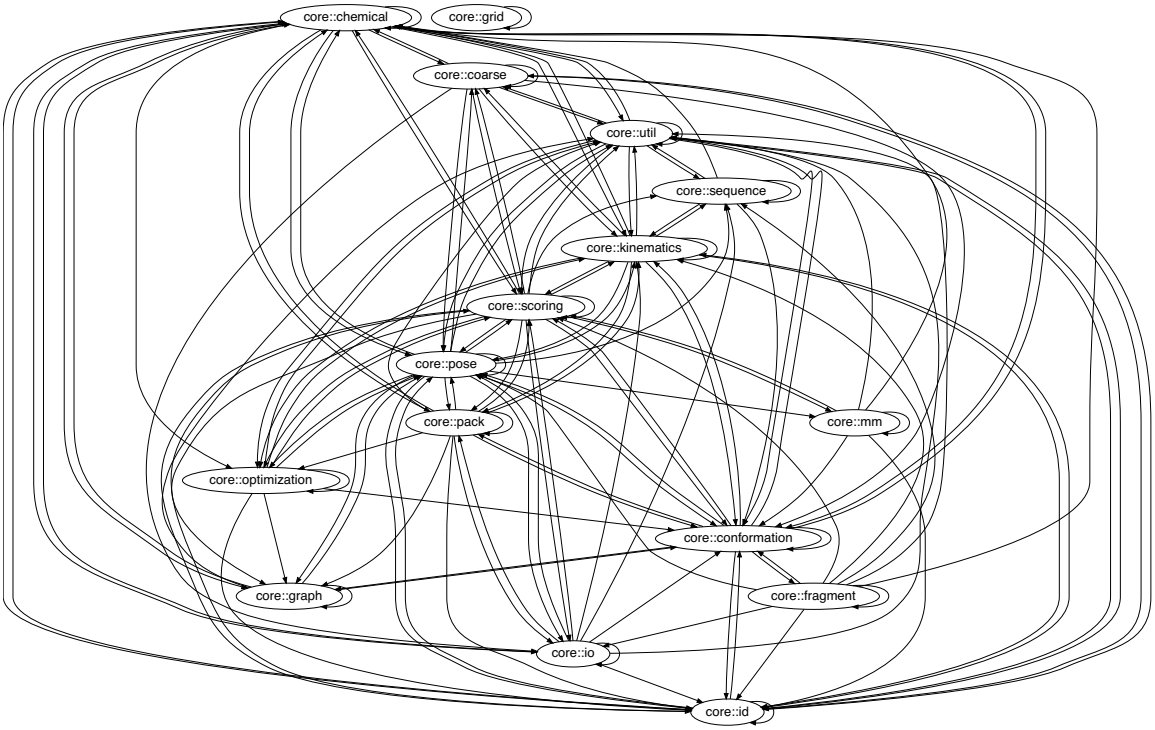

Figure 2.1: The dependency graph for Rosetta's `core` library. The nodes are namespaces and are depicted as ovals. Edges are directed arrows pointing from a dependent namespace to the namespace upon which it depends. This graph illustrates the complexity of the `core` library, includes many cycles.

Figure 2.1 shows the dependency graph for Rosetta's `core` library as it existed in the Fall of 2010, which contains the code responsible for representing protein structures, scoring functions and side-chain optimization. Because of the complexity of this graph, the Rosetta community decided to attempt to reorganize this library into a set of smaller

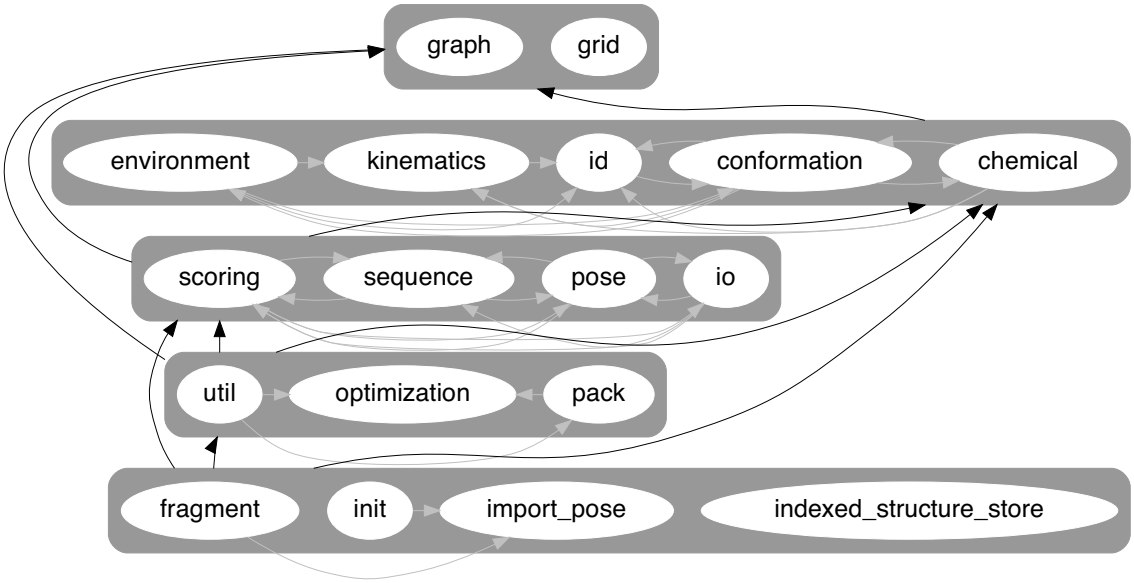

Figure 2.2: The dependency graph for Rosetta's `core.1`(top)–`core.5`(bottom) libraries. The libraries are depicted as gray shaded boxes with the namespaces contained within are shown as white ovals. Namespace dependencies within the same library are explicitly represented with light gray arrows, while interlibrary dependencies are shown with black arrows connecting the libraries. Note the black arrows only point up, *i.e.* toward lower numbered `core` libraries.

libraries. A team of eight individuals was assembled to take on this task in what was deemed the eXtreme Rosetta Workshop (XRW). With the community's support, we spent a full week splitting the `core` library into five libraries. Figure 2.2 shows the dependencies of the newly formed `core` libraries. The resultant libraries are arranged in a directed acyclic graph (DAG), which allows code within that library to include anything in the same library or lower, but not higher. This simple restriction led to a dramatic speed up in compilation time, as well as reduced memory requirements for compiling. The latter effect has been shown to be extremely important for building Rosetta on some high-performance computing (HPC) resources with small amounts of memory dedicated to each processor.

A second XRW resulted in splitting the `protocols` library, which resulted in further

improvements to compilation time and memory usage. The XRWs led to a more modular codebase that encouraged the community to adopt rapid build–test–debug workflows, which in turn has contributed to increased stability and performance of Rosetta.

2.3 Designing the RosettaDock code

The text in this section is primarily taken from the methods section of Chaudhury *et al.* "Benchmarking and analysis of protein docking performance in Rosetta v3.2",[42] with permission under the terms of the Creative Commons Attribution License.

When Rosetta was migrated to an object-oriented implementation,[32] existing applications needed to be ported, which provided an opportunity to reconsider underlying assumptions about how the software would be used. One such application was RosettaDock. At the time, there were already several flavors of RosettaDock: (1) RosettaDock,[43] used for rigid body docking; (2) EnsembleDock,[31] which incorporates backbone flexibility via an ensemble of backbone conformations; and (3) SnugDock,[44] which explicitly samples the conformational degrees of freedom in antibody–antigen complexes. Based on the current applications, we settled on two major goals for a new implementation: first, to allow for easier use of built-in Rosetta functionality, such as constraints or ligand modeling; and second, to give developers greater flexibility when developing their own protocols that use docking functions. My role in this study was to design and implement the various classes that would enable current and future applications to reuse the individual components. Figure 2.3 diagrams the structure of the major classes associated with docking. Docking has been split into three major classes: `DockingProtocol`, `DockingLowRes` and `DockingHighRes`. DockingProtocol is responsible for handling user-specified docking-options, appropriately configuring various objected associated with docking, and applying `DockingLowRes` and `DockingHighRes` objects.

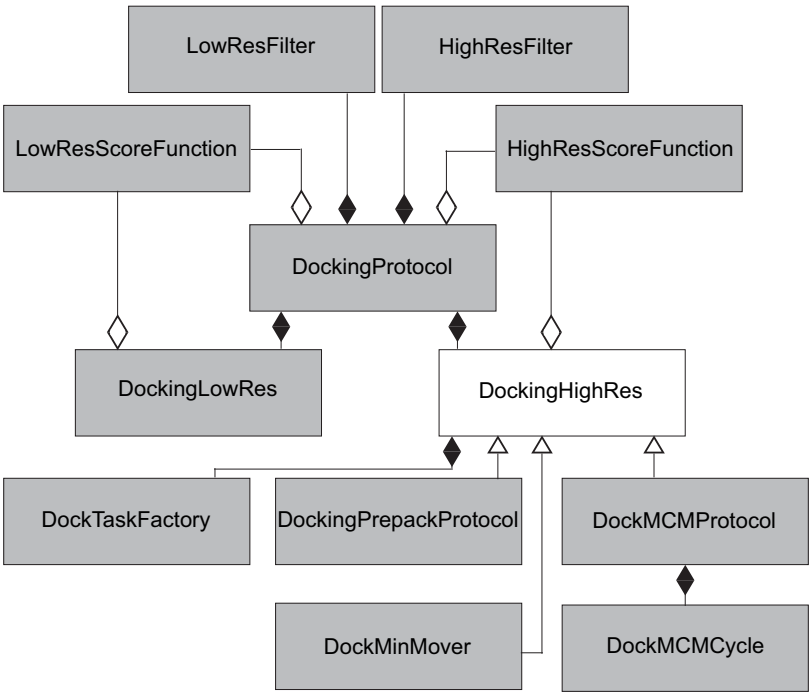

Figure 2.3: A shaded diamond indicates composition (the object the diamond points towards is responsible for the lifecycle of the other object); an open diamond indicates aggregation (the object the diamond points towards has an instance of the other object but it may not be solely responsible for that instance's lifecycle); and an open triangle indicates a class hierarchy with the triangle pointing towards the parent class. Reprinted with permission under the terms of the Creative Commons Attribution License: PLOS ONE Chaudhury *et al.* "Benchmarking and analysis of protein docking performance in Rosetta v3.2"[42]

`DockingLowRes` and `DockingHighRes` contain all the data and functions associated with the low-resolution docking and high-resolution refinement stages, respectively, including the score functions, sampling functions (including translation/rotation parameters and side-chain packing), and Monte Carlo data. Both objects are independent of the Rosetta options system and can be called directly within the Rosetta source code or through Rosetta interfaces such as PyRosetta[45] and RosettaScripts.[46] Given the wide range of minimization and side-chain packing strategies that might be utilized in the high-resolution docking stage, `DockingHighRes` is designed as an abstract class that underlies a diverse set of high-

resolution docking functions including standard high-resolution docking, pre-packing, as well as extensions of docking such as peptide docking and protein interface design. This versatility is achieved through the `DockTaskFactory` class within `DockingHighRes`, which handles all docking side-chain packing options and allows subclasses of `DockingHighRes` to be able to create a tailored set of packing instructions (Figure 2.3). All docking objects contain default parameters that allow them to be run with minimal setup; users only need to specify docking parameters for non-default behavior.

As a testament to the flexibility of this design, the current versions of EnsembleDock and SnugDock are implemented using derived classes of `DockingLowRes` and `DockingHighRes`, docking is configurable via RosettaScripts and the docking classes are widely used throughout the Rosetta source code.

2.4 Development of new visualization techniques

Visualizing structural models of proteins is often a critical step in evaluating the performance of a new method, formulating new hypotheses or identifying bugs that lead to systematic errors. Although there are many excellent visualization tools available, successful visualization requires outputting the atomic coordinates of the model at a particular point in a simulation. In the case of debugging a new method, determining when to output coordinates can be tedious and time consuming. Conversely, even when one is working with an established method, the modes of motion can be mystifying in the absence of visualization.

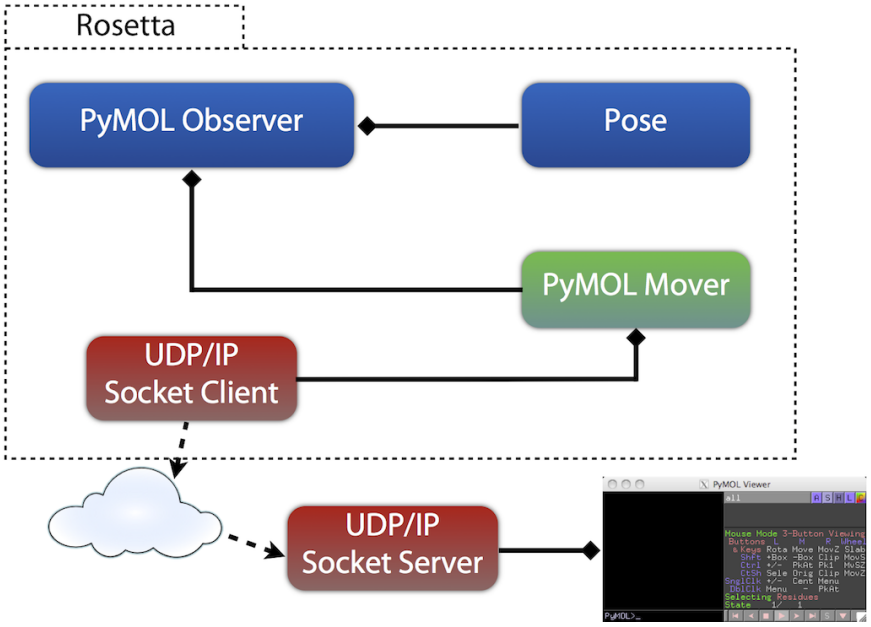

Figure 2.4: Rosetta transmits data through the PyMOL_Mover's UDP/IP socket client to an IP address. Dotted arrows represent network communication and diamonds represent composition (*i.e.* the PyMOL Observer contains a PyMOL Mover and an owning pointer to a Pose). The PyMOL Observer monitors changes in a Pose and uses the PyMOL Mover to transmit this information to PyMOL. The UDP/IP socket server running in PyMOL listens for network traffic and translates appropriate packets. Once the data is translated, PyMOL displays biomolecular structures. Reprinted with permission under the terms of the Creative Commons Attribution License: PLOS ONE Baugh *et al.* "Real-time PyMOL visualization for Rosetta and PyRosetta"[47]

Inspired by the interactive nature of PyRosetta,[45] we developed a novel real-time visualization solution that links PyMOL[48] and Rosetta/PyRosetta. My role in this project was determining the set of features that would be supported within Rosetta and developing the API that could be used by developers. As shown in Figure 2.4, the solution relies on running a UDP server within PyMOL and transmitting data via UDP from Rosetta. UDP does not require implicit "hand-shaking" and tolerates lost transmissions, which enables running each program separately. Maintaining the separation of these programs prevents either from losing focus; Rosetta performs simulations and calculations while PyMOL

performs visualization.

In addition to being a useful tool to display protein structures, PyMOL is well suited for the production of informative movies. Rosetta simulations are often presented visually to demonstrate or explain the principles underlying Rosetta algorithms. Previously, making movies of Rosetta protocols required significant work. When sending data to PyMOL, the user may simply retain output history to produce PyMOL movies. The history feature also allows the user to inspect protocols that are otherwise inaccessible.

2.5 Modeling new classes of proteins

Membrane proteins are critical functional molecules in the human body, but experimental methods to determine their structures are fraught with difficulty. Thus, computational tools tailored for membrane proteins lag behind those intended to be used for soluble proteins. In contrast to the enormous structural diversity of soluble proteins, the structural motifs in the membrane environment are either α-helical bundles or β-barrels, which, coupled with constraints imposed by the membrane, leads to a restricted conformational search space. This reduction in conformational sampling space is offset by the larger size of membrane protein complexes, which necessitates the development of efficient sampling methods. The distinction between native-like from non-native models requires accurate scoring functions, which have proven difficult to formulate for the heterogeneous environment of the lipid bilayer. To provide a central access point for protocol development, we sought to create a class that stores all information necessary for representing the Pose in the membrane bilayer.

For this project, I assisted in designing this class.

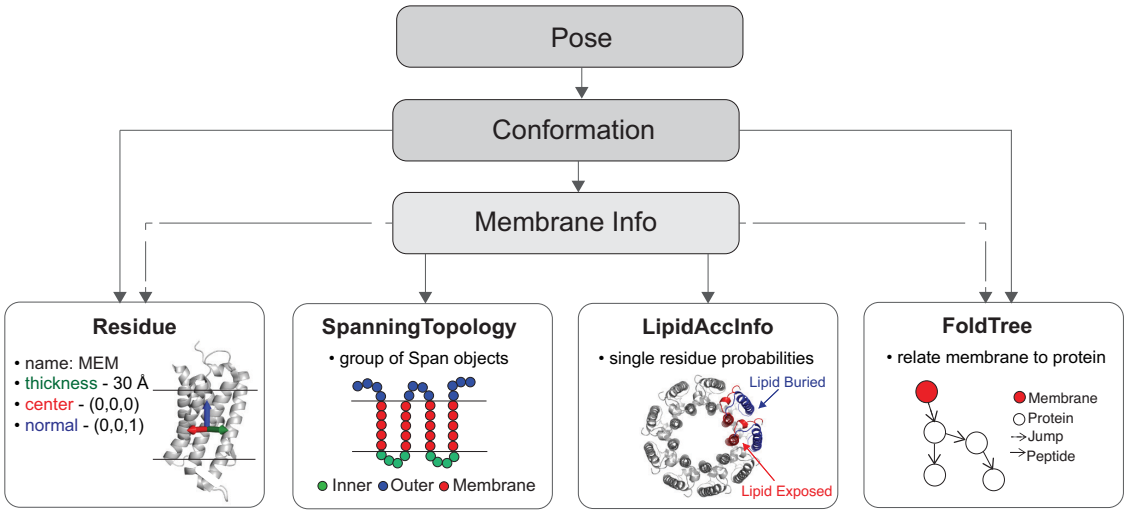

Figure 2.5: The Membrane Framework represents the membrane bilayer using four main components connected to a central `MembraneInfo` object (light gray). This object stores information needed to represent the membrane (solid arrows) and tracks information already in the `Pose` coordinates (dashed arrows). A special membrane residue is added to the `Pose`, whose coordinates indicate the center, normal and thickness of the bilayer (bottom left). A connection between the membrane residue and protein is established through the `FoldTree` object using a `Jump` edge (bottom right). A `SpanningTopology` object is used to describe regions of the Pose that span the membrane bilayer (second from left). Finally, a `LipidAccInfo` object is used to describe single-residue probabilities of lipid exposure or burial (second from right). Reprinted with permission under the terms of the Creative Commons Attribution License: Alford *et al.* "An integrated framework advancing membrane protein modeling and design"*sub juice*[49]

The information is organized in the `MembraneInfo` object, which stores descriptors of sequence- and structure-based protein properties, such as membrane protein topology and lipophilicity, and manages the attachment of a virtual membrane residue to the `Pose` to represent the membrane bilayer. The `MembraneInfo` object is a member of the `Conformation` object, which is part of the `Pose` (Figure 2.6). Because the `Pose` is the central object in Rosetta protocols, the information in `MembraneInfo` is readily available to novel protocols. Thus, by developing a robust solution to provide access to these properties, we have enabled Rosetta

to model a new class of proteins.

2.6 Controlling access to degrees of freedom

The text in this section is primarily taken from of Porter *et al.* "A framework simplifying combined sampling modes in Rosetta",[50] with permission under the terms of the Creative Commons Attribution License.

One of the core implications of the Levinthal Paradox[33] for the development of computational structure prediction methods is the necessity of directing sampling algorithms toward reasonable conformations. In Rosetta, such protocols frequently draw upon knowledge of physical chemistry and, sometimes, experimental observations about the specific system under consideration. As a result, the most effective sampling schemes for a particular system are not the established, benchmarked protocols but rather variants of those protocols that incorporate all the available information about the protein of interest. For example, incorporating explicit sampling of β-sheet pairing has been successful in a number of contexts.[51] Most protocols, however, have not been developed with target-specific optimizations in mind, making such modifications time consuming for experienced Rosetta developers and impossible for others.

To address this we developed a framework, the BrokeredEnvironment, for the rapid combination of sampling strategies, which reduces the burden on both developers and users when combining various sampling strategies. The BrokeredEnvironment operates by taking control over shared resources (*e.g.* the fold tree and control over simulated degrees of freedom (DoFs)) within the simulation system and requiring participating sampling algorithms to declare the required degree of control over these resources at application launch time.

I developed the methodology by which docking applications can be represented within the BrokeredEnvironment. A virtual residue is positioned at the center of mass of a specified region of a protein, and its position is updated automatically as necessary. This approach enables docking to coexist with other tasks, such as loop modeling, in a single simulation.

Using this information, a consensus fold tree and DoF accessibility are generated and enforced. This system allows a number of useful but technically demanding features to be incorporated into existing simulations without additional C++ development, including procedural generation of fold tree based on decoy-specific data, trivial composition of movers, and the ability to use a single fold tree in simulations with complex sampling behavior. Together these features allow for a level of algorithm rapid prototyping previously unavailable in Rosetta.

2.7 Summary

The tools described in this chapter are useful in their own right, and developing them has imbued me with the ability to rapidly develop robust, correct code. Developing frameworks, visualization methods and widely-used applications provided me with the expertise that is required to tackle the daunting challenge of developing a tool to accurately model antibodies.

CHAPTER III

BLIND PREDICTION PERFORMANCE OF ROSETTAANTIBODY 3.0

Adapted from Weitzner BD*, Kuroda D*, Marze N, Xu J & Gray JJ, "Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization," *Proteins* 82(2), 1611–23.
 *Joint First Authors.

3.1 Overview

Antibody Modeling Assessment II (AMA II) provided an opportunity to benchmark RosettaAntibody on a set of 11 unpublished antibody structures. RosettaAntibody produced accurate, physically realistic models, with all framework regions and 42 of the 55 non-H3 CDR loops predicted to under an Ångström. The performance is notable when modeling H3 on a homology framework, where RosettaAntibody produced the best model among all participants for four of the 11 targets, two of which were predicted with sub-Ångström accuracy. To improve RosettaAntibody, I investigated the causes of model errors. The most common limitation was template unavailability, underscoring the need for more antibody structures and/or better *de novo* loop methods. In some cases, better templates could have been found by considering residues outside of the CDRs. *De novo* CDR H3 modeling remains challenging at long loop lengths, but constraining the C-terminal end of H3 to a kinked conformation allows near-native conformations to be sampled more fre-

quently. I also found that incorrect V_L–V_H orientations caused models with low H3 RMSDs to score poorly, suggesting that correct V_L–V_H orientations will improve discrimination between near-native and incorrect conformations. These observations will guide the future development of RosettaAntibody.

3.2 Introduction

Antibodies are vital immunological molecules, protecting their hosts by binding to their infectious targets, antigens, and triggering a directed immune response. In addition to their biological role, antibodies serve as protein therapeutics.[52,53] Advances in computational protein modeling and a growing understanding of the sequence–structure relationship in antibodies have fueled development of methods to engineer improved affinity,[54–57] stability[58] and solubility.[58–63]

Antibody structure prediction typically focuses on modeling the variable fragment (F_V), which is composed of the N-terminal domains from the heavy and light chains (V_H, V_L). The F_V contains the antigen-binding site,[64] composed of the six complementarity determining region (CDR) loops (L1–L3, H1–H3) that are responsible for antigen recognition and binding. While the structure of the V_L and V_H domains is highly conserved, the CDR loops, especially CDR H3, vary considerably both in terms of sequence and structure, prompting many studies, both computational and experimental, focused on CDR loops[17–19,24,26,28,65–72] and their interactions with antigens.[30,73–81] Several antibody prediction methods are available as servers on the web.[82–84]

In 2011 the first Antibody Modeling Assessment (AMA I) was conducted, in which

some of these servers and commercial software tools were benchmarked against nine newly determined antibody crystal structures.[23] Templated regions were predicted to about 1 Å RMSD, and CDR H3 loops were predicted to about 3 Å RMSD. Since that time, efforts in the antibody structure prediction field have included updated databases,[85] additional examination of the V_L–V_H orientation,[86] reassessment of canonical loop clusters,[19] designs of antibodies for thermal resistance[59] and non-canonical residue antigen crosslinking.[4] In addition, progress has been made in *ab initio* loop modeling.[87–90] In response to these developments, a second antibody modeling assessment was organized in 2013.

In this report, I discuss the performance of RosettaAntibody implemented in the Rosetta 3 framework[32] when blindly predicting the structure of eleven unpublished antibody crystal structures as a part of Antibody Modeling Assessment II (AMA II). This experiment is the second blind test of several antibody-modeling methods and the first test of an updated Rosetta-based antibody modeling method still under active development. The eleven targets provided to me represent a diverse set of antibodies, including a rabbit antibody (Ab01), a human antibody with a λ light chain (Ab05), antibodies derived from phage display libraries (Ab03 and Ab05) and CDR H3 loops ranging from 8–14 residues in length (Kabat/Chothia definition). Modeling these targets enabled me both to test new methods and to incorporate the results from AMA I into our workflow. In addition to our overall performance, I discuss sampling and scoring issues that can guide future improvements to RosettaAntibody.

3.3 Methods

3.3.1 Target Sequences

The target dataset consisted of the sequences for 11 unpublished antibody F_V structures crystallized in the free state with a maximum resolution of 2.8 Å comprising 6 mouse antibodies, 4 human antibodies and 1 rabbit antibody.

3.3.2 Construction and relaxation of the crude F_V models

I used a new Python script for the first step of antibody modeling to build a crude F_V model and relax it to remove grafting anomalies. The script inputs light and heavy chain sequences and calls BLAST for the template selections and several Rosetta applications for the template grafting and refinement. Then, I assessed the model geometry and torsion angles by MolProbity. If the MolProbity score[91] for the model was poor, the problematic templates were removed from the database and the process repeated. This process produced a Chothia-numbered intermediate structure and a constraint file for CDR H3 loop *de novo* modeling.

3.3.3 Kinematic loop modeling and simultaneous V_L–V_H optimization

After the initial model was refined, the CDR H3 loop was modeled de novo while simultaneously refining the V_L–V_H orientation using the Rosetta docking algorithm[42] (stage 1). Next-generation KIC (NGK)[89] without two-body Ramachandran sampling and legacy

KIC[88] were used to sample CDR H3 loop conformations. The conventional sequence-based classification rules[26] predicted all targets other than Ab07 to have a kinked CDR H3 loop. The sequence of the Ab07 is featureless, and since the majority of antibodies have a kinked CDR H3 conformation, it was also presumed to adopt a kinked conformation. The kink prediction is incorporated into the sampling routine by restricting the pseudodihedral angle of the four consecutive C_α atoms of residues H100X, H101, H102 and H103 to -10° to 70°, a range consistent with the kink.[24]

3.3.4 CDR H3 loop modeling on a crystal framework

For stage 2, I was given the crystal structures for the targets with the CDR H3 loop coordinates removed. After repacking the side chains, I ran NGK with two-body Ramachandran sampling as well as legacy KIC without any constraints for 7 targets (Ab04/05/06/07/09/10/11), and legacy KIC with the kink constraint for 3 targets (Ab02/Ab03/Ab08). Given the rapid turnaround required for this challenge, the protocols for each target were chosen based on the estimated computational time required and available resources. For Ab05/Ab06/Ab10, however, the kinked conformations were rarely sampled in the unconstrained simulations, so I employed legacy KIC with the kink constraint as described above.

3.3.5 CDR loop definitions

RosettaAntibody uses the Chothia numbering scheme.[18] CDRs L1–L3, H2 and H3 follow the Kabat definitions (L1: L24–L34, L2: L50–56, L3: L89–L97, H2: H50–H65, H3: 95–102), while CDR H1 is defined as residues H26-H34. FRL and FRH are defined as the whole V_L and V_H domains except for the CDR loops.

3.3.6 V_L–V_H packing angle calculation

The V_L–V_H packing angle, α, is calculated using a Rosetta implementation of the protocol described in Abhinandan and Martin,[92] which defines the packing angle as a pseudo-torsion angle between four non-atomic points at the V_L–V_H interface. These points are identified using two pairs of conserved β-strands at the V_L–V_H interface, one pair located in the V_L framework (L35-L38, L85-L88), the other in the V_H framework (H36-H39, H89-H92). For each β-strand pair, C_α coordinates were extracted, and the centroid and best-fit line (first principal component) of the coordinate set were identified. Points 2 and 3 in the pseudo-torsion calculation are defined as the V_L centroid and V_H centroid, respectively, while points 1 and 4 are defined as points along the V_L and V_H best-fit lines, respectively, that lie to the same side of the centroid as the CDRs.

3.3.7 RMSD calculation

As reported in AMA I,[23] all RMSDs for model assessment were calculated over the backbone atoms (C, C_α, N, O). The RMSDs of CDR-H and L are computed after superposing the corresponding FR, while the RMSDs used to assess domain orientation are defined as the RMSD of FRH and FRL after superposing FRL or FRH, respectively. The RMSD of template availability was examined based on the CDRs in the Chothia definition, which excludes structurally conserved regions from our CDR definitions. All RMSDs were computed using the McLachlan algorithm[93] as implemented in the ProFit software.[94] All antibody models generated by RosettaAntibody 3.0 are available upon request or on the web (`http://www.Abmodeling.com`).

3.3.8 MolProbity

I used MolProbity version 3.[91] To ensure fair comparisons between crystal structures and models, all hydrogen atoms are removed from the models before calculating MolProbity scores.

3.3.9 Algorithm Availability

All methods used for this work are included in the Rosetta biomolecular modeling suite, distributed freely for academics and non-profits through the Rosetta Commons (`http://www.rosettacommons.org`). Along with compiled Rosetta executables, there are pre-and post-processing scripts and tools. The initial template grafting and refinement is driven by a master python script as follows:

```
./antibody.py --light-chain <L.fasta> --heavy-chain <H.fasta>
```

This script generates several PDB files and a constraint file called cter_constraint, which is for the kink constraint. The grafted.relaxed.pdb is recommended to use in the H3 modeling step below. The script is included in the latest Rosetta release (Rosetta/tools/antibody/antibody.py),

The H3 modeling jobs above can be run using the Rosetta command. For NGK with kink constraint, the command line is:

```
./antibody_H3.macosclangrelease
  -s ./grafted.relax.pdb
  -antibody::remodel perturb_kic
  -antibody::snugfit true
  -antibody::refine refine_kic
  -antibody::cter_insert false
  -antibody::flank_residue_min true
  -antibody::bad_nter false
```

```
-antibody::h3_filter true
-antibody::h3_filter_tolerance 20
-antibody:constrain_cter
-antibody:constrain_vlvh_qq
-constraints:cst_file ./cter_constraint
-ex1
-ex2
-extrachi_cutoff 0
-kic_bump_overlap_factor 0.36
-corrections:score:use_bicubic_interpolation false
-loops:legacy_kic false
-loops:kic_min_after_repack true
-loops:kic_omega_sampling
-loops:allow_omega_move true
-loops:ramp_fa_rep
-loops:ramp_rama
-loops:outer_cycles 5
-run:multiple_processes_writing_to_one_directory
-nstruct 2000
```

For legacy KIC with the kink constraint, the command line is:

```
./antibody_H3.macosclangrelease
  -s grafted.relaxed.pdb
  -antibody:remodel perturb_kic
  -antibody:snugfit true
  -antibody:refine refine_kic
  -antibody:flank_residue_min true
  -antibody:bad_nter false
  -antibody:h3_filter false
  -antibody:cter_insert false
  -ex1
  -ex2
  -constraints:cst_file ./cter_constraint
  -nstruct 2000
```

These flags are compatible with the public Rosetta release 2013-wk48 (3-Dec-2013).

3.4 Results

The basis of our approach used for AMA II is the original RosettaAntibody algorithm described by Sivasubramanian *et al.* in 2009,[22] with each component revisited and updated. Briefly, RosettaAntibody uses templates from other antibody structures for the framework regions (FRs) and non-H3 CDR loops.[95] Because CDR H3 does not form canonical conformations, it is modeled *de novo* while optimizing the V_L–V_H orientation and minimizing CDR loop torsions. In this chapter, I analyze successes and shortcomings at each step of the process, starting with template selection, then *de novo* CDR H3 modeling on both a homology and a crystal framework, and finally V_L–V_H orientation.

Target	FRL	FRH	L1	L2	L3	H1	H2	H3	light_heavy
Ab01	2hwz	1mvu	3ghb	2aj3	2otu	3s34	2ojz	1uz6	1dql
Ab02	1mvu	3cvi	3o2d	3t65	3o2d	2ok0	1ktr	1p7k	3q3g
Ab03	3eo9	3qot	1rzg	3ncj	1yqv	3nps	3qot	1fgn	2cmr
Ab04	1ncc	2jel	1ztx	1h8n	1kb5	1nlb	2jel	1uz6	1ztx
Ab05	1aqk	3njc	4d9l	3c2a	1rzf	2b1h	3ncj	1uz6	2xwt
Ab06	3uc0	3kdm	1vge	3idg	1vge	3s34	2hrp	1dql	3bn9
Ab07	2xqy	1ft8	2vl5	3phq	2aab	3rvv	1hq4	1uj3	1f58
Ab08	2ap2	1q0x	1mvu	3t65	1mvu	2q76	1d5i	3phq	1mvu
Ab09	3eo9	3njc	1hez	3ncj	3qot	2h32	2xwt	3qot	3nab
Ab10	3t65	1e4x	3qot	3t65	3oz9	1kb5	1ktr	2xzq	3o2d
Ab11	1yy8	3cvi	1yy8	2ih3	1bm3	3cvi	1igj	1s3k	2oz4

Table 3.1: PDB accession codes of the source of the template used for each antibody structural component. FRL and FRH, light and heavy variable domain framework templates; L1. . . H3, complementarity determining loops L1 through H3 templates; light_heavy; template for initial FRL–FRH orientation.

3.4.1 Template based modeling is accurate. . . except when it's not

RosettaAntibody begins by searching curated databases containing sequences of structural components (FRL, FRH, L1–3, H1–3) from high-quality antibody crystal structures (resolution $\leq$ 2.8 Å CDR C_α B-factors $\leq$ 50) in the Protein Data Bank (PDB)[15,96] as of June 2012. Templates are selected by BLAST[97] bit-score. A template for the initial V_L–V_H orientation is similarly identified by using an overall sequence similarity to a complete F_V. The template structures are then assembled into a crude model and refined in Rosetta. The templates for each structural component are listed in Table 3.1.

Table 3.2 shows RMSDs of the templates for each submitted model. Excluding the rabbit Ab01, the average backbone RMSDs of the L1, L2, L3, H1 and H2 CDR loops were 0.61 ± 0.27 Å, 0.48 ± 0.12 Å, 1.02 ± 0.54 Å, 1.00 ± 0.63 Å, and 0.99 ± 0.64 Å, respectively. All of our submitted models have sub-Ångström FRL and FRH regions relative to the crystal structure including Ab01. So as in previous works,[16–19] the CDRs, other than H3, form canonical conformations, which, when identified correctly, provide high-quality backbone atom coordinates for the loop. In this assessment, 42 of the 55 non-H3 CDR loops submitted were predicted to sub-Ångström accuracy.

I examined the validity of our hypothesis of choosing templates based on BLAST bit score by investigating the causes of modeling errors in the 13 non-H3 loops that were not predicted accurately Figure 3.1 shows the RMSD of all possible candidate templates to the crystal loop structure for three representative loops. Figure 3.1A shows the successful case of the CDR L1 loop for Ab06; there are many accurate (low RMSD) loop templates in the database, and those with the highest sequence identity or BLAST bit score are accurate. In

Target	Model	FRL	FRH	L1	L2	L3	H1	H2	FRL-frh	FRH-frl
	1	0.43	0.72	3.27	0.44	4.13	0.69	1.82	2.08	2.09
Ab01	2	0.44	0.72	3.28	0.44	4.17	0.72	1.85	2.14	2.33
	3	0.44	0.72	3.27	0.42	4.38	0.7	1.99	1.86	1.88
	1	0.33	0.72	0.66	0.44	1.1	0.86	1.04	2.28	1.99
Ab02	2	0.33	0.72	0.59	0.45	1.12	0.94	1.16	1.93	1.9
	3	0.33	0.71	0.56	0.43	1.08	0.83	0.99	2.31	1.77
	1	0.37	0.51	0.36	0.46	1.32	2.58	2.24	2.89	1.93
Ab03	2	0.37	0.5	0.36	0.47	1.28	2.63	2.04	2.89	1.93
	3	0.37	0.51	0.37	0.45	1.32	2.59	2.05	1.34	1.49
	1	0.63	1.1	0.58	0.81	1	0.71	0.24	1.46	1.37
Ab04	2	0.63	1.1	0.58	0.66	1.03	0.73	0.24	2.19	2.14
	3	0.63	1.1	0.58	0.84	0.97	0.78	0.24	1.59	1.89
	1	0.67	0.29	1.16	0.51	2.17	0.92	0.71	1.95	1.5
Ab05	2	0.67	0.29	1.15	0.52	2.21	1.13	0.89	1.55	1.75
	3	0.67	0.28	1.2	0.47	2.13	0.96	0.67	1.7	1.82
	1	0.47	0.5	0.32	0.46	0.71	1.09	0.51	1.32	1.32
Ab06	2	0.47	0.5	0.34	0.44	0.72	1.05	0.58	1.66	1.33
	3	0.47	0.5	0.32	0.46	0.84	1.05	0.51	1.53	1.49
	1	0.31	0.5	0.82	0.36	0.64	0.46	0.37	1.25	0.84
Ab07	2	0.31	0.51	0.78	0.37	0.6	0.5	0.52	2.01	0.88
	3	0.31	0.51	0.82	0.37	0.64	0.44	0.57	1.96	1.36
	1	0.52	0.46	0.8	0.55	0.58	0.55	0.89	2.27	3.28
Ab08	2	0.52	0.46	0.9	0.55	0.59	0.47	1.01	0.82	0.9
	3	0.52	0.45	0.84	0.57	0.61	0.46	0.84	1.54	2.99
	1	0.34	0.27	0.35	0.44	0.42	0.37	0.38	1.18	0.99
Ab09	2	0.34	0.26	0.36	0.41	0.41	0.37	0.48	0.81	0.78
	3	0.34	0.26	0.36	0.41	0.36	0.34	0.78	1.07	0.96
	1	0.42	1.02	0.71	0.38	1.7	1.36	0.96	1.55	1.44
Ab10	2	0.42	1.02	0.68	0.42	1.7	1.43	1	1.56	0.84
	3	0.42	1.02	0.76	0.37	1.69	1.65	1.31	1.05	0.55
	1	0.42	0.56	0.39	0.42	0.62	1.04	1.99	1.41	0.95
Ab11	2	0.42	0.56	0.38	0.43	0.55	0.85	2.23	1.56	0.95
	3	0.42	0.56	0.37	0.32	0.59	0.89	2.17	1.39	1.29
Ab01-11	**Mean**	**0.45 ± 0.11**	**0.60 ± 0.26**	**0.85 ± 0.82**	**0.47 ± 0.11**	**1.31 ± 1.07**	**0.97 ± 0.61**	**1.07 ± 0.66**	**1.70 ± 0.51**	**1.54 ± 0.62**
Ab02-11	**Mean**	**0.45 ± 0.12**	**0.59 ± 0.27**	**0.61 ± 0.27**	**0.48 ± 0.12**	**1.02 ± 0.54**	**1.00 ± 0.63**	**0.99 ± 0.64**	**1.67 ± 0.52**	**1.49 ± 0.62**

Table 3.2: RMSD of heavy and light variable domain framework regions and non-H3 CDR loops for all models in stage 1.

Table 3.2: RMSD of heavy and light variable domain framework regions (FRH/FRL) and non-H3 CDR loops (L1...H2) for all submitted models in stage 1. The model FRL and FRH were superposed onto the corresponding crystallographic framework before computing the RMSD, while CDR loop RMSDs were computed after superposing the FR (*i.e.* to compute the RMSD of CDR L1-3, the model FRL was superposed onto the crystallographic FRL, and FRHs were superposed before computing the and the RMSDs for H1 and H2). In order to measure the effect of V_L–V_H orientation, the RMSD of each framework was computed after superposing the other framework (*i.e.* the RMSD of FRL was computed after superposing FRH). These values are annotated as FRL-frh and FRH-frl where the FR written in lowercase is the FR that was superposed. All RMSDs are reported in Å.

this case our models were based upon the 1vge template and, after all refinements, resulted in loop RMSDs of 0.71–0.84 Å.

Figure 3.1B shows accuracies of all template candidates for CDR loop H2 in Ab11. In this case, many low-RMSD templates were available, but the algorithm chose a template from 1igi that was less accurate. Four cases in total exhibited this failure mode (Ab03-L3, Ab10-L3, Ab06-H1, Ab11-H2), missing potential sub-Ångström templates. These failures suggest that incorporating other environmental effects into template selection may be necessary. The structural determinants of the canonical CDR conformations includes some residues in the framework regions,[19,65] and the identity of these residues has a species dependence. This information has been used to guide the humanization of antibodies[98] and would likely be useful in building a more sophisticated template selection scheme.

Figure 3.1C shows a third example, namely the CDR H1 loop of Ab03. In this case, there are no near-native templates in the structural database, even though there are three templates with an exact loop sequence match. Five failure cases in total suffer from lack of accurate (sub-Ångström) templates (Ab01-L1/L3/H2, Ab03-H1, and Ab05-L3). Unsurprisingly, three of these five loops are in the rabbit antibody and have uncommon loop lengths (Ab01-L1, L3, and H2). The other targets in this category, Ab03 and Ab05,

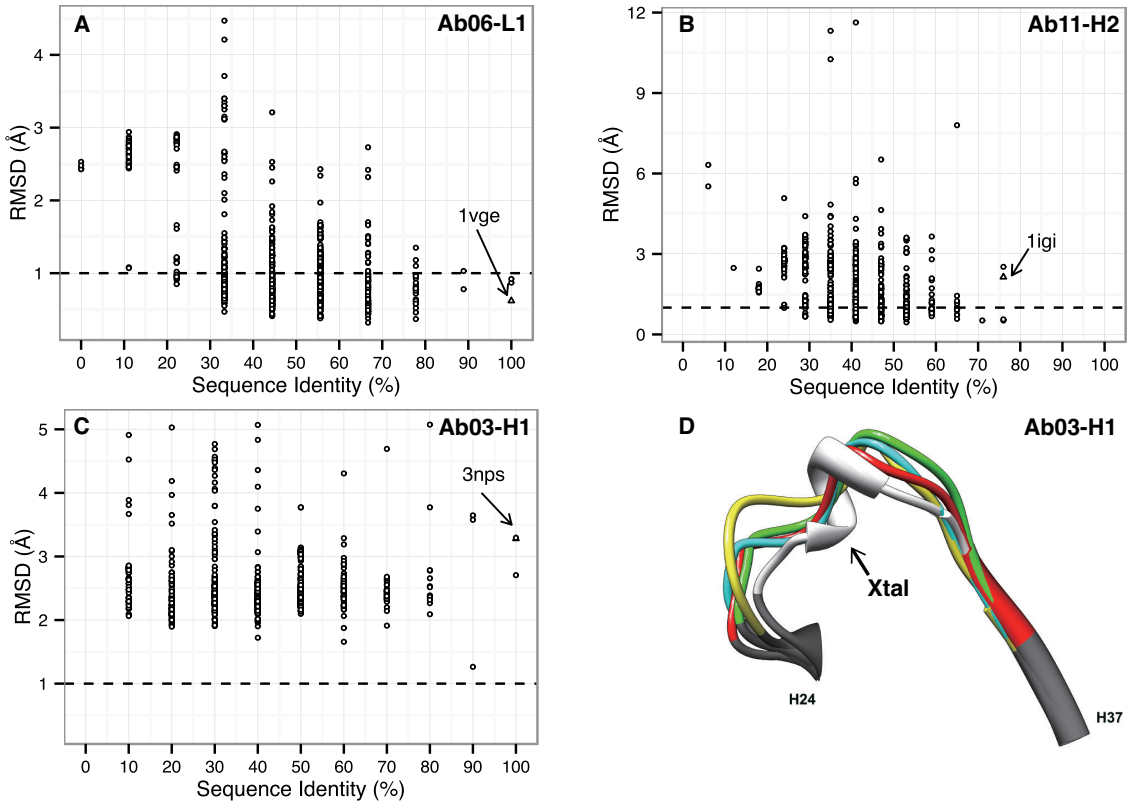

Figure 3.1: CDR loop template selection. Scatterplot of the RMSD versus sequence identity (SID) for (A) a typical success case (Ab06-L1) where a sub-Ångström template is selected; (B) a case where a good template is in the database but RosettaAntibody does not select it (Ab11-H2); and (C) a case where there are no templates in our structural database with an RMSD $\leq$ 1.0 Å of the target, which results in a modeling failure (Ab03-H1). Dashed line at RMSD = 1.0 Å for reference. The superposition of the models on the crystal structure (D) shows the result of this modeling failure.

are human antibodies derived from phage display libraries, suggesting that phage display may yield structures that depart from those in biologically derived antibodies. Ab01, the rabbit antibody, requires separate discussion. During the challenge, only 1 rabbit antibody (4HBC) was available in the PDB, and I used it for templates for the frameworks (FRL, FRH), the CDRs that matched in length (L2, H1, H2) and the initial V_L–V_H orientation. However, the RMSD values are much higher than the other targets. Thus, more rabbit antibody crystal structures are needed for more diverse templates, or I must recognize

challenging, non-template loops and resort to de novo loop building.

Ab03 CDR H1 has templates with excellent sequence identity (100% SID; 3NPS/3QOT/1RZI) but incorrect structures (Figure 3.1D). Based on the BLAST bit-score, I choose 3NPS, which is a complex of a human antibody and a membrane-type serine protease with some contacts between the H1 and the antigen. Although the unbound-state antibody is not available in the current PDB, the H1 can be classified into a known canonical conformation, and it is typical that the backbone of the H1 conformation is not influenced by the contact with the antigen. In the case of the human germline antibody 3QOT, the B-factors of the H1 region are high, but it still forms the same canonical conformation as 3NPS. 1RZI is a crystal structure of anti-HIV human antibody, which contains eight unique F_Vs in the asymmetric unit. I included only the first heavy and light chains in the file (B and A) during the database construction process, resulting in a candidate template with 2.7 Å RMSD (Figure 3.1D). The H1 loop of chain L, which was not included in our database, is closer to the target with 2.3 Å RMSD, indicating that structural differences between different F_Vs in the asymmetric unit can be significant. Thus, if all asymmetric unit chains were included in our database, it would have been possible to identify a better loop template. Further, H1 conformations both with (2CMR, 1.3 Å RMSD) and without a helical region are present in the database with 90% sequence identity to Ab03-H1, indicating that canonical conformations are heavily influenced by the local environment and highlighting the difficulty of selecting the best template for this target.

Ab05 has a λ light chain, which is underrepresented in the PDB (*i.e.* 61 of 415 light chains in our curated non-redundant database). Although there are 56 templates of

11-residue CDR L3 loops available in the database, the highest sequence identity is only 55% (2J6E; RMSD 1.7 Å), and no template is structurally similar to the CDR L3 loop of Ab05.

The remaining four poorly predicted non-H3 CDR loops (Ab02-L3, Ab03-H2, Ab05-L1, Ab10-H1) have low-RMSD templates in the database which our protocol can select correctly, but minimization of the loop perturbed the coordinates to an incorrect conformation. As discussed in the previous antibody modeling assessment,[23] relaxation to improve the physical realism of a model can destroy the accuracy originally present in a crystallographically-derived template. In fact, an antibody modeling server, PIGS, often generates better non-H3 CDR backbones, but several steric clashes and bad geometries are observed as reflected in poor MolProbity score.[23]

In summary, three scenarios led to failures in CDR template selection: 1) no availability of low-RMSD templates, (six cases); 2) inability to identify the best template (four cases); and 3) perturbation of the template away from the native structure by energy minimization and refinement (three cases). Fortunately these scenarios were relatively rare, and 42 of the submitted non-H3 CDR loops were predicted correctly (*i.e.* backbone RMSD $\leq$ 1.0 Å).

3.4.2 Template refinement improves physical realism of models

In the first Antibody Modeling Assessment,[23] some Rosetta antibody structures suffered from poor MolProbity scores. MolProbity tests structure files for reasonable backbone torsion angles and clashes.[91] I determined that some of these issues arose from uncommon template backbone angles and odd torsion angles at the graft points. To improve these ratings and make the RosettaAntibody models more physically realistic, I tested new

relaxation methods after the template grafting and before the CDR refinement steps.

The new template refinement steps are as follows. After grafting the selected template structures, the bond angles and bond lengths are set to standard values[99] to alleviate artifacts at the graft points. The model is refined by running side-chain repack and minimization cycles, where I gradually increase the weight of the repulsive component of the Lennard-Jones potential and enforce all-atom constraint to prevent large distortions from the original templates.[100] Figure 3.2 shows the improvement in geometry as assessed by MolProbity scores. Initial models after grafting have clashes and strained backbone dihedral angles, but the refinement process results in models with geometries that score well within the range of the crystal structures.

3.4.3 β-sandwich assembly is accurate for antibodies with a near-average packing angle

The accuracy of the β-sandwich assembly process can be assessed by superposing the model and crystal FRH domains and examining the RMSD of the FRL domains. This RMSD is sub-Ångström for five out of the eleven targets (Ab07, Ab08, Ab09, Ab10, Ab11).

All five of the correctly-predicted targets have a V_L–V_H packing angle,[92] α(see Methods), within one standard deviation of the mean packing angle of antibodies in the PDB (-52.3° ± 3.9°). Among the six targets without a sub-Ångström model, three (Ab01, Ab02, and Ab05) have packing angles further than one standard deviation from the PDB average: -46.6°, -56.8° and -47.2°, respectively. The initial orientation of these three targets was taken from 1DQL (α = -50.7°), 3G3G (α = -49.9°), and 2XWT (α = -52.4°), respectively; each of these initial orientations was closer to the PDB average than to the target packing

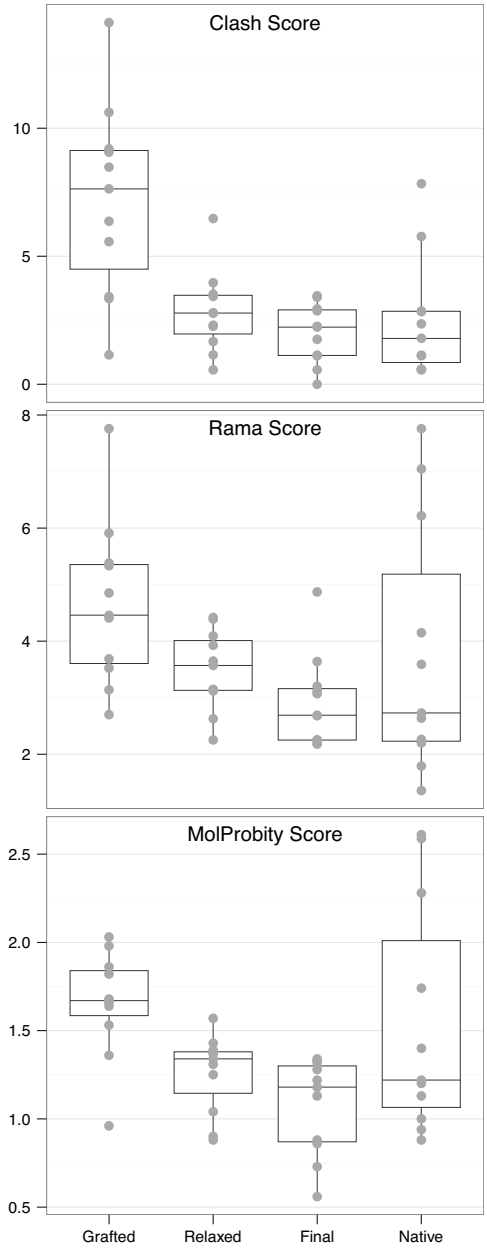

Figure 3.2: MolProbity scores of model quality at various stages. Scores for each target are plotted (1) after the initial grafting; (2) after relaxation using the Rosetta force field; and (3) for the final model. The crystal structure score is shown for reference. Hydrogen atoms were omitted. Overall, MolProbity scores improve throughout refinement, ending up within the range of the crystal structures.

angle. Similarly, our final models for these three targets have packing angles ranging from -49° to -52°, also closer to the PDB average packing angle than to the target.

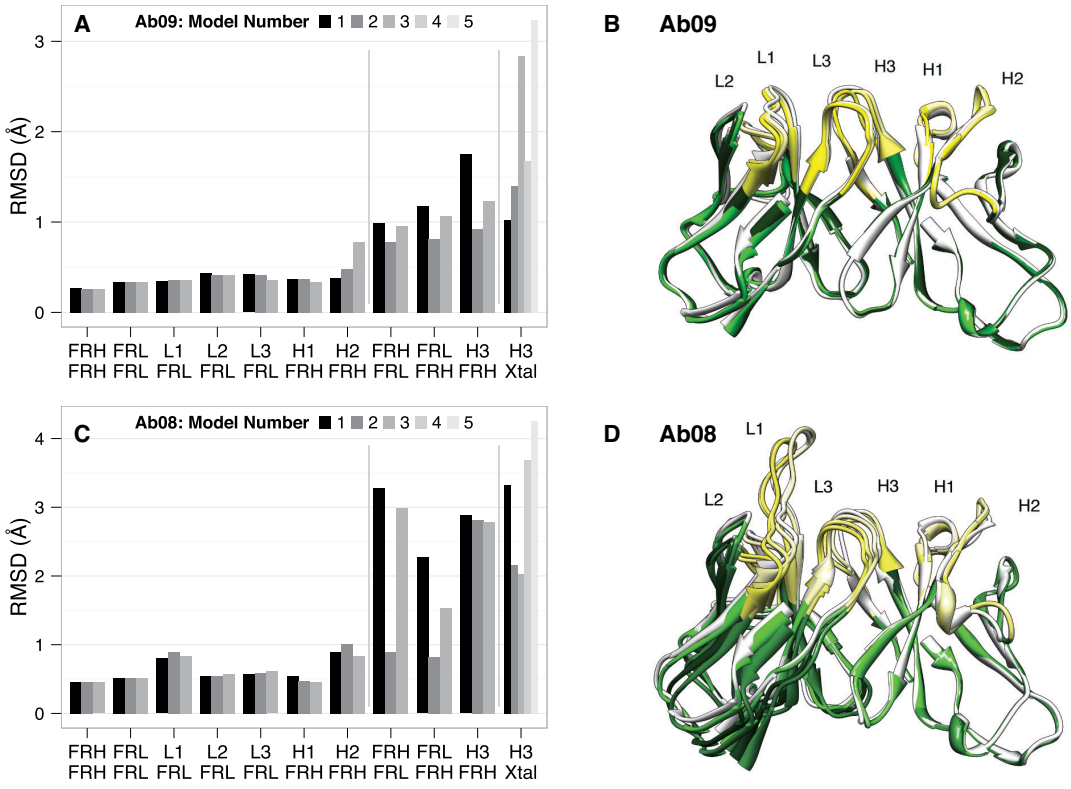

Figure 3.3: Examples of convergent (A, B) and divergent (C, D) modeling attempts. RMSDs are plotted for each structural component (top line along horizontal axis) when a particular alignment (bottom line along horizontal axis) is performed. (A) Ab09; (C) Ab08. When CDR H3 and the V_L–V_H orientation varies between models, top-scoring models diverge (C). The top-scoring models superimposed on the crystal structure are shown for Ab09 (B) and Ab08 (D). Figures generated in UCSF Chimera.[101]

Figure 3.3A shows the RMSD for each structural component when a particular alignment is performed for Ab09, a typical success case. The non-H3 CDR loops and the self-aligned frameworks show little variation between models because these regions are not explicitly sampled after they are grafted. Figure 3.3B shows the top-scoring models

superimposed on the crystal structure, showing the variation is localized to the H3 loop.

In contrast, for Ab08, the top scoring models diverge (Figure 3.3C), and the FRL RMSDs when superposing FRH are 3.28 Å, 0.90 Å and 2.99 Å for the three submitted models. Although 1MVU (α = -53.0°) was used for the initial orientation of Ab08 (α_{xtal} = -49.0°), the submitted models have packing angles ranging from -49° to -55°. Figure 3.3D shows the top scoring structures superposed on the crystal structure (white), showing that models 1 and 3 have a significantly different V_L–V_H orientation than the crystal. In this view it is clear these models would be difficult to use for additional simulations such as docking because the variation within the antigen binding site may prevent important atomic interactions from forming. Comparison of the hydrogen bonds of the H3 loop of the these models and the native structure reveals a possible explanation of the discrepancy (Figure 3.4). In the native structure there are two notable hydrogen bonds involving the H3 loop: (1) the side chain of the Ser H100A points back toward the V_H domain and forms a hydrogen bond with the backbone of the Tyr H98 within the H3 loop; and (2) the backbone of the Gly H99 forms a hydrogen bond with the side chain of the Trp L50 in the L2 loop. Conversely in the submitted models, the Ser H100A points toward the V_L domain, and in models 1 and 3, it forms a hydrogen bond with the backbone of Trp L50. In model 2, Ser H100A forms a hydrogen bond with the side chain of Trp L50. As a result, the V_L domain in models 1 and 3 is tilted back compared to the native structure, resulting in the increased RMSDs.

These issues can be classified into sampling and scoring problems. Sampling problems can occur both in the initial orientation template selection as well as in *de novo* H3 modeling, while scoring problems arise from favorable scores of non-native interactions

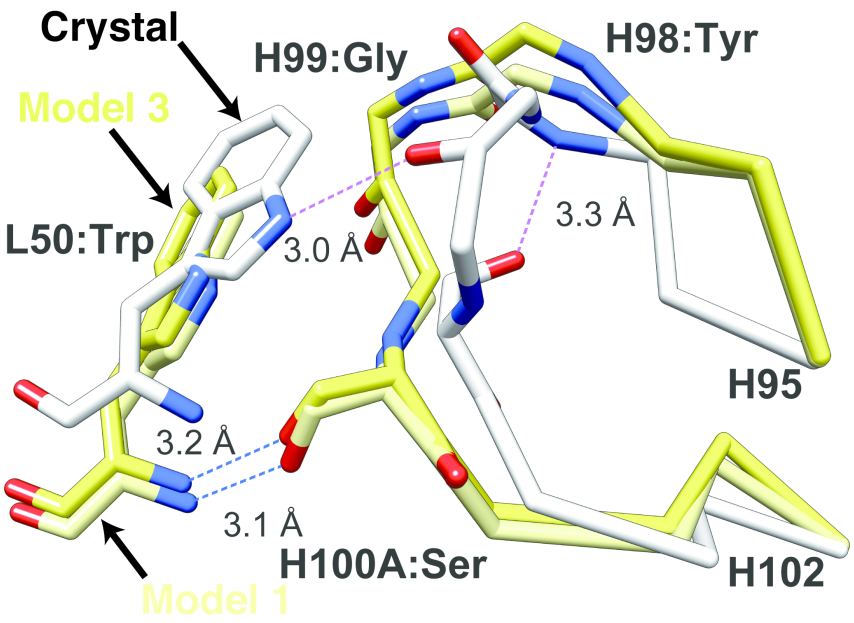

Figure 3.4: Non-native contacts arising from errors in V_L–V_H orientation lead to scoring complications. The crystal structure for Ab08 (white) forms a hydrogen bond between H99 Gly and the side chain of L50 Trp. However, two of the submitted models have an incorrect V_L–V_H orientation that is incompatible with this hydrogen bond and instead allows the side chain of Ser H100A to form a hydrogen bond with the backbone N of Trp L50. These non-native hydrogen bonds cause these structures to score favorably. Figure generated in UCSF Chimera.[101]

with the CDR H3 loop. These observations suggest a relationship between H3 conformation and the packing angle, which is discussed further below.

3.4.4 New loop modeling methods and constraints for CDR H3 prediction

The classic RosettaAntibody algorithm performs de novo CDR H3 modeling by inserting small fragments of residues from known structures followed by loop closure using the cyclic coordinate descent (CCD)[102,103] algorithm. Recently, new loop prediction algorithms have shown promising results. I updated RosettaAntibody to take advantage of two of these

new methods: Kinematic Closure (KIC)[88] and next-generation KIC (NGK).[89] KIC randomly perturbs several loop angles and then solves for the remaining six torsions of three 'pivot' residues to close the loop using a fast analytical formulation. NGK improved the KIC approach by incorporating annealing via ramping of van der Waals energy and Ramachandran potential weights, including neighbor-dependent Ramachandran propensities[104] and ω angle sampling.[105] Unfortunately the large neighbor-dependent Ramachandran propensity arrays exceeded the memory available on our standard supercomputing nodes, so I disabled this feature.

When building H3 on a homology framework (AMA II stage 1), I used KIC and NGK in conjunction with a constraint to favor kinked[24,26] C-terminal conformations. For CDR H3 conformations on the crystal environment (AMA II stage 2), I used NGK and KIC with and without kink constraints. I first summarize the results of modeling short H3 loops (8–10 residues) on a crystal framework, long H3 loops (11–14 residues) on a crystal framework, and then H3 loops on a homology framework.

MODELING SHORT H3 LOOPS (8–10 RESIDUES) IS MODERATELY ACCURATE (AB03/04/05/07/09/11)

I was able to build a model of all short CDR H3 loops with an RMSD < 2.0 Å for all targets except Ab11. Among the submitted models, the average loop RMSDs were 1.66 $\pm$ 0.96 Å and 1.58 $\pm$ 0.97 Å for the top-ranked and lowest-RMSD models, respectively (Table 3.3).

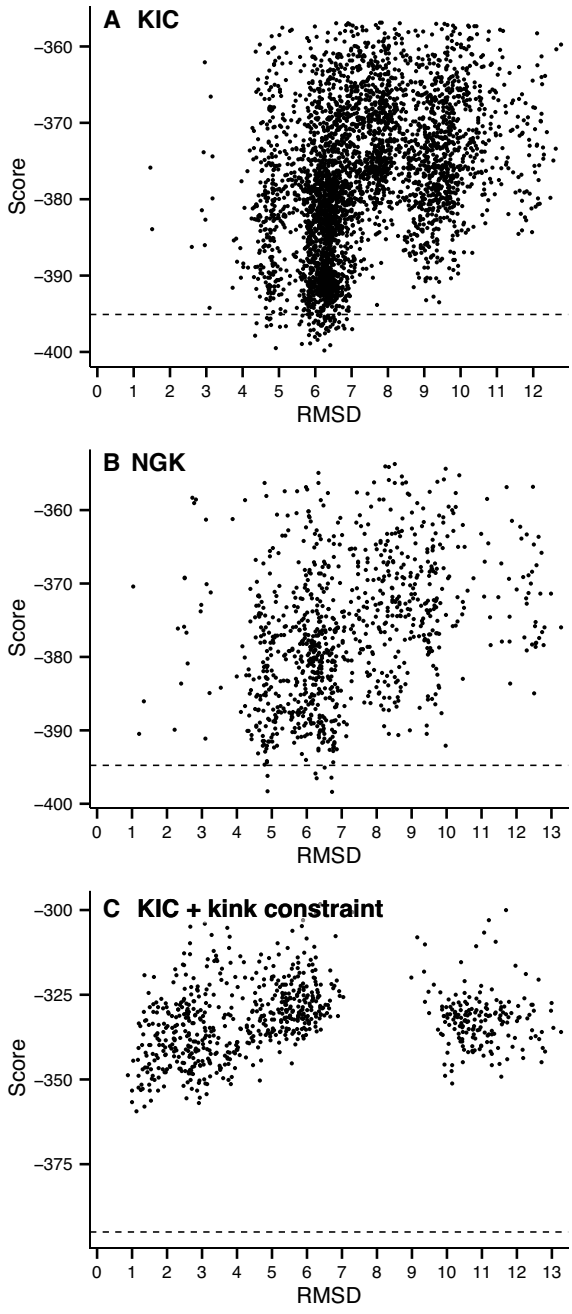

Figure 3.5: Score vs. RMSD plots for unconstrained de novo modeling the CDR H3 loop of Ab10 (A) shows that near-native conformations of CDR H3 are rarely sampled. Utilizing next-generation KIC (B) results in more near-native conformations sampled, but lowest scoring models are still far from the native structure. Including a constraint to prefer the C-terminal kink of the H3 loop (C) greatly improves the result.

Target	H3 length	Best scored (I)	Best RMSD (I)	Best scored (II)	Best RMSD (II)
Ab02	11	3.52	2.35	2.85	1.42
Ab03	8	2.48	2.31	1.82	1.36
Ab04	8	1.62	1.62	1.2	1.17
Ab05	8	3.02	2.92	1.86	1.86
Ab06	14	3.9	3.88	3.77	3.7
Ab07	8	1.27	1.27	0.68	0.68
Ab08	11	2.88	2.79	3.33	2.03
Ab09	10	1.75	0.92	1.02	1.02
Ab10	11	2.21	1.68	1.13	1.13
Ab11	10	3.3	0.91	3.39	3.39
Short H3s	8–10	2.24 ± 0.82	1.66 ± 0.81	1.66 ± 0.96	1.58 ± 0.97
Long H3s	11–14	3.13 ± 0.74	2.68 ± 0.92	2.77 ± 1.16	2.07 ± 1.15

Table 3.3: H3 RMSDs for top ranked and lowest RMSD models in stages I and II. All RMSDs reported in Å.

LONG H3 LOOPS (11–14 RESIDUES) BENEFIT FROM CONSTRAINTS (AB02/06/08/10)

For long CDR H3 loops built on the crystal frameworks, the lowest RMSD models have RMSDs of 2.07 ± 1.15 Å (Table 3.3). For insight, I examine the loop sampling and scoring through a plot of candidate structure score vs. distance from the native structure. Figure 3.5 compares A) unconstrained KIC, B) unconstrained NGK, and C) KIC using a kink constraint for Ab10 (11-residue CDR H3 loop). Although unconstrained KIC samples a couple conformations as low as 2 Å, those models score worse than other structures with RMSD $\sim$5.0 Å (Figure 3.5A). NGK samples more near-native conformations than unconstrained KIC, but some non-native structures still score better (Figure 3.5B). Enforcing the kink constraint drastically alters the results, with the best scoring model having an H3 RMSD of 1.1 Å (Figure 3.5C).

When using the kink constraint, the lowest-scoring structure scores approximately 40 units higher than the lowest-scoring structures when using unconstrained KIC or NGK

(Figure 3.5), suggesting that our choice of constraints is preventing formation of the lower-scoring near-native loop structures. Further, the constrained algorithm created a cluster of structures with RMSDs between 10 and 13 Å (Figure3.5C) that satisfy the constraints with a kink rotated into a conformation inconsistent with antibodies. Therefore, predictions might be further improved by a more precise constraint defining the kink (Chapter 5.

3.4.5 CDR H3 prediction on a homology framework can produce models with sub-Ångström accuracy

Figure 3.6 shows the RMSD of the closest-to-native model submitted by each group in stage 1 of the challenge for Ab02–Ab11. Our lab contributed the lowest-RMSD models for four targets (Ab02, Ab09, Ab10, Ab11), two of which have sub-Ångström RMSDs (Ab09, Ab11). For short (7–10 residue) CDR H3 loops, the average RMSD of the best submitted H3 model is 1.66 ± 0.81 Å, while for long (11–14 residue) H3 loops the average H3 RMSD of the lowest RMSD model is 2.68 ± 0.92 Å.

Even on a homology framework, RosettaAntibody built models of 4 of 6 short CDR H3 loops with an RMSD < 2.0 Å. For the two failures (Ab03 and Ab05), near-native H3 conformations were sampled (1.6 Å and 0.5 Å, respectively) but scored poorly, so they were not submitted. Retrospective analysis revealed that the poor scores of the models with near-native H3 conformations are due to the lack of low-RMSD templates for non-H3 CDRs (Ab03-H1–2, Ab05-L3) as discussed above. These failures in particular highlight the importance of accurately predicting all of the CDR conformations in order to model H3 successfully.

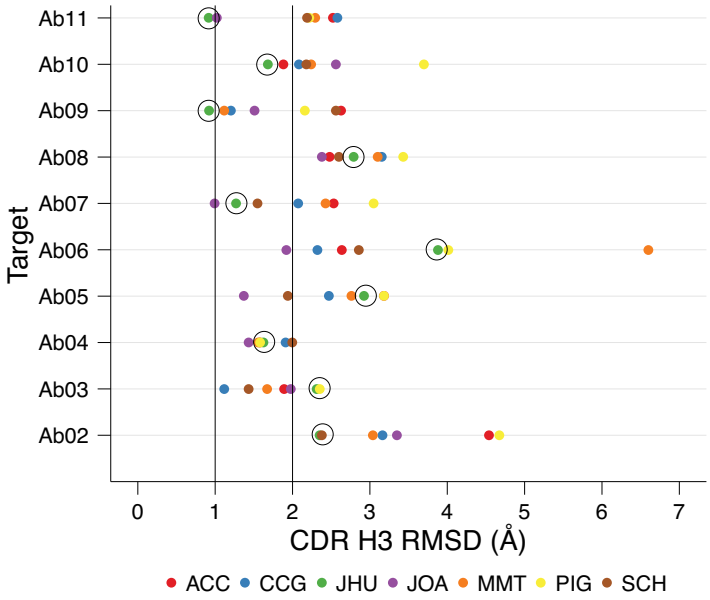

Figure 3.6: Modeling CDR H3 RMSD on a homology framework. The lowest-RMSD model for targets Ab02-Ab11 from each participant in AMA II stage I reveals the progress that has been made toward accurately modeling CDR H3. For four targets (Ab02, Ab09, Ab10, Ab11), RosettaAntibody (green plots, circled) produced the best CDR H3 models, two of which are sub-Ångström (Ab09, Ab11). Chemical Computing Group (blue plots) produced the best model for Ab03, and the collaboration between Astellas and Osaka University (purple plots) produced the best models for the remaining five targets.

3.4.6 Effect of V_L–V_H orientation on CDR H3 modeling

Even when the packing angle significantly deviates from the crystal structure, near-native CDR H3 conformations were sampled (*e.g.* candidate structures for Ab04 deviate from the crystal packing angle by as much as ∼10° yet still maintain an H3 RMSD of ∼1.0 Å). However, these structures do not score as low as those with a near-native packing angle and a sub-Ångström H3 RMSD. When such a decoy is produced, as it is for Ab05, the near-native decoys can clearly be distinguished from those with deviating packing angles. This suggests that important inter-chain atomic contacts are not present in the latter models and that correctly identifying the V_L–V_H orientation is a critical factor for model ranking.

3.5 Discussion and Conclusions

Computational antibody structure prediction algorithms have the potential to dramatically alter the development of new antibody products, including therapeutics. The performance of RosettaAntibody in AMA II demonstrates the progress made toward predicting atomically accurate models solely from a query sequence. This community-wide challenge provided me with the opportunity to test our knowledge of antibody sequences and structures with newly developed RosettaAntibody components and related methods of Rosetta 3.

An important lesson learned is the degree to which template availability is still a limiting factor. Attempting to predict Ab01, a rabbit antibody, resulted in abject failure due to the dearth of appropriate templates for the CDR loops and the fact that these algorithms require templates. Both phage display antibodies, Ab03 and Ab05, also proved difficult due to template availability. Ab05 prediction was complicated by the paucity of templates for λ light chains. In humans the populations of κ and λ light chains are almost equivalent, but κ antibodies are more abundant in the PDB since murine antibodies dominate the PDB and mice have predominantly κ light chains. λ light chains can have a longer CDR L3 loop than κ light chains,[106] and thus accurately predicting the L3 loop may prove to be a bottleneck for predicting and designing many human antibodies.[107]

Analysis of failures where RosettaAntibody did not select the best template in the database revealed that more sophisticated search criteria may need to be developed that include residues outside the target loop and use of all templates from crystal structures

with multiple copies in an asymmetric unit. Additionally, some templates retain poor MolProbity scores after refinement, indicating that *a priori* filtering of bad templates may improve model quality. Finally, in cases where templates are clearly not adequate (such as species not represented in the antibody database), *de novo* modeling might be used.

The scoring function used can also cause some systematic problems as evidenced by situations where energy minimization of the template caused deviations from the target crystal coordinates. This can result in an inaccurate model of non-H3 CDR loops even when the best template structure in the database is selected, and these deviations can lead to inaccuracies in the H3 modeling steps. Other sources of error in the H3 modeling stage stem from the infrequent sampling of the native-like conformations and, in some cases, the inability of the Rosetta score function to effectively discriminate native-like conformations from incorrect ones. Using a constraint to penalize non-kinked conformations results in significantly better sampling, and I am pursuing alternate kink constraint formulations, as described in Chapter 5.

The difficulty of CDR H3 loop prediction is demonstrated by Ab06, which has the longest H3 loop (14 residues; Kabat/Chothia definition) in the challenge set. Even when building the loop in the crystal environment, near-native models are sampled rarely. The difficulty of predicting long CDR H3 loops is problematic when considering that the average human CDR H3 length is 12 residues (Kabat/Chothia definition). The conformational space accessible by the large number of degrees of freedom in long loops remains the central challenge for *de novo* loop prediction for CDR H3 modeling. Accurately modeling the non-H3 CDR loops is critical to create the environment in which to model H3, so I believe

that continuing to improve our template-based modeling efforts is a necessary aspect of H3 modeling. These improvements may involve incorporating multiple templates, as well as metadata for each template to provide genetic information such as germline genes, species, and length and conformations of the other CDR loops in the parent structure.

Although incorrect V_L–V_H orientations do not preclude sampling of near-native CDR H3 conformations, the packing angle affects the score of the model such that the correct loop conformation cannot be recognized. Our current method for selecting an initial homology template for V_L–V_H orientation does not fare well when the native packing angle is far from the PDB average, nor is the packing angle adequately corrected during modeling. New approaches may require multiple V_L–V_H templates to capture a wider range of orientations, a packing angle constraint during modeling to direct orientation sampling, better V_L–V_H orientation predictions from sequence, or the development of statistical filters for the V_L–V_H interface.

A major weak point with RosettaAntibody models generated for AMA I was their poor MolProbity scores. By relaxing the structures after grafting the templates in AMA II, I was able to build models with MolProbity scores within the range of the crystal structures of the targets in the assessment. Because the differing target sequences, length of H3 loops and the number of models considered in AMA I and AMA II, it is difficult to directly gauge the difference of H3 modeling accuracy between AMA I and AMA II. However, there is a trend toward improvement in H3 accuracy. In AMA I, the average RMSD of the medium–long H3 (10–12 residue loops, 8 targets) was 3.3 ± 1.3 Å whereas that of the rank 1 models in the AMA II (10–11 residue loops, 5 targets) is 2.7 ± 0.8 Å. Notably, when considering the lowest

3 scored models in AMA II, the average H3 RMSD decreases to 1.7 $\pm$ 0.8 Å, highlighting the importance of using multiple models for further applications such as computational protein–protein docking,[31,108] design[109] and drug discovery.[110,111]

The object-oriented design in Rosetta 3 was critical during the challenge as it enabled me to quickly incorporate new modeling routines into RosettaAntibody. The continued interest in antibodies and rapidly increasing number of antibody crystal structures in the PDB contributes to the improvement of the method. RosettaAntibody 3 is available through the web server ROSIE (`http://rosie.rosettacommons.org/`).[112]

CHAPTER IV

THE ORIGIN OF CDR H3 STRUCTURAL DIVERSITY

Adapted from Weitzner BD, Dunbrack RL Jr & Gray JJ, "The origin of CDR H3 structural diversity," *Structure* 23(2), 302–11 Copyright 2015 Elsevier, Inc. Reproduced with permission.

4.1 Overview

Antibody complementarity-determining region (CDR) H3 loops are critical for adaptive immunological functions. Although the other five CDR loops adopt predictable canonical structures, H3 conformations have proven unclassifiable, other than an unusual C-terminal "kink" present in most antibodies. To determine why the majority of H3 loops are kinked and to learn whether non-antibody proteins have loop structures similar to those of H3, I searched a set of 15,679 high-quality non-antibody structures for regions geometrically similar to the residues immediately surrounding the loop. By incorporating the kink into the search, I identified 1,030 H3-like loops from 632 protein families. Some protein families, including PDZ domains, appear to use the identified region for recognition and binding. My results suggest that the kink is conserved in the immunoglobulin heavy chain fold because it disrupts the β-strand pairing at the base of the loop. Thus, the kink is a critical driver of the observed structural diversity in CDR H3.

4.2 Introduction

Structural diversity of antibodies is achieved through a highly coordinated, intricate process of genetic recombination and hypermutation through which a relatively small number of genes are able to produce antibodies against an immense array of pathogens. Antibodies consist of two pairs of heavy and light chains linked by disulfide bonds. The N-terminal domains of each chain compose the variable fragment (F_V). The F_V differs from antibody to antibody and contains the antigen-binding site, which is composed of three complementarity determining region (CDR) loops connecting β-strands from each of the two variable domains on a conserved framework.[10–12,113] Five of the CDR loops (L1–3, H1–2) form a limited number of distinct conformations, while the third CDR loop on the heavy chain (H3) has remained unclassifiable.[16–19] High structural conservation among antibodies makes it possible to model the framework and the five CDR loops that adopt canonical conformations, but the exceptionally diverse CDR H3 loop evades current methods, thus making structure prediction of the antigen binding region difficult.[23,40]

Because the F_V is highly conserved, antibodies are an ideal system for both library-based protein engineering techniques and computational protein structure prediction methods.[23,40,114–116] Library screening and directed evolution techniques have enabled the successful production of engineered antibodies used for sensors and assays as well as novel therapeutics.[4,52,117,118] However, discovery and development of such antibodies remains challenging. Because the CDR H3 loop is largely responsible for the diversity among antibody structures, it is typically critical to antigen binding. Indeed, studies analyz-

ing antibody–antigen complexes noted that CDR H3 was responsible for one third of the antigen-binding contacts and binding energy.[10,30] Increased understanding of the factors that govern CDR H3 conformations is vital to the continued development of engineered antibodies.

Because of their high-throughput and low cost, computational methods hold promise to decipher recently developed antibody sequence libraries obtained by high-throughput sequencing techniques[9,13,14] and usher in an era of rationally designed antibodies, but these methods require accurate antibody structure prediction, especially for CDR H3. To date, there have been several antibody structure prediction methods developed to begin to address this issue.[22,82,84] Most of these algorithms consist of three major steps: (1) identification of reasonable structural templates for the framework region and the five CDR loops that form canonical conformations; (2) assembly of these templates; and (3) *de novo* prediction of the H3 loop. The major source of error is the final step.[23,40]

The failure of *de novo* CDR H3 loop modeling is surprising in many cases because of the modest loop lengths at which they occur. It remains unclear why CDR H3 is such a challenging target for current loop modeling algorithms, but one possible explanation is that V(D)J recombination[7] can produce loops that access conformations that are extremely rare in existing protein structural databases. An alternate hypothesis is that the environment formed by the V_H and V_L domains stabilizes CDR H3 loop conformations that existing methods do not detect as favorable. In either scenario, loop modeling algorithms may not have been trained for, or proven capable of, predicting these structures.

The five non-H3 CDR loops can each be clustered into a small number of "canon-

ical" conformations for each loop length.[17,19] While CDR H3 loop structures cannot be described by such canonical conformations, the loop's C-terminus often contains an unusual "kink" or "bulge," with the remainder of the structures continuing the β-strand pairing into the loop ("extended"). I refer to these broad categories as having a kinked or extended base geometry. Several studies have been conducted to develop a framework to predict this kink's presence to aid structure prediction methods.[24–29] However, it was recently shown that the rules used for this prediction have not held up as the number of solved antibody structures has grown; the majority of structures contain the kink even when the sequence-based rules would classify the CDR H3 loop as extended.[19] More generally, rules intended to aid structure prediction of CDR H3 loops developed from structural analyses are complicated by the fact that the set of solved structures is not a representative set of antibodies.[69]

I recently participated in Antibody Modeling Assessment II (AMA II)[40] and found that Rosetta rarely samples kinked CDR H3 conformations unless I exploited a geometric kink constraint based on Shirai *et al.*'s description.[24,39] Other participants in AMA II[40,119] and the Web Antibody Modeling server[84] also use constraints to favor the kinked geometry. In contrast to antibodies, the available score functions prefer the extended base geometry.

In this study I investigate the physical and biological reasons for the majority of CDR H3 loops being kinked, and I determine whether or not the underlying genetic mechanism favors loops capable of adopting conformations not typically observed in non-antibody proteins. To accomplish this, I compared the geometry of the CDR H3 loop anchor regions (not including the residues involved in the kink) to all same-length segments in

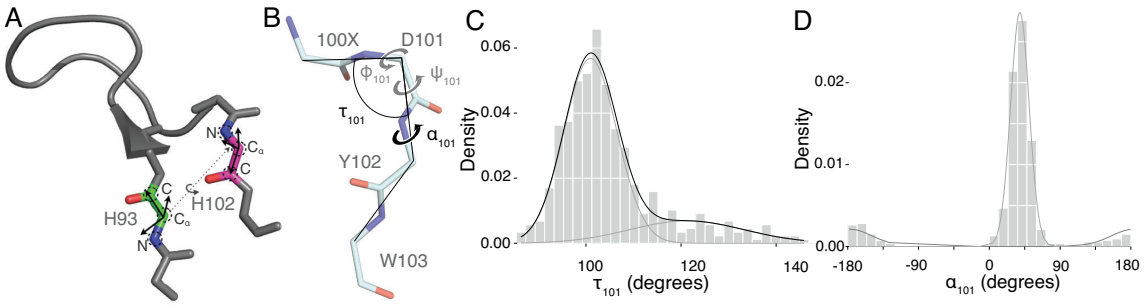

Figure 4.1: Loop anchor transform and C-terminal kink description. (A) An example CDR H3 loop showing the construction of the loop anchor transform (LAT). Coordinate frames (black) are constructed based on the backbone heavy atom coordinates (black dashed circles) of the N-terminal (green) and C-terminal (magenta) loop anchors. The six degrees of freedom (three translational and three rotational) required to perfectly superimpose the coordinate frames constitute the LAT and are represented as a dashed line connecting the coordinate frames. (B) Annotated antibody kink geometry showing the two angles I defined to describe the kink: (1) τ_{101} , the C_α–C_α–C_α pseudo bond angle for the three C-terminal residues in CDR H3 loops; and (2) α_{101}, the C_α–C_α–C_α–C_α pseudo dihedral angle for the three C-terminal residues in CDR H3 loops and one adjacent residue in the framework. (C) A histogram of τ_{101} reveals a skewed right distribution. A Gaussian mixture model fitted to the data with an expectation maximization algorithm showed the data can be partitioned into two states with roughly 80% of the data belonging to one distribution, centered at 101°. (D) A histogram of α_{101} is well represented by a two-state mixture model of von Mises distributions. Approximately 85% of the data lies in the distribution centered at 39°.

over 15,000 polypeptide chains. I found that a vast majority of the structures I identified adopted an extended strand-turn-strand conformation, but by incorporating the kink into the search criteria, I identified a diverse set of loops across a wide range of lengths. These loops show that the kinked conformation of CDR H3 loops is common and constitute a starting point for training new loop modeling routines or templates for antibody design. Moreover, my results suggest that the kink is a critical part of the immunoglobulin heavy chain fold that serves to disrupt the β-strand pairing at the base of the CDR H3 loop in order to create structural diversity among loops of the same length. Thus, I believe the C-terminal kink is a key component in generating CDR H3 structural diversity.

4.3 Results

4.3.1 Description of CDR H3 base geometry using a 3D transformation from the beginning to the end of the loop

I curated a set of 444 high-quality, non-redundant IgG heavy chains and a set of 15,769 high-quality, diverse chains from the Protein Data Bank (PDB).[15] For each heavy chain I computed the three-dimensional transformation between the backbone heavy atoms of the residue immediately preceding the conserved tryptophan after the CDR H3 loop (residue 102 using the Chothia numbering scheme[16]) and the residue immediately following the cysteine before the CDR H3 loop (residue number 93), and I stored the six degrees of freedom in a relational database for future analysis. I refer to these six parameters collectively as a Loop Anchor Transform (LAT). Figure 4.1A shows a CDR H3 loop with the relevant residues annotated. Similarly, I calculated the three-dimensional transformation for every 5 to 31-residue window in each chain in the non-antibody set (see 4.6).

The range of structural variation in the CDR H3 LATs is significantly more constrained than that of the non-antibody set from the PDB (Figure 4.2 shows 13-residue loops), which is a result of having selected H3 definitions extending to a structurally conserved position of the F_V (to facilitate comparisons among such loops). After confirming that the degrees of freedom have a negligible covariance and that the antibody LATs do not vary with length, I fitted a Gaussian distribution to each parameter of the LATs of all of the antibodies across all lengths. I then selected all regions from the PDB set with LAT parameters within 3.0 σ of the mean of each antibody degree of freedom, resulting in 45,940

matches.

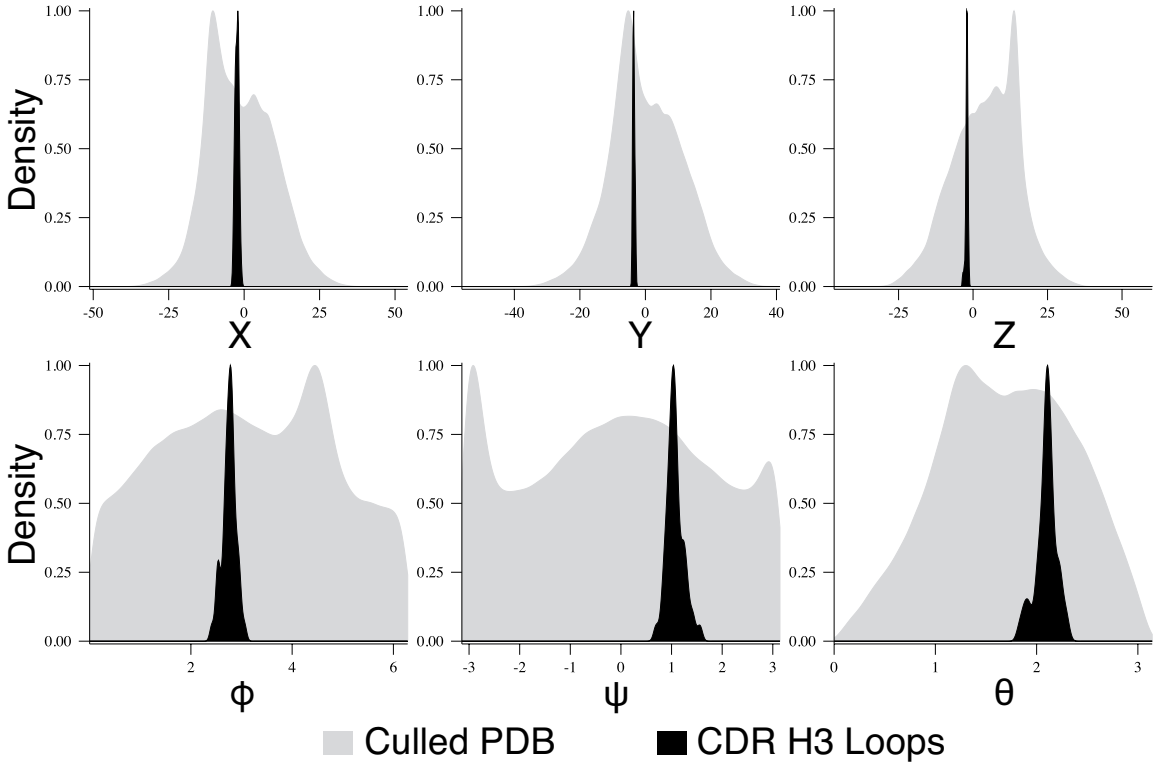

Figure 4.2: Density estimates for each of the six degrees of freedom of the loop anchor transform for each 13-residue segment from (1) the culled PDB set (gray); and (2) the known CDR H3 loops (black) show the relative structural diversity between the two sets. The tightness of the H3 distributions indicates the defined CDR H3 anchor points at structurally conserved positions, while the diffuse distributions from the PDB show the structural diversity of 13-residue segments in the other proteins. Because the six degrees of freedom are not covariant, each can be considered independent and modeled with a Gaussian distribution. These six Gaussian distributions are used to extract a set of non-antibody regions that match the span and orientation of the CDR H3 loop anchors.

4.3.2 Geometric parameters defining the C-terminal kink

I sought a quantitative description of the previously observed C-terminal kink.[19,24–26,28]

I first measured the pseudo bond angle of the C_α atoms of the three C-terminal residues (Chothia residue numbers 100x, 101, 102), termed τ_{101} based on the nomenclature introduced

by Levitt in 1976.[120] Shirai *et al.*[24] described the kink using ϑ_{base}, a pseudo dihedral angle of the C_α atoms from Chothia residue numbers 100x, 101, 102 and 103, which I will call α_{101} (Figure 4.1B). Figures 4.1C and 4.1D show the distribution of τ_{101} and α_{101} for the antibody set. The τ_{101} distribution is skewed right and can be accurately modeled as a mixture of two Gaussians, the larger of which encompasses roughly 80% of the data. Structural measurements and visual examination confirmed that the larger distribution is consistent with kinked or bulged structures. The peak of the smaller distribution is consistent with a β-strand or extended conformation. Thus, this parameter is effectively identifying the geometry of the kink.

Because α_{101} has density near 0° and $\pm$ 180°, I modeled it as a mixture of von Mises distributions[121] to account for the periodicity. Similar to the model for τ_{101}, the larger distribution represents about 85% of the structures, but unlike τ_{101}, the distributions constituting α_{101} have almost no overlap. Thus, these geometric parameters capture somewhat distinct structural features, and I sought to find a combination of the parameters that enables me to classify the base geometry of CDR H3 structures.

Previous sequence-based rules for predicting kinked vs. extended base geometries posit that these residues' ability to form hydrogen bonds at key positions is the underlying cause for the formation of the kink.[24–26,28,122] Specifically, the interactions that are considered are: (1) a salt bridge between the side chains of Arg94 and Asp101; (2) a backbone-backbone hydrogen bond between Arg94 and Asp101 that occurs in kinked structures but not in extended structures, where the hydrogen bond is between residues 94 and 102 (typically Tyr102); (3) a hydrogen bond between the Trp103 side-chain and residue 100x

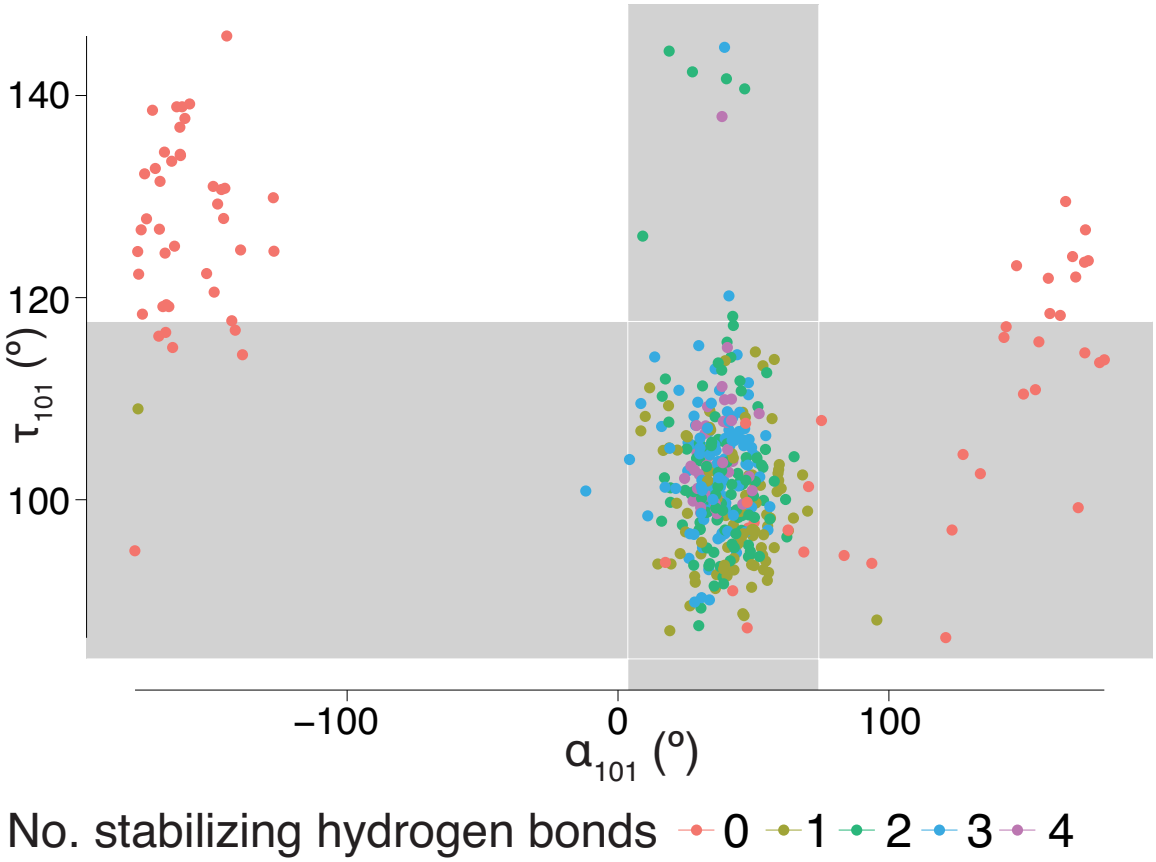

Figure 4.3: Scatterplot of τ_{101} vs. α_{101} for the antibody set. The gray, shaded regions represent $\pm$ 3.0 σ from the mean of the distribution presumed to represent the kinked subpopulation. Each point is colored by number of stabilizing hydrogen bonds in the structure. Although α_{101} is useful for isolating structures with these hydrogen bonds, there is a small subpopulation of well-hydrogen bonded structures with high values of τ_{101} ($\sim$140°), suggesting that neither τ_{101} nor α_{101} alone suffices to describe the kinked conformation. Structures in this region possess a β-bulge at position 101 but resume β-sheet strands C-terminal from the bulge.

(typically Phe100x) carbonyl oxygen; and (4) a second bulge that sometimes occurs further into the loop evidenced by a backbone-backbone hydrogen bond between residues 96 and the fourth residue before the conserved Trp at position 103. I refer to these four interactions as the stabilizing hydrogen bonds, and in Figure 4.3, I show a scatterplot of τ_{101} vs. α_{101} for the antibody set colored by the number of stabilizing hydrogen bonds. Overall there

is a strong correlation between a structure's τ_{101} and α_{101} values and the presence of the stabilizing hydrogen bonds, with the majority of the structures that deviate from the most common values having none of these interactions. However, there is a cluster of structures with τ_{101} and α_{101} values of roughly 140° and 30° respectively that form several of the aforementioned hydrogen bonds, demonstrating that these hydrogen bonds alone do not cause the H3 loop to adopt the typical kink formation.

Visual inspection of individual antibodies in the kinked, extended, and high-τ_{101} populations reveals the roles of α_{101} and τ_{101}. α_{101} positions the carbonyl group of residue 100x such that it lies in the plane of the base of the loop and points away from it. More generally, this parameter positions the kink relative to the framework of the antibody. τ_{101} is a measure of the degree to which the loop is kinked; if the loop is not kinked enough (large values), a strand pairing can still occur and if it is too kinked (small values) the stabilizing hydrogen bonds at the base of the loop are not disrupted. Thus, these two parameters describe the kink better when used together, and indicate that 79% of non-redundant antibodies in the PDB contain a kinked H3. Figure 4.4 shows the τ_{101} vs. α_{101} for the non-antibody loops set and reveals that the kink parameters describe a small subset of these structures.

4.3.3 CDR H3-like regions in non-antibody proteins

I constructed conformation logos—seqLogos made using the DSSP secondary structure assignments[123]—to compare the conformational diversity of sets of structures. Figures 4.5A and 4.5B show the conformation logos for all 12-residue H3 loops and all of the 12-residue structures from the PDB set with a LAT consistent with CDR H3 loops. The H3 loops begin

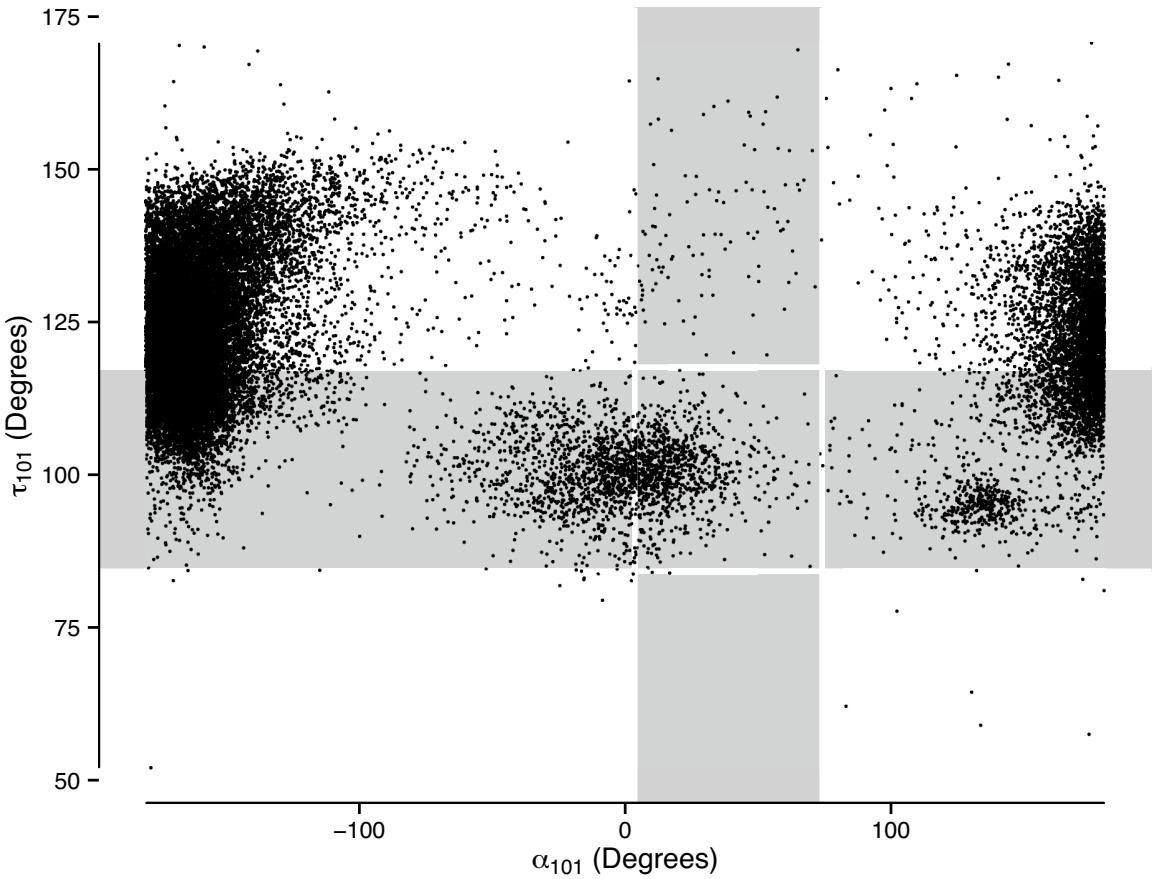

Figure 4.4: Scatterplot of τ_{101} vs. α_{101} for the LAT matches. The gray, shaded regions represent $\pm$ 3.0 σ from the mean of the τ_{101} and α_{101} distributions from antibodies. Unlike the antibody set, there are a considerable number of structures within the range of one of the parameters and not the other.

and end in an extended conformation, but are very diverse further into the loop, with a majority of structures having loop/coil, turn or 3_{10}-helix conformations at each position and very few residues adopting repeating secondary structure conformations (H or E). The set of non-antibody matches identified using the LAT alone does not resemble the CDR H3 loop set structurally, with the set of matches from the PDB consisting almost entirely of strand-turn-strand segments. This is not surprising considering the loop anchor residue locations are in paired β-strands. Because of this, many extended β-strand motifs lacking

long coil regions can match the LAT parameters.

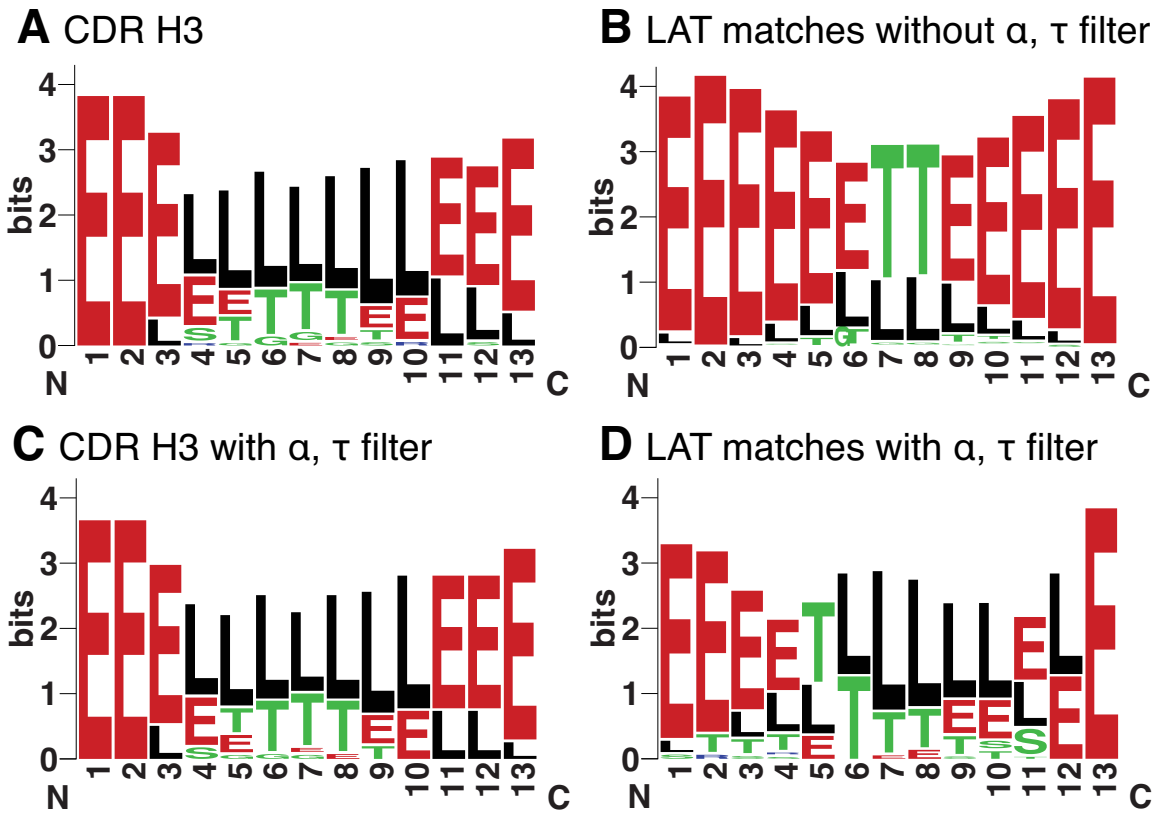

Figure 4.5: "Conformation Logos" for CDR H3 loops and LAT matches with and without a kink. WebLogo[124] was used with DSSP codes to produce a distribution of secondary structure elements in antibodies and the LAT matches using "E" for extended strand, "G" for 3_{10} helix, "H" for α-helix, "I" for π-helix, "T" for hydrogen bonded turn and "S" for bend, "R" for β-bridge and "L" (loop) for unassignable conformations. Using the LAT parameters alone to select the set of structures results in a set of antibodies with diverse conformations (A) and set of structures from the PDB that are largely consist of strand-turn-strand motifs (B). Including the additional constraint of the τ_{101} and α_{101} angles results in a set of LAT matches in the PDB that more closely resembles the distribution in antibodies (D), while the constraint has little effect on the antibody distribution (C).

Since the kink defined by α_{101} and τ_{101} is present in a large majority of CDR H3 structures, I restricted the search of the non-antibody structures to include only segments that have a C-terminal kink. The antibodies and the LAT matches from the PDB were filtered to remove structures with τ_{101} or α_{101} values beyond 3.0 σ of the mean of the distribution

associated with the kink (τ_{101} = 101° [σ = 5.6°] and α_{101} = 39° [σ = 11.8°]), which reduced the number of PDB LAT matches by roughly 90% (24,885 LAT matches to 2,207 LAT+kink matches). Figures 4.5C and 4.5D show the result of this filtering process. The conformation logo for the antibodies is nearly unchanged, while the results from the PDB display a very different conformation logo that is now very similar to the antibody set.

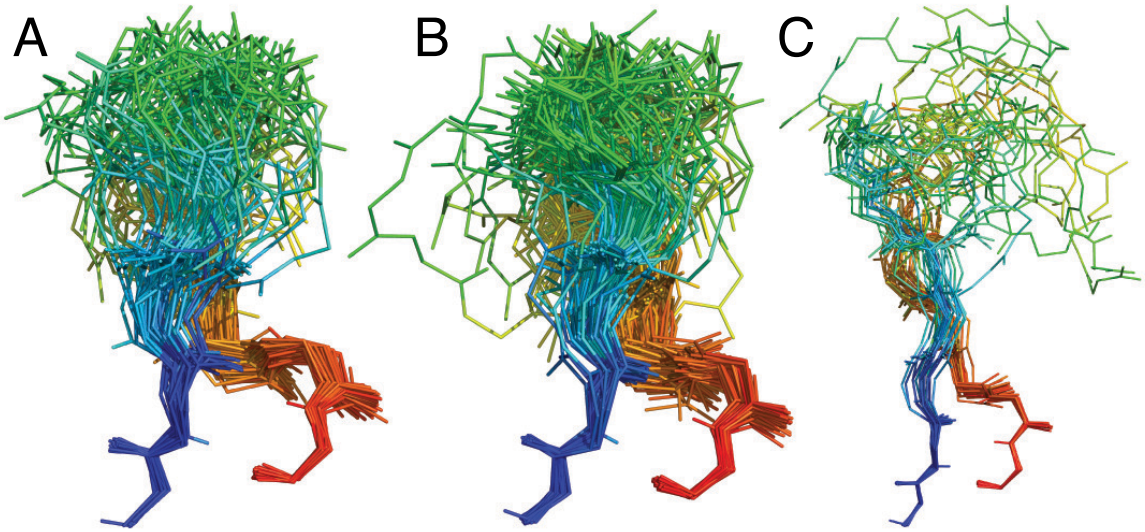

Figure 4.6: Comparison of CDR H3 and LAT+kink matches. Aligned, superimposed 12-residue CDR H3 loops (A) and 12-residue LAT+kink matches (B) show the similarity between the two sets of structures. The PDZ domain LAT+kink matches across all lengths (C) are included to show the diversity spanned by this particular Pfam alignment. The kink (red-orange) can be clearly seen and both sets occupy similar regions of space. Although some of the outliers may clash with the F_V framework, the PDB set could be included in a template-based H3 modeling algorithm.

4.3.4 Comparison of CDR H3 loops and loop anchor transform matches

Having a similar distribution of secondary structural elements does not mean the LAT matches are necessarily structurally similar to the CDR H3 loops. To illustrate the diversity

of the identified PDB segments, Figure 4.6 shows structures of 12-residue loops from the antibody H3 set (Figure 4.6A) and the 12-residue LAT+kink matches from the PDB (Figure 4.6B). The 12-residue segments were chosen for this visual comparison because they are the most common H3 loop length in the dataset. The C-terminal kink can be seen in both sets, and nearly all of the segments identified using the LAT and kink constraint appear to adopt a reasonable H3-like backbone conformation. To assess the degree to which the matches cover the structures of the H3 loops, I computed the root-mean-square deviation (RMSD) of the backbone heavy atom coordinates between the matches and the H3 loops. Figure 4.7 shows a cumulative density estimate of the lowest RMSD of a match to each CDR H3 loop. Approximately 10% of CDR H3 loops have a match within 1.0 Å RMSD, and 50% have a match within 2.0 Å RMSD, indicating that the LAT matches do in fact represent CDR H3-like conformations.

Although there are LAT+kink matches that are structurally similar to CDR H3 loops, it is not clear if they are more similar to CDR H3 loops than other CDR H3 loops. Figure 4.7 shows the cumulative density estimate of the minimum RMSD of an H3 loop to another H3 loop, and Figure 4.8 shows cumulative density estimates for each loop length being considered. I restricted the loop lengths to 9–20 residues and imposed a maximum sequence identity of 30% to prevent the comparison of different H3 loops that differ only by a small number of point mutations. Figure 4.7 shows that roughly 50% of CDR H3 loops are within 1.9 Å RMSD of another CDR H3 loop across all lengths. This may be compared with a figure of 2.1 Å for comparison of H3 structures with LAT+kink matches (Figure 4.7). In order to assess the degree to which the kink factors into selecting close

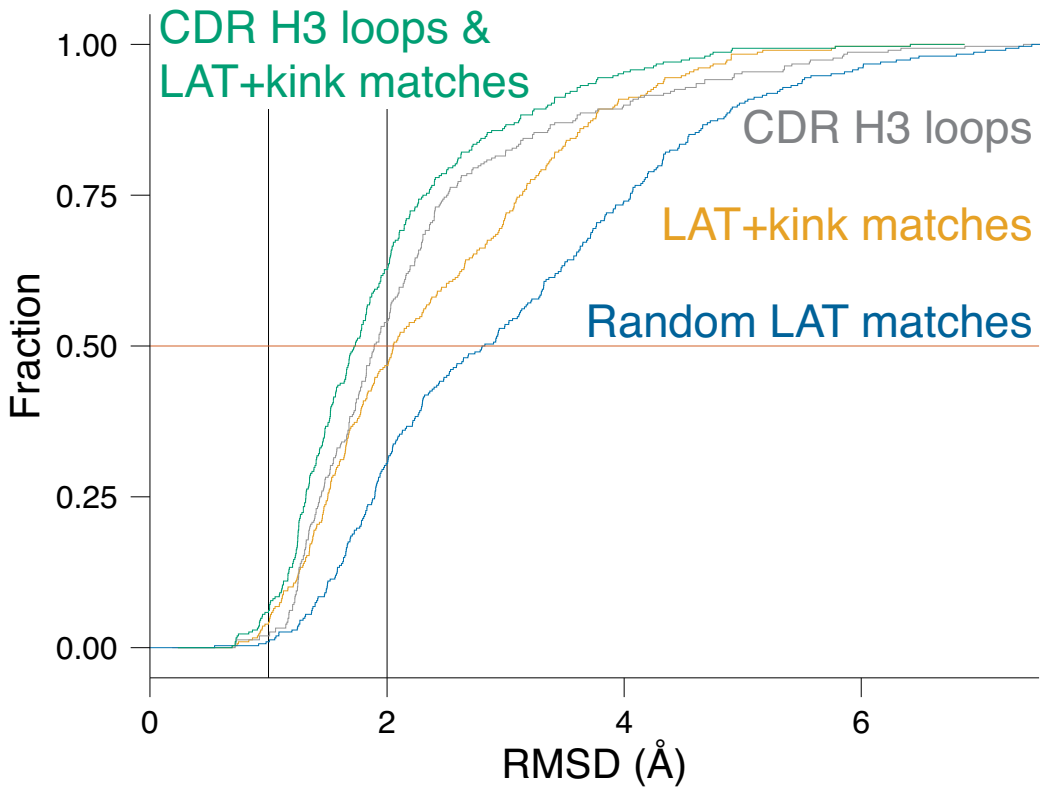

Figure 4.7: Structural similarity of CDR H3 and LAT+kink matches. A cumulative density estimate of the lowest root-mean-square deviation (RMSD) of backbone atomic coordinates of each H3 loop relative to all other H3 loops with a maximum sequence identity of 30% (gray curve), the minimum RMSD of any LAT+kink match relative to each antibody CDR H3 loop (yellow curve) and a random set of LAT matches of the same size and length distribution as the LAT+kink matches (blue curve). The green curve is a cumulative density estimate of the combination of the CDR H3 and LAT+kink sets. Comparisons were limited to H3 loops of 9–20 residues in length (296 H3 loops) to avoid kinematic constraints in loop conformations and to ensure there were a sufficient number of reference CDR H3 structures. Dashed vertical lines at 1.0 and 2.0 Å indicate the frequency of finding a PDB segment that closely matches a known CDR H3 loop conformation. The red dashed line shows that for 50% of H3 loops from length 9–20, there is a structure from the LAT+kink set under 2.1 Å RMSD, and within CDR H3 loops, there is a match within 1.9 Å RMSD in contrast to the 2.8 Å RMSD that would be expected from a set of random loops. Using the combined H3 and LAT+kink set results in the lowest RMSDs overall.

structural matches, I constructed a set of random LAT matches of the same size and length distribution as the set of LAT+kink matches. In Figure 4.7, the blue curve shows that 50% of CDR H3 loops are within 2.8 Å RMSD of random loops, indicating that requiring the presence of the kink greatly improves the structural similarity to CDR H3 loops. Figure 4.8 shows that this relationship is strongly related to the length of the loop being examined. The distribution begins to shift dramatically when the length of the CDR H3 loop exceeds 12 residues. The reasons for this are twofold: (1) longer loops have access to a significantly larger conformational space; and (2) there are fewer solved structures of longer CDR H3 loops. This result shows that a template-based CDR H3 loop modeling routine using only other known CDR H3 loops is unlikely to be successful for long loops. To gain insight on how the LAT+kink matches may lead to improvements in CDR H3 structure prediction, I also include a cumulative density estimate for the combined set of CDR H3 loops and LAT+kink matches (green curve), which shows that identifying templates from non-antibody proteins provides a path to obtaining a set of useful templates for longer CDR H3 loops. The combined set contains more low-RMSD structures than the CDR H3 set or LAT+kink set alone, with 50% of CDR H3 having a match with RMSD $\leq$ 1.7 Å.

4.3.5 Summary of loop anchor transform matches

To assess the degeneracy of the non-antibody LAT matches, I examined the proteins and protein families from which they originate. To determine whether matches originated in similar positions of homologous proteins, I assigned each matching chain a Pfam chain architecture[125,126] and recorded the positions within the Pfam alignments[127] for each LAT match. Table 4.1 compares the number of LAT matches to the number of H3 loops as

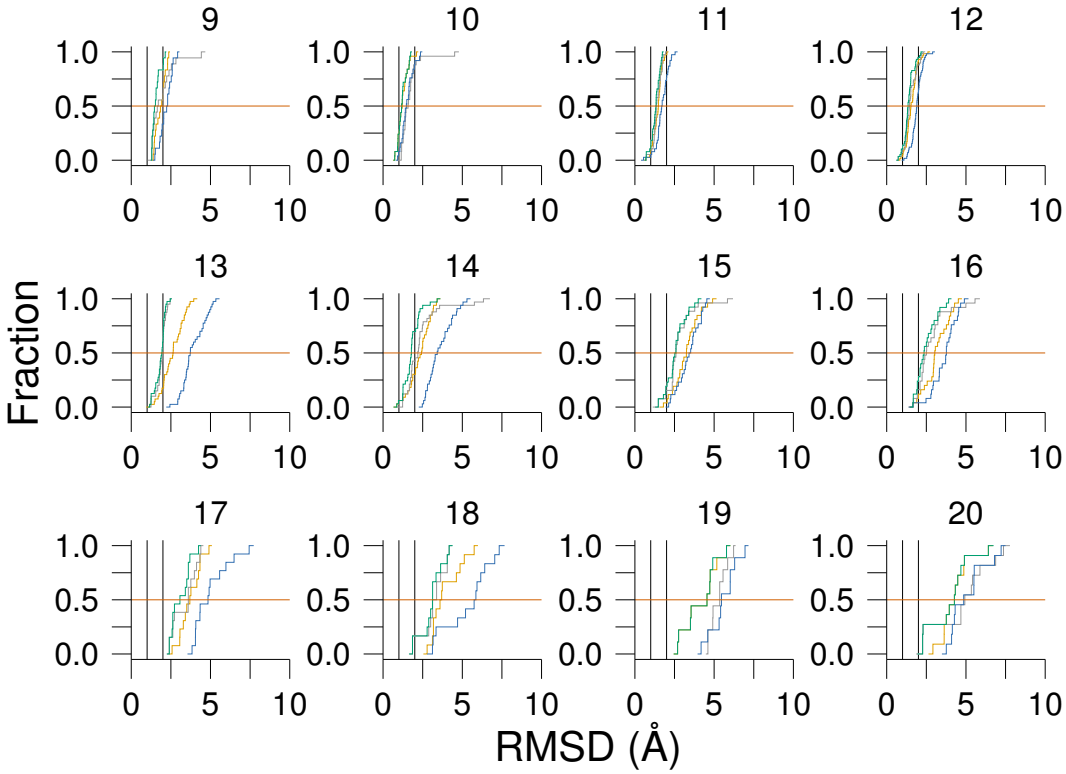

Figure 4.8: Cumulative density estimate for minimum root-mean-square deviation (RMSD) of backbone atomic coordinates for each H3 loop relative to all other H3 loops ranging from 9–20 residues with a maximum sequence identity (SID) of 30% (gray curve) and the minimum RMSD of any LAT+kink match relative to each antibody CDR H3 loop (yellow curve), a random set of LAT matches of the same size and length distribution as the LAT+kink matches (blue curve) and the union of the H3 loop and LAT+kink match sets (green curve) split up by length. Dashed vertical lines at 1.0 and 2.0 Å indicate the fraction of structurally similar H3 loops at each cutoff. As the loop length increases, so does the distribution of minimum RMSDs, showing that longer H3 loops are more diverse. These data underscore the difficulty of developing a template-based modeling method for long H3 loops and show that using the LAT+kink matches improves the coverage of known CDR H3 loops.

well as the number of unique Pfam alignments at each length. Whether the LAT matches are broken down by length or taken as a whole, nearly all of the LAT matches originate from a unique match position in a Pfam hidden Markov model. However, when multiple

matches originate from the same Pfam, they nearly always align to the same positions in the Pfam, indicating that antibodies are not the only proteins to select for loop structures with the C-terminal kink. There are more than three times as many non-antibody loops as H3 loops with kinked base geometry, with over 200 matches at very long loop lengths ($\geq$ 20 residues). The complete list of matches and their Pfams is available as a downloadable text file (Appendix A). An example of how to use this file to generate a set of coordinates is provided in 4.7.

Seven percent of the matches do not align to any Pfam, indicating that the match does not originate from a structurally conserved region of the protein or that it is beyond the bounds of the Pfam domain definition. Forty percent of the matches align to a Pfam, but this Pfam alignment only occurs once. The remaining 53% come from repeated alignments to the same Pfam, with the most common Pfam alignments being PDZ (23 matches) and peptidase C1 (17 matches). Appendix B contains a list of all of the Pfams that occur more than once, and lists the number of LAT+kink matches, the number of unique alignment positions as well as the corresponding tags from the Gene Ontology server.[128]

Figure 4.6C shows the PDZ LAT+kink matches. The N-terminal strand of the kinked loop forms an anti-parallel β-sheet pairing with the C-terminus of PDZ substrate proteins and, along with a conserved helix, forms the binding region of PDZ domains.[132] Several structures of PDZ domains in complex with their binding partners confirm that this CDR H3-like region is involved in binding (Figure 4.9). In the case of heterodimeric protein substrates (*i.e.*, not peptide substrates or homodimers), residues in the loop region of the kinked H3-like anchor segments are directly involved in domain-domain interactions with

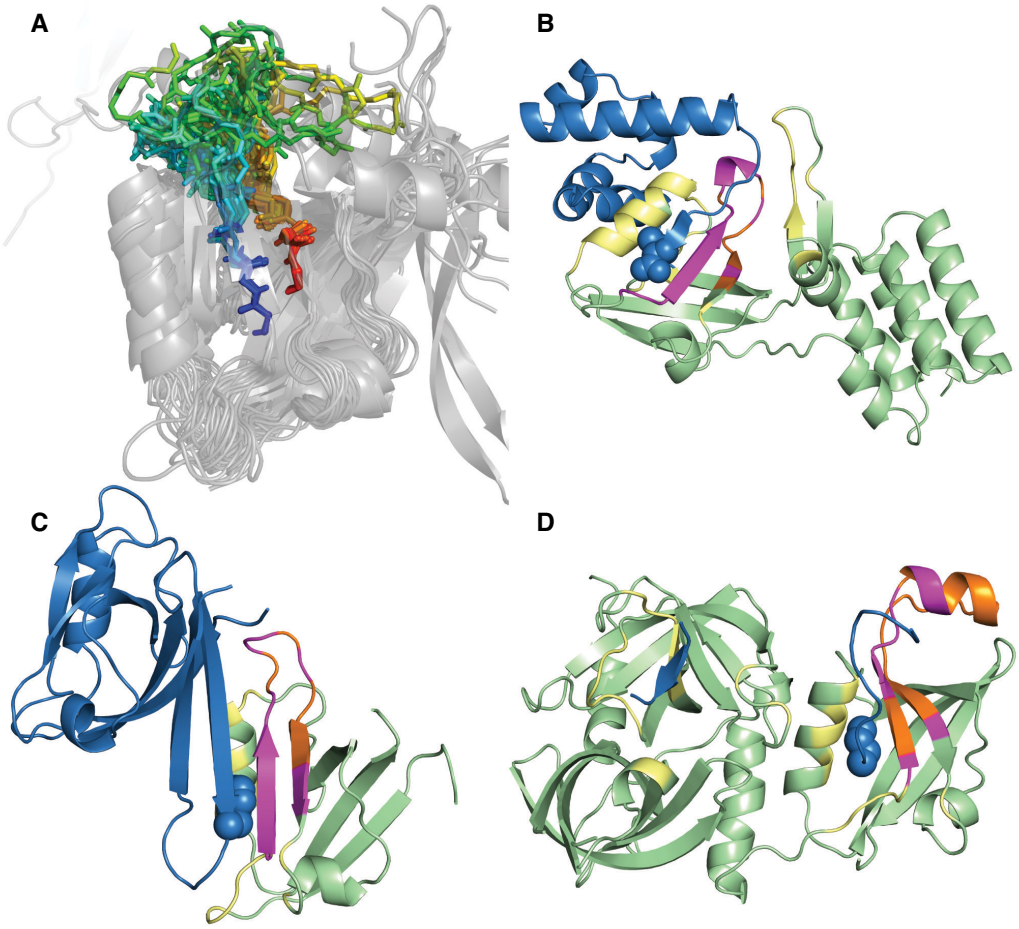

Figure 4.9: PDZ domains interacting with substrates through a kinked loop. (A) Superposition of PDZ domains with LAT+kink matches shows that the kink is in a structurally conserved position. The matching region is colored in rainbow with blue at the N-terminus and red at the C-terminus of the loop. The structural diversity of the identified loop is on display. I searched the PDB for PDZ–protein substrate heterodimers and found examples of the matching loop being involved in binding: (B) the N-terminal PDZ domain of harmonic in complex with Usher syndrome type-1G protein (3k1r)[129] (C) Alpha-1 Syntrophin (PDZ containing) in complex with neuronal nitric oxide synthase (1qav)[130] (D) Periplasmic serine endoprotease DegP (PDZ containing) in complex with lysozyme C (3otp).[131] In this view, the substrate is blue with the C-terminal residues shown in spheres and the PDZ containing chain is pale green. The matching loop is shown in orange and all contacts between the substrate and the loop are shown in magenta. Other contacts within 5.0 Å between the PDZ domain and the substrate are colored yellow.

Length	CDR H3 Loops	LAT+kink Matches	Unique Pfams
9	18	27	18
10	24	221	131
11	34	143	103
12	58	123	80
13	40	25	19
14	32	72	58
15	26	49	35
16	24	57	48
17	11	34	27
18	12	26	23
19	9	22	21
20	8	32	25
>20	13	199	118
Total	309	1030	632

Table 4.1: Number of CDR H3 loops, LAT matches and unique Pfam alignments at each loop length. Because I am using alignments to a consensus sequence for each Pfam, matches of different lengths can have the same Pfam description. Note that the total number of unique Pfams is not the sum of the number of unique Pfams broken down by length.

the substrates. Interestingly, the matching regions in both PDZ and peptidase C1 domains appear to be involved in recognition and/or binding. Thus, C-terminal kinks are present in a wide variety of non-antibody proteins, and some other protein domain families use this feature for binding and selectivity in the same way as antibodies.

Using my description of the kink, I tested the predictive power of the identity of the base residues at positions 94 and 101, which are frequently Arg and Asp respectively in antibodies. Table 4.2 shows the percentage of kinked CDR H3 loops with all combinations of the presence or absence of the supposed stabilizing base residues. In agreement with North *et al.*, who used a Ramachandran-based criterion for identifying the kink, I find that the majority of CDR H3 loops are kinked even when none of these residues are present. I also applied the rules developed in a study by Kuroda *et al.*,[26] which constitutes the

	No. Abs	% Kinked
R/K and D both present	228	91.2
R/K or D present	111	73.9
R/K present; D absent	68	85.3
R/K absent; D present	43	55.8
R/K and D absent	45	60.0

Table 4.2: The number of and percentage of antibodies that are kinked for all combinations of residues implicated in kink formation. As found earlier,[19] regardless of the presence of the stabilizing residues, the majority of CDR H3 loops are kinked.

most detailed analysis of explicit interactions among the H3-base residues, residues within the kink, and tertiary interactions with light chain residues (Table 4.3). The accuracy of these rules is 88.9%, which agrees with the published value of 89%. However, when one classification dominates a population, balanced accuracy (BACC) is a more meaningful measurement of the performance of a model.[133] While 94.2% of kinked structures are correctly predicted, only 46.2% of extended structures are identified as such, which results in a balanced accuracy of 70.3%. Because the percentage of correctly predicted extended structures is less than 50%, I conclude that the sequence-based rules do not fully explain the presence or absence of the kink.

Additionally, I examined the flanking regions of the LAT and LAT+kink matches and found that the LAT effectively constrains the environment to a β-strand scaffold (Figure 4.12). I investigated the CDR H3-like non-antibody loops for the presence of these stabilizing residues and observed neither the Arg–Asp combination nor the tryptophan at the equivalent of position 103. In fact, the sequences of the LAT matches and the LAT+kink matches do not show any preferences at the base of the loops that would explain the presence or absence of the kink (Figure 4.13).

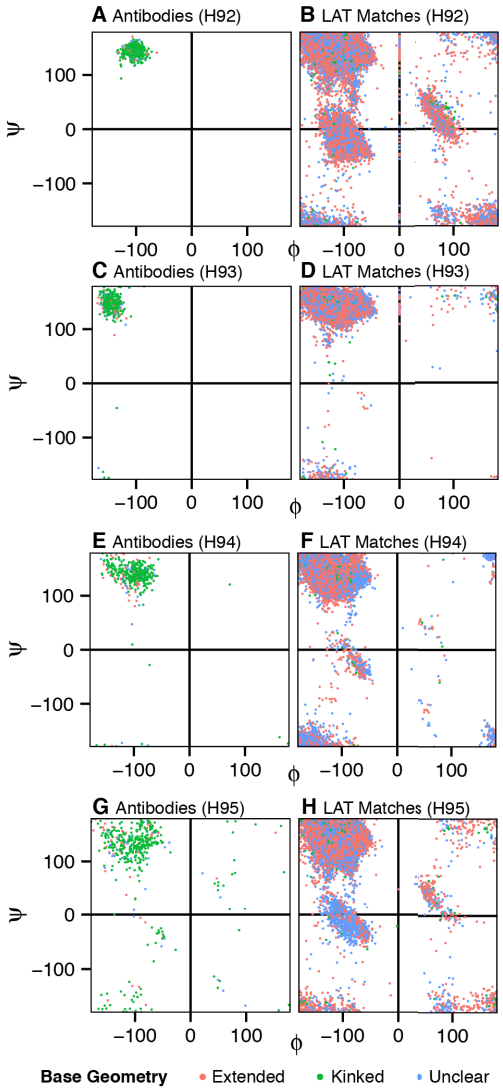

Figure 4.10: Ramachandran plots for four N-terminal residues of H3 loops (left) and LAT matches (right), beginning with the residue preceding the N-terminal loop anchor. In antibodies (A) this residue is the Cysteine that precedes H3 loops and is structural conserved, but in the LAT matches (B), this residue is free to adopt an extremely wide variety of conformations. At the H93 (or equivalent) position, the antibodies (C) are again structurally conserved, but now the LAT matches (D) are almost entirely restricted to the β region of the plot because this is the anchor residue that I use to identify LAT matches. The next two positions show the antibodies (E and G) beginning to broaden as further into the loop and the LAT matches (F and H) return to an extremely diverse set of conformations.

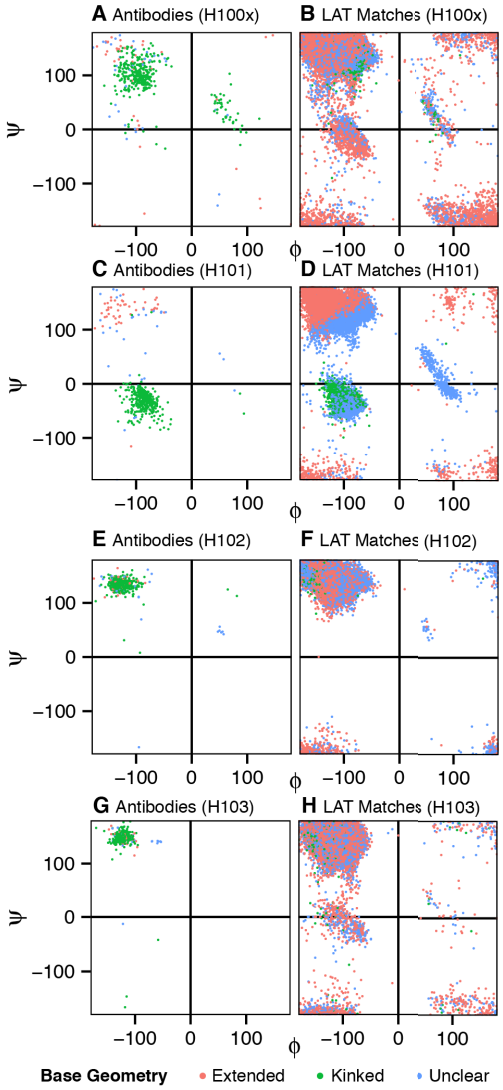

Figure 4.11: Ramachandran plots for four C-terminal residues of H3 loops (left) and LAT matches (right), ending with the residue following the C-terminal loop anchor. At the position furthest into the loop, the antibodies (A) and the LAT matches (B) both show conformational diversity. However, the penultimate residue shows the antibodies (C) and the LAT matches (D) confined to different regions of Ramachandran space based on the loop base geometry. Extended loops lie in the β region while kinked loops are in the α conformation. Interestingly, loops with indeterminate base geometry lie in the β, α and the L_α regions, showing that dihedral angles alone cannot be used to classify base geometry. Position H102 is the C-terminal loop anchor, and, as with the N-terminal loop anchor, most structures are confined to the β region of the plot. After the loop ends, the antibodies (G) exhibit almost no conformational diversity while the LAT matches (H), while biased toward β have density in a much larger region of the plot.

Metric	Description	Value
True Positive Rate	% of extended structures that are correctly predicted	46.5%
True Negative Rate	% of kinked structures that are correctly predicted	94.2%
Positive Predictive Value	% of predicted-extended structures that are extended	50.0%
Negative Predictive Value	% of predicted-kinked structures that are kinked	93.4%
Accuracy	% of correct predictions	88.9%
Balanced Accuracy	Expected accuracy on a balanced dataset	70.3%
Matthews Correlation Coeff.	Balanced correlation ranging from -1 to + 1	0.42

Table 4.3: Statistical analysis of H3-classification rules proposed by Kuroda *et al.*[26] Because the majority of structures are kinked, predicting an extended structure is considered a positive prediction. In my analysis I have a third base geometry: unclear. Those structures are excluded from this calculation to ensure that only structures where both analyses agree on how the base geometry should be classified are considered. Also left out are 23 structures that only have a heavy chain in the crystal (either because only the F_V was crystallized or because it is the structure of a V_HH) and require using light chain residues to apply the rules. Within this group, 20 structures are kinked and the other 3 are extended, but I cannot apply the rules to these structures so they are not included in the analysis below.

4.3.6 Conformation of base residues in CDR H3 loops

North *et al.*[19] proposed an alternate description of the kink based on the conformation of residue 101 and 102. If residue 101 is in the β region of the Ramachandran plot, the CDR H3 loop is considered extended, whereas if residue 101 is in the α region, it is deemed kinked. Residue Trp 102 is in the βregion in both cases. Based on this description, North *et al.* reported that regardless of the identity of the base residues at positions 94 and 101, the majority of CDR H3 loops are kinked.

To test this description, I generated Ramachandran plots for the terminal residues of my set of CDR H3 loops and LAT matches. Figures 4.10 and 4.11 show Ramachandran plots for the N and C-terminal loop residues including an additional residue on each side of the loop of both antibodies and the LAT matches. The antibodies show much less diversity at the anchor points and beyond the loop (Figures 4.10A,C and 4.11E,G) than the LAT

matches (Figures 4.10B,D and 4.11F,H), but within the loops there is considerable diversity. Only at the penultimate residue (residue 101) is there a clear distinction that can be made between loops with kinked and extended base geometries. Loops with "unclear" base geometries, those that match only τ_{101} or α_{101} and not both, populate all regions of the Ramachandran map at residue 101 (or its equivalent) in both antibodies and LAT matches (Figure 4.11C,D). From these plots I conclude that the dihedral angles for the residue at position 101 (or equivalent) are not sufficient to classify loop base geometries and that different base geometries are not confined to specific regions of Ramachandran space for the other residues.

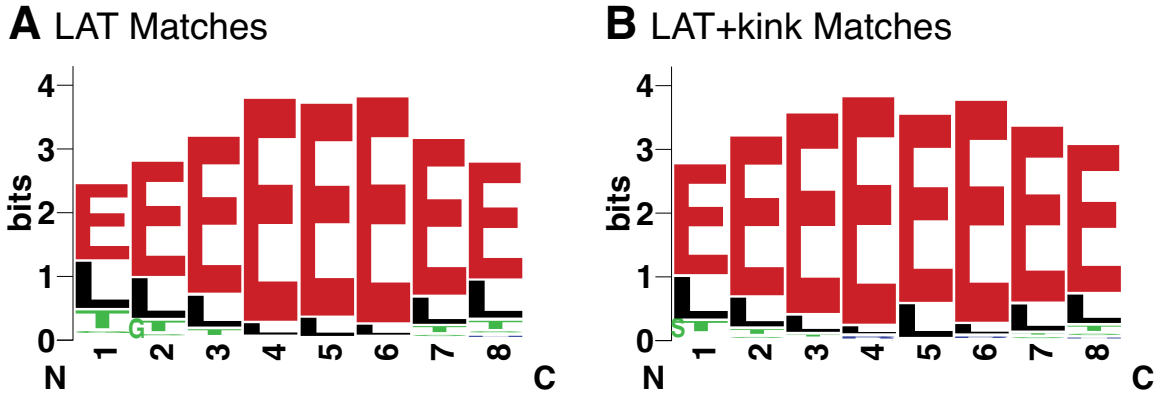

Figure 4.12: "Conformation Logos" for the loop anchors and three flanking residues on either side for (A) all LAT matches, and (B) LAT+kink matches. In both cases, positions 1, 2 and 3 correspond to N - 3, N - 2 and N - 1; position 4 is the N-terminal loop anchor; position 5 is the C-terminal loop anchor; and positions 6, 7 and 8 correspond to C + 1, C + 2 and C + 3. Approximately 84% of LAT match environments and 90% of LAT+kink match environments are extended β strands. The distributions are nearly identical in LAT matches and LAT+kink matches.

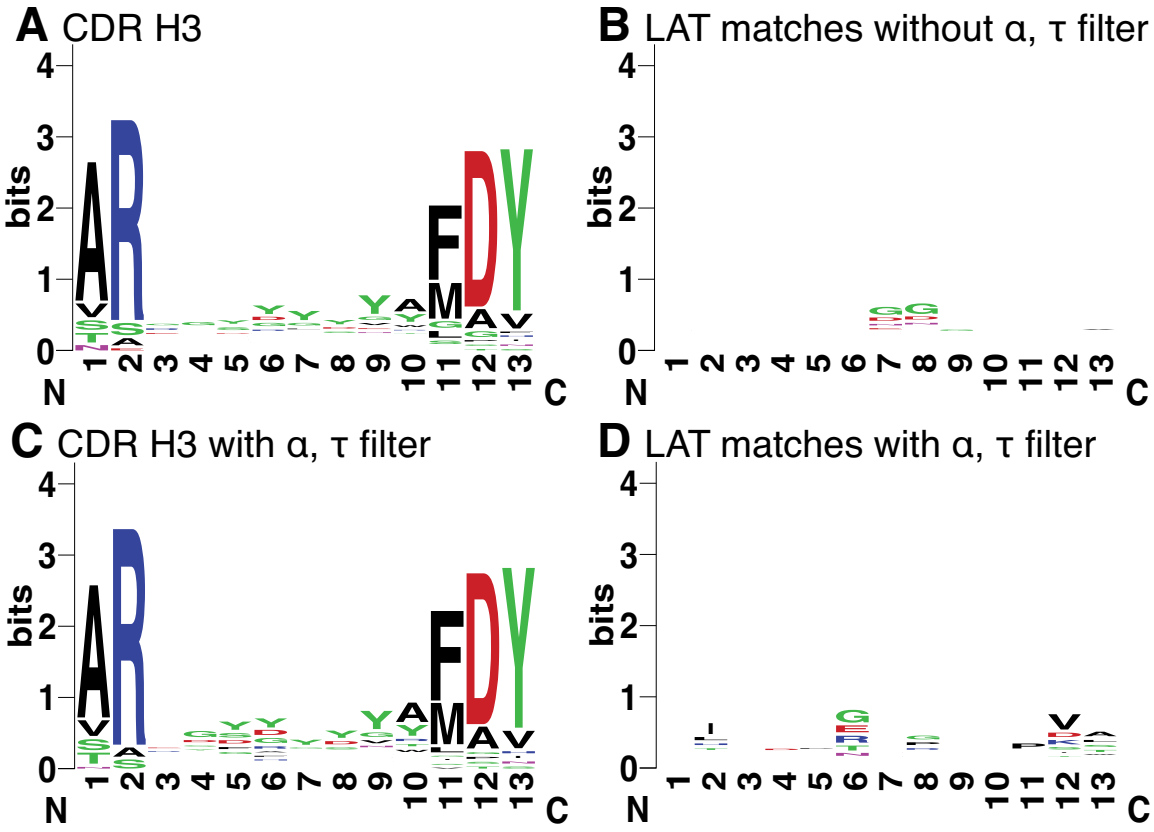

Figure 4.13: Sequence Logos for CDR H3 loops and LAT matches. Using the LAT parameters alone to select the set of structures results in a set of antibodies with clear sequence conservation at the termini and very little conservation in the central positions (A) and set of structures from the PDB that have no clear sequence preferences save for a small Glycine signal in the central-most positions (B). Including the additional constraint of the τ_{101} and α_{101} angles does not alter the antibodies (C), while the constraint has a small, but noticeable effect on the PDB sequence distribution (D). However, there is still no detectable signal and the residues typically associated with kink formation in antibodies are not observed in the LAT+kink set.

4.3.7 The effect of loop apex glycine residues on base geometry

Extended CDR H3 conformations often consist of a continuation of the β-strands at the base of the loop. As shown in Figure 4.5, this extended conformation is much more common than a kinked base geometry in most proteins. It has been established that β-strands are geometrically compatible with the "mirror image" turn types (types I′ and II′) that strongly prefer glycine in central positions.[134,135] Since all of the loops must change direction to maintain a continuous backbone, nearly all of them contain at least one β-turn, but the position of the β-turn may be restricted by the base geometry. Thus, I hypothesized that glycine in a central position may be indicative of an extended conformation. The effect of glycine position in extended CDR H3 loops has been incorporated into previous CDR H3 classification rules,[24–26] but the predictive significance of glycine in central positions has not been investigated.

The paucity of extended H3 loop structures demands that I analyze all loop-lengths simultaneously, but I restricted the loops to be 9–20 residues in length to remove geometric loop continuity constraints. I recorded the position of every glycine in all of the H3 loops and the LAT matches scaled from 0–1, with each count placed at the center of the bin. To account for possible register shifts in H3 loops with an even vs. odd number of residues, I treated them as separate groups. In this scheme, shorter loops have wider bins.

Figure 4.14 shows density estimates of glycine position for even and odd-length H3 loops and LAT matches split up by base geometry.. In the H3 loops, glycine residues are favored more on the N-terminal side of the loop in kinked structures while for structures with extended or unclear base geometries a more central position is preferred. The unclear

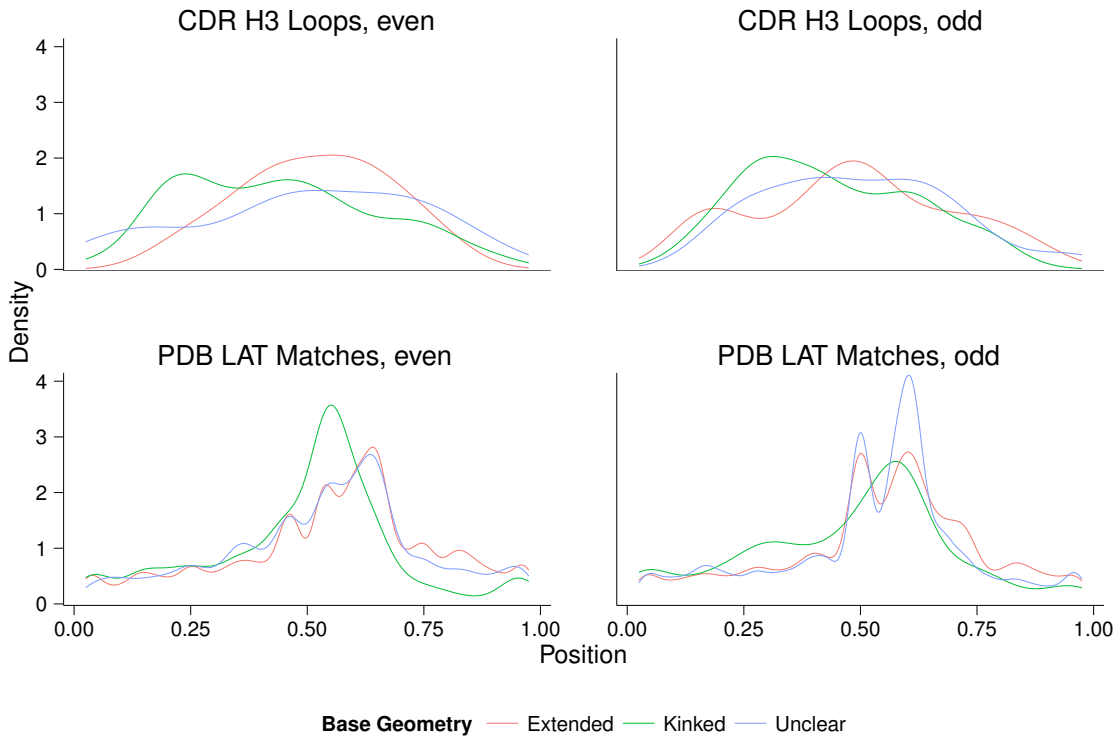

Figure 4.14: Density estimates for glycine positions in CDR H3 loops and LAT matches. The data are split up by base geometries using the τ_{101} and α_{101} values for the loop. The data are further split up by the parity of the loop length because defining a central position differs between the two sets.

structures appear to have a much broader distribution than either kinked or extended structures and this is likely because of the small sample size (13 even-length loops, 13 odd-length loops) and because the unclear base geometries may have a glycine position distribution somewhere between kinked and extended structures. In the LAT matches, the structures with unclear and extended base geometries have very similar glycine distributions, with the peak of the distribution biased toward the C-terminal end. In the kinked structures, the preference for glycine is also slightly biased toward the C-terminal end, but is more central than for the other base geometries. The difference between the kinked and unclear struc-

tures provides additional evidence that using both τ_{101} and α_{101} to define the kinked base geometry identifies a subset of structures that is distinct from the extended conformations. That the PDB LAT matches prefer glycine in more central positions for kinked structures is likely a consequence of averaging the result of disparate evolutionary pathways. This result supports my hypothesis, but the dearth of extended CDR H3 structures at various lengths precludes using this result predictively.

4.4 Discussion

CDR H3 is the most diverse region in antibodies due to its position relative to the V(D)J recombination sites, junctional diversification at these sites, and somatic hypermutation. Accordingly, the CDR H3 loop often plays a central role in antigen recognition and is a major contributor to binding strength. The success of several therapeutic antibodies and the advent of next-generation sequencing techniques have led to an increased interest in computational antibody structure prediction and design. While there has been progress in these efforts, accurate modeling of CDR H3 has remained challenging, leading me to question whether (1) the diversification of CDR H3 can lead to extremely rare conformations; or (2) there are environmental factors encoded into the F_V. My results indicate that CDR H3-like conformations, while not common, occur with some regularity, occurring in 7.4% of 5,783 Pfams and 6.0% of the 15,769 chains in the non-antibody set. Environmental factors are most likely responsible for kink stabilization.

I identified 1,030 protein segments of at least 9 residues from 632 distinct Pfam alignments that match the same 3D transformation as the anchors of the H3 loop and

include the C-terminal kink motif that is common in antibodies. Without the inclusion of the kink in my search criteria, most of the matches are extended strand-turn-strand conformations, suggesting that adopting CDR H3-like conformations is unusual. This is helpful for understanding why *de novo* loop structure prediction of CDR H3 tends to produce models with extended base geometry and indicates that using constraints for this purpose is likely a wise course of action. In fact, when prediction algorithms use fragment or template-based approaches, the libraries are predominantly composed of structures that do not adopt the kinked base geometry, making it challenging to identify appropriate conformations. The data presented here can be used to enrich fragment or template libraries effectively.

For example, RosettaAntibody accounts for the kink either by using a curated set of fragments or by filtering H3 loops with poor kink geometry.[22,39] Here I have established a more detailed geometric description of the kink and identified a significantly larger set of structures from which fragments can be selected. Both results can be used as a starting point for improving *de novo* CDR H3 loop structure prediction.

The set of identified loops with LAT and kink matches contains close structures ($\leq$ 2.0 Å) for roughly 50% of H3 loops 9–20 residues in length, showing that CDR H3 loops do not adopt conformations that are inaccessible to loops in other proteins. In most Pfams, kinked loops appear to arise only in some family members, while in others they are highly conserved structural features. One such protein family, PDZ domains, has evolved a motif for protein recognition and binding that is strikingly similar in structure and function to CDR H3. The appearance of the kink irrespective of the presence of the stabilizing residues

indicates that environmental factors are crucial to kink formation.

Furthermore, I have produced a set of H3-like structures of a wide variety of lengths from non-antibody proteins. Across all loop lengths, and especially for long loop lengths, there are more potential template loops from non-antibody structures than from antibody structures. If the quality and homology constraints that were used to cull the PDB were relaxed, it is likely I would identify even more, albeit lower quality, H3-like regions in non-antibody proteins. This set of structures could be incorporated into a database that could be used to assist CDR H3 structure prediction by threading the sequence of interest onto many possible H3-like backbones, analogous to successful database-based methods for loop structure prediction.[136–141] The green curve in Figure 4.7 shows that supplementing known CDR H3 loops with the LAT+kink matches results in a set of template structures that contains more structures with low-RMSDs to CDR H3 loops than either set alone.

Another possible use for this set of structures is in the field of computational antibody design. The extremely large sequence and conformational spaces of long loops often make incorporating backbone motions into design methods infeasible. Effective sampling is further complicated if docking simulations are desired, as may be the case in designing a binding region such as CDR H3. The large number of PDB matches at long loop lengths for which there are few or no H3 loops provides an opportunity to present multiple H3-like scaffolds for fixed and flexible backbone design routines. Using the provided scripts and instructions included in Supplemental Information, a set of all of the backbone coordinates of the LAT+kink matches can be extracted and used for novel design routines. Thus, it is expected that the identified structures will improve antibody design.

4.5 Conclusion

This is the first study to my knowledge that uses non-antibody loops to analyze CDR H3 structures, which required developing the most detailed description of the CDR H3 loop to date (LAT+kink). While the kink has been discussed in the past,[19,24–26,28] previous descriptions were more useful for classifying CDR H3 loops than as a rigorous description of the geometry, as demonstrated by various failures in CDR H3 prediction attempts. For example, I observed that the previous kink geometrical description can be satisfied in multiple ways.[39] My work shows that the residues that had been previously indicated in kink formation are not present in kinked structures from non-antibody proteins (Figure 4.13). In fact, no local interactions among the loop residues fully explain the presence of the kink. Instead, I am led to the conclusion that the Ig heavy chain fold stabilizes the kink, and thus it is the extended H3 structures that are the exceptions and not the kinked loops. Whereas previous studies have explained the presence of the kink as a "strange" structural feature, I show here that the kink is not strange; it is found in a wide range of proteins, and some other proteins even conserve it and use it in diverse loops that are involved in binding.

All of my results lead to my hypothesis of why the kinked base geometry is preferred: it is an agent of loop diversification. The C-terminal kink in H3 loops disrupts the β-strand pairing, allowing increased structural diversity with the same number of residues. In other words, if it were not for the kink, most sequences would form extended strand-turn-strand conformations, giving little structural diversity, but with a kink, many structures of similar free energy can form instead. Such a feature is advantageous to an

antibody undergoing somatic hypermutation to improve affinity and specificity to a newly introduced antigen. For this reason, I believe the heavy chain fold has been selected to form the kink, and it is only in rare circumstances that the extended geometry is energetically favorable compared to the kinked conformation.

4.6 Methods

4.6.1 Datasets

A set of IgG heavy chain V domains, constructed and filtered as described by North *et al.* (resolution ≤ 2.8 Å backbone B-factor ≤ 80.0 Å^2, no missing coordinates, no *cis*-non-Proline residues, conformational energy ≤ 9.5),[19] was further filtered for redundancy by removing structures with CDR loops of identical length with either a single residue difference or no differences in sequence. Using the PISCES web server[142] a diverse set of high quality non-antibody protein chains was obtained by searching the PDB[15] for chains with maximum sequence identity of 70%, a resolution of 2.2 Å or better, and a maximum R-value of 0.25. Before recording results, segments with high B-factors in backbone atoms (> 80.0 Å^2) were filtered out.

4.6.2 Loop anchor transform calculation

Unlike other investigations of CDR H3 structures,[19,24–29,143,144] this study focuses on comparing CDR H3 loops to non-antibody proteins rather than restricting the comparison to other antibodies. For this reason, I developed a description of the CDR H3 loop environment based on structure independent of sequence. The definitions used by North *et al.*[19]

(residue numbers 93–102 using the Chothia numbering scheme[16]) were used to identify the terminal residues on the CDR H3 loop. A coordinate frame was defined using the main chain backbone atoms (N, C_α, C) of each of these residues such that the z-axis is the unit vector along the C_α–C bond, the y-axis lies in the N–C_α–C plane and the x direction is the vector product of the y and z directions. The six degrees of freedom of the 3D transformation of the C-terminal coordinate frame onto the N-terminal coordinate frame together compose what I term the loop anchor transform (LAT). The covariance for each pair of degrees of freedom revealed that each degree of freedom could be treated independently.

4.6.3 Loop Anchor Transform Parameters

A Loop Anchor Transform (LAT) is the three-dimensional transformation between the anchor residues of the loop, *i.e.* the transformation matrix required to perfectly superimpose the backbone heavy atoms of the residues immediately preceding and following the loop. Figure 1A shows an example H3 loop annotated to highlight the anchor residues and coordinate frames.

I compute the LATs using homogeneous coordinates because they incorporate translation and rotation into a single matrix. For each anchor residue, I define the z-axis ($\hat{z}$) as the unit vector pointing from C_α to the carbonyl carbon,

$$\frac{\vec{C} - \vec{C}_\alpha}{\left\|\vec{C} - \vec{C}_\alpha\right\|}, \tag{4.1}$$

then $\hat{y}$ as the unit vector normal to $\hat{z}$, that lies in the N–C_α–C plane:

$$\hat{y} = \frac{\vec{a} - (\vec{a} \cdot \hat{z})\hat{z}}{\left\|\vec{a} - (\vec{a} \cdot \hat{z})\hat{z}\right\|}, \tag{4.2}$$

where

$$\vec{a} = \left(\vec{\mathrm{N}} - \vec{\mathrm{C}}_\alpha\right),$$

and $\hat{x}$ is simply the cross product of $\hat{y}$ and $\hat{z}$.

$$\hat{x} = \hat{y} \times \hat{z}$$

The carbonyl carbon coordinates, $\vec{\mathrm{C}} = (\mathrm{C}_x, \mathrm{C}_y, \mathrm{C}_z)$, are used in conjunction with the orthonormal basis vectors determined above to construct a homogeneous coordinate transformation matrix,

$$\mathbf{F} = \begin{pmatrix} \hat{x}_1 & \hat{y}_1 & \hat{z}_1 & \mathrm{C}_x \\ \hat{x}_2 & \hat{y}_2 & \hat{z}_2 & \mathrm{C}_y \\ \hat{x}_3 & \hat{y}_3 & \hat{z}_3 & \mathrm{C}_z \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.3}$$

A single point p defined relative to the global origin can be related to the point p' defined relative to the N-terminal coordinate frame ($\mathbf{F}_1$) and the point p'' defined relative to the C-terminal coordinate frame ($\mathbf{F}_2$) as follows:

$$p = \mathbf{F}_1 p' = \mathbf{F}_2 p''. \tag{4.4}$$

By right multiplying by $\mathbf{F}_1^{-1}$, the relationship between points defined in coordinate frames $\mathbf{F}_1$ and $\mathbf{F}_2$ is found:

$$p' = (\mathbf{F}_1^{-1}\mathbf{F}_2)p'' \tag{4.5}$$

To invert $\mathbf{F}_1$, I take advantage of the fact that its upper 3×3 submatrix is orthogonal, so its inverse is its transpose. The inverse of the fourth column is computed by negating the product of the submatrix and the carbonyl carbon coordinates.

I extract the LAT parameters from $\mathbf{F}_1^{-1}\mathbf{F}_2$ by noting that the three translational degrees of freedom (X, Y, Z) are defined by the fourth column in the matrix. The three rotational degrees of freedom, represented by the Euler angles (ϕ, ψ, θ), can be computed from the upper 3×3 submatrix. Euler angles describe an arbitrary rotation as the successive elemental rotations about $\hat{z}$ (Φ) followed by a rotation about $\hat{x}$ (Θ) and another rotation about $\hat{z}$ (Ψ). Elemental rotations by an angle θ about $\hat{x}$ and $\hat{z}$ are represented as

$$\mathbf{R}_{\hat{x}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \text{ and} \tag{4.6}$$

$$\mathbf{R}_{\hat{z}} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.7}$$

respectively. Rotating by an angle ϕ about $\hat{z}$, then by an angle θ about $\hat{x}$ and then by ψ about

$\hat{z}$ results in the rotation matrix

$$\mathbf{A} = \Phi\Theta\Psi, \tag{4.8}$$

$$\mathbf{A} = \begin{pmatrix} \cos\phi\cos\psi - \cos\theta\sin\phi\sin\psi & -\cos\phi\sin\psi - \cos\theta\cos\psi\sin\phi & \sin\phi\sin\theta \\ \cos\psi\sin\phi + \cos\phi\cos\theta\sin\psi & \cos\phi\cos\theta\cos\psi - \sin\phi\sin\psi & -\cos\phi\sin\theta \\ \sin\theta\sin\psi & \cos\psi\sin\theta & \cos\theta \end{pmatrix}. \tag{4.9}$$

Using the values from the upper 3×3 submatrix of $\mathbf{F}_1^{-1}\mathbf{F}_2$, I solve for the Euler angles

$$\phi = \text{atan2}(\mathbf{A}_{1,3}, -\mathbf{A}_{2,3}), \tag{4.10}$$

$$\psi = \text{atan2}(\mathbf{A}_{3,1}, -\mathbf{A}_{3,2}), \text{and} \tag{4.11}$$

$$\theta = \arccos \mathbf{A}_{3,3} \tag{4.12}$$

Figure 4.2 shows density estimates of each of the six parameters for 13-residue H3 loops and segments from the culled PDB set. ϕ and ψ are on a domain of 0–2π and θ is on a domain of 0–π.

4.6.4 Features analysis

LATs were calculated using the feature analysis framework[145] within the Rosetta software suite.[32] A custom feature reporter was developed to compute (1) LATs for every 5–31 residue window in each chain in the non-antibody dataset and (2) the C_α–C_α–C_α pseudo bond angle of the last three residues in each window. The results were saved to a relational database (http://www.sqlite.org). Analysis scripts were developed to display distributions of the results using the ggplot2[146] library in R.[147] The resulting database was then queried

to identify regions of proteins with LATs and bond angles within $\pm$ 3.0 σ of the mean of the distributions developed from the antibody dataset.

4.6.5 Primary & secondary structure analysis

Sequence and secondary structure comparisons were performed using a local copy of WebLogo.[124] When comparing secondary structures, the DSSP code[123] is used in place of the one-letter amino acid abbreviation. Due to limitations of WebLogo, the "B" DSSP code (β-bridge) and a blank DSSP code are represented as "R" and "L", respectively.

4.7 Extracting the LAT+kink matches from the PDB

In order to facilitate use of the data sets described in this manuscript, I have provided several supplemental files than can be used to generate a local set of the backbone coordinates of all of the LAT+kink matches. Readers interested in obtaining these structures will need access to a computer running a POSIX-compliant operating system (Unix, GNU/Linux, etc.), rsync, perl and python2.7.

1. First, the reader will need a local mirror of the PDB (approximately 30 GB). This is most easily accomplished by using the script `rsyncPDB.sh`, which is provided by RCSB here: http://www.rcsb.org/pdb/static.do?p=download/ftp/index.html
   It is recommended that readers create a directory within the directory they would ultimately like to use and set the script to mirror to the inner-most directory. For example, if you want your nicely named files to be in `/pdb`, set `rsyncPDB.sh` to mirror to `/pdb/data`. You will need to create these directories before running the script.

2. Next, configure the `pdbName.pl` script (provided with this manuscript) to point to the directories used in the previous step. Following the directory names above, the correct configuration is `$MIRRORDIR="/pdb";` and `$data_dir ="data";`.

3. Run `pdbName.pl` by typing `perl pdbName.pl`.

4. Run `extract_lat_kink_matches.py` (provided with this manuscript) to read in `supporting_file1.txt` (provided with this manuscript) and point to the directory that contains the PDB mirror and an output directory. An example command line for this script is:

   ```
   python extract_lat_kink_matches.py -f supporting_file1.txt
   -p /pdb -o outdir
   ```

   Descriptions of the flags can be accessed by typing:

   ```
   python extract_lat_kink_matches.py --help
   ```

CHAPTER V

IMPROVEMENTS IN CDR H3 STRUCTURAL MODELING

5.1 Overview

Antibody structure prediction has made great strides, but accurately modeling CDR H3 loops remains elusive. Unlike the other five CDR loops, CDR H3 does not adopt canonical conformations and usually must be modeled *de novo*. Recent advances in *de novo* loop modeling methods have shown success in modeling longer loops and showed promise during Antibody Modeling Assessment II (AMA II). In Chapter 3, my coworkers and I found that simulations needed to be biased toward kinked CDR H3 conformations to generate low-RMSD models, and in Chapter 4, I presented new geometric parameters, τ_{101} and α_{101}, that define the kink conformation. In this chapter, I use these parameters to develop a new constraint that can applied during the simulation to bias toward kinked conformations. The functional form of the constraint is selected to ensure that it is differentiable, to enable minimization, and to avoid over-constraining the possible solutions. When applied to a benchmark set of high-quality CDR H3 loops, the average minimum RMSD sampled is 0.93 Å, compared to 1.34 Å without the constriant. The constraint also enables Rosetta

to find conformations that score closer to the native structure. The average RMSD of the top-ranked model is 2.0 Å for the constrained simulation and 3.2 Å without constraints. The performance of the constrained *de novo* method is then tested in the context of homology modeling and rigid-body docking.

5.2 Introduction

The adaptive immune system in vertebrates is capable of raising antibodies against a countless number of antigens. More recently, however, engineered antibodies have been used as therapeutic molecules[52,53] and biosensors.[4–6] The source of these antibodies varies across specific applications. In order to optimize specific modes of interactions, rational engineering techniques must be developed. Rational engineering of antibodies requires accurate structural models, but crystallization is not always practical or even possible. Additionally, expressing a large library of mutants in order to assess the energetic implications of specific mutations is time consuming, resource intensive and, in some cases, technically challenging. Computational methods, namely antibody homology modeling, are poised to enable the realization of rational design.

RosettaAntibody's approach to modeling[22] is to break the structure into eight distinct structural components: the heavy- and light-chain frameworks; CDR loops L1–3; and CDR loops H1–3. Because the non-H3 CDR loops adopt canonical conformations,[17,19] accurate backbone conformations for them can usually be found in known structures. RosettaAntibody exploits this by selecting templates from curated structural databases by BLAST[97] bit-score for CDRs L1–3, H1 and H2 and the framework regions. Each structural

component is defined such that they have overlapping residues that can then be superposed to create a grafted model. An initial V_H–V_L orientation is also selected from databases, and the grafted heavy and light chains are each superposed to the corresponding chain in the orientation template. After this, the CDR H3 loop is modeled *de novo* while sampling the V_H–V_L orientation.

In Chapter 3,[39] I presented the performance of RosettaAntibody in Antibody Modeling Assessment II (AMA II).[40] With few exceptions, RosettaAntibody selects templates for the framework regions and the non-H3 CDR loops with sub-Ångström RMSD from the native structure. The most difficult aspect of antibody homology remains accurately predicting the V_H–V_L orientation and the CDR H3 conformation.

A large majority of CDR H3 loops have a C-terminal kink,[19,24–28,41] and in AMA II (Chapter 3) I found that producing low-RMSD models required filtering out non-kinked H3 conformations. However, the scores of the kinked structures was higher than some of the extended structures that Rosetta produced. In response to these findings, I developed new geometric parameters that describe the kink in Chapter 4.[41]

Because the CDR H3 loop lies at the interface between the heavy and light chains, incorrect V_H–V_L orientations can frustrate identifying correct CDR H3 conformations. In the time that has elapsed since AMA II was conducted, progress has been made in predicting V_H–V_L orientation[148] from sequence by training a random forest model[149] on a set of "fingerprint" residues at the V_H–V_L interface using ABangle's six degree-of-freedom description of orientation.[86] Similarly, effort has been made to develop a CDR H3-specific loop modeling routine,[150] but successful predictions require extremely accurate atomic

coordinates for the rest of the F_V,[150,151] which may make these tools better-suited for refining crystal structures with poor electron density around the CDR H3 loop than for homology modeling.

De novo loop modeling has endured as a challenging problem in part because of the large number of degrees of freedom that need to be sampled, as well as the challenges associated with accurately ranking different structures that may appear to be very similar when using a coarse-grained measurement such as RMSD. Additionally, side-chain interactions may play key roles in stabilizing observed loop conformations, potentially complicating low-resolution searches. Complicating the task even further is the most common source of the reference coordinates: crystal structures. Crystals are extremely crowded environments in which each protein molecule is surrounded by several by others; this may or may not influence the observed conformation within the asymmetric unit. Without the existence of a crystal structure of the same protein in more than one distinct crystal form, it cannot be determined if these "crystal contacts" perturb the conformation of any region of the protein.

Similarly, another complication of loop modeling is the search for a single set of coordinates. Proteins in physiological conditions are not completely rigid, and estimating the conformational entropy of a loop requires supplying a model to describe the describe the modes of flexibility accessible to the loop.[152] Nevertheless, the possible existence of multiple degenerate-energy conformations cannot be dismissed.

In this chapter, I use the parameters defend in Chapter 4[41] to constrain the kink during the course of a simulation. To limit the uncertainty in the crystallographic coordinates, I constructed a set of extremely high-resolution H3 loops. Given the high degree

of confidence in the atomic coordinates, computed RMSD values are also better-defined. The constraint is tested by predicting H3 conformations on the crystal framework structure across the set of benchmark structures. Finally, to test the utility of the constraint, I also assess the ability to dock an antibody with a modeled H3 loop and CDR H3 modeling on a homology modeled framework.

5.3 Methods

5.3.1 Dataset

A set of F_Vs with accurate CDR H3 coordinates was constructed by querying the backend databases of PyIgClassify[153] for structures with a resolution of 2.5 Å or better, a maximum R-value of 0.2, B-factor $\leq$ 80.0 Å^2 for every atom in the structure, only one copy of the F_V in the asymmetric unit, and CDR H3 loop-lengths ranging from 9–20 residues. To ensure the set has diverse chemical environments, no two heavy-chain CDR loops are permitted to be identical in sequence. The structures were further filtered to remove antibodies from species other than humans and mice, and modified residues (namely pyroglutamic acid (PCA), a cyclized form of glutamine or glutamic acid). The resulting set of structures contains 49 F_Vs and is summarized in Table 5.1.

5.3.2 Kink constraint

In *de novo* loop modeling simulations it is impossible to exhaustively sample all of the structural degrees of freedom. To increase the likelihood of generating a model with a near-native structure, Rosetta has the ability to add an arbitrary potential that is evaluated using

the value of the distance of two atoms, angle of three atoms, or torsion angle of four atoms. These potentials, referred to as "constraints" in Rosetta parlance, allow experimental data or homology information to be exploited to improve the accuracy of structure prediction.[32] In the case of CDR H3, there are two parameters that can be constrained: (1) τ_{101}, the C_α–C_α–C_α pseudo bond angle for the three C-terminal residues; and (2) α_{101}, the C_α–C_α–C_α–C_α pseudo dihedral angle for the three C-terminal residues in the CDR H3 loop and one adjacent residue in the heavy chain framework.

Because an objective of this study is to determine whether or not Rosetta can correctly identify native H3 conformations, it is important not to over-constrain any of the simulations. With this in mind, a `FLAT_HARMONIC` potential, which has a region wherein no penalty is applied, is a natural choice. The `FLAT_HARMONIC` potential is of the form

$$f(x) = \begin{cases} 0, & \text{if } |x-\mu| \leq t \\ (\frac{|x-\mu|-t}{\xi})^2, & \text{if } |x-\mu| > t \end{cases}, \tag{5.1}$$

where μ is the mean, t (tolerance) is the distance from μ with no penalty, and ξ is the scaling factor that controls the penalty that is applied.

For the kink parameters α_{101} and τ_{101}, the following penalty schedule was devised: no penalty should be applied when the value is within 1.0 σ of the mean, and a penalty of 1.0 at 3.0 σ. This will encourage Rosetta to generate models with kinked H3 loops without forcing the geometry toward the mean values of both parameters.

Because the penalty should begin after 1.0 σ, t can be set to σ. Now I solve for ξ in

order to produce the desired penalty schedule. First, I solve for ξ at 3.0 σ as follows:

$$\begin{aligned} f(\mu + 3t) = 1.0 &= \left(\frac{\mu + 3t - \mu - t}{\xi}\right)^2 \\ \left(\frac{2t}{\xi}\right)^2 &= 1.0 \\ \xi &= 2t \end{aligned} \tag{5.2}$$

and then plug in $\xi = 2t$ and evaluate the penalty at 2.0 σ to check if this intermediate value produces a reasonable penalty

$$\begin{aligned} f(\mu + 2t) &= \left(\frac{\mu + 2t - \mu - t}{2t}\right)^2 \\ \left(\frac{1}{2}\right)^2 &= 0.25 \end{aligned} \tag{5.3}$$

and we find that setting ξ to 2σ will exactly produce the desired penalty schedule with the useful feature of being a factor of four larger at 3.0 σ than at 2.0 σ.

Using the values determined in Chapter 4,[41] the kink constraint for AHo-numbered antibodies is written as:

```
# alpha: pseudo dihedral - last 3 residues in H3 and the following W
# mean: 38.85 degrees; SD: 11.75 degrees (in radians)
Dihedral CA 136H CA 137H CA 138H CA 139H FLAT_HARMONIC 0.678 0.41 0.205

# tau: pseudo bond angle of the last 3 residues in H3
# mean: 100.9 degrees; SD: 5.57 degrees (in radians)
Angle CA 136H CA 137H CA 138H FLAT_HARMONIC 1.761 0.194 0.0972
```

Figure 5.1A shows the functional form of the `FLAT_HARMONIC` potential used for each of the constraints, and Figure 5.1B shows a contour plot of the combined value of the τ_{101} and α_{101} constraints, with each line representing an increase in score of 2.0 Rosetta

Energy Units (REU). Figure 5.1B also shows the regions of τ_{101} and α_{101} that define kinked (orange; $\pm$ 3.0 σ of the mean of both parameters), unclear (gray; $\pm$ 3.0 σ of mean of one of the parameters), and extended (white; beyond 3.0 σ of both parameters) conformations. These definitions are used throughout this study.

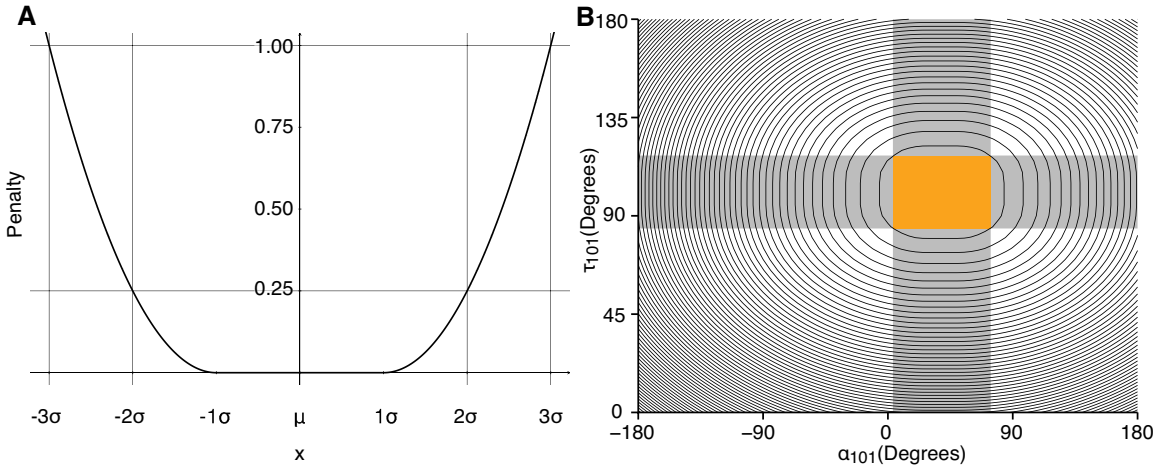

Figure 5.1: Functional form of the constraint used to bias *de novo* CDR H3 loop modeling simulations toward generating kinked conformations. (A) A plot of the `FLAT_HARMONIC` potential with the parameters that were found in equations 5.3 and 5.2. (B) A contour plot showing the value of the kink constraint across all values of τ_{101} and α_{101}. Each line represents a 2.0 Rosetta Energy Unit (REU) increase in penalty. The orange box demarcates $\pm$ 3.0 σ of the mean of the τ_{101} and α_{101} distributions. Throughout the rest of this text, models falling within this region are classified as "kinked", models with τ_{101} and α_{101} in the gray, shaded regions are classified as "unclear", and all other models are classified as "extended".

5.3.3 *De novo* loop structure prediction

Rosetta has several loop modeling routines integrated into a unified framework. The most popular loop modeling methods are cyclic coordinate descent (CCD),[154] kinematic closure (KIC)[88] and next-generation KIC (NGK).[89] Like most Rosetta applications, the loop modeling methods consist of low- and high-resolution stages, where side chains are represented as a single pseudoatom and in full atomic detail, respectively. The low-

resolution stage of CCD consists of insertion of fragments of known structures followed by cyclic coordinate descent loop closure[102] to ensure loop continuity, and the high-resolution stage uses small perturbations to the backbone dihedral angles, CCD loop closure and side-chain packing.

KIC generates candidate backbone conformations by sampling ϕ and ψ dihedral angles from a Ramachandran distribution[155] for all but three "pivot" residues. The ϕ and ψ dihedral angles for the pivot residues are solved analytically from a 16-order polynomial[156] using resultants.[157] The process of generating backbone conformations is the same method in both stages, but the all-atom stage also includes side-chain optimization. KIC has been shown to generate more near-native models on a benchmark set of 12-residue loops than the CCD-based method.[88]

More recently, NGK has been developed to further improve the performance of loop modeling in Rosetta. NGK is similar in approach to KIC but uses neighbor-dependent Ramachandran maps,[104] explicitly samples ω backbone dihedral angles and introduces a simulated annealing strategy for repulsive and Ramachandran score terms in the all-atom stage. On the same set of loops used to benchmark KIC, NGK generates substantially more near-native models,[89] which is why it is used as the starting point in this study.

The initial implementation of the neighbor-dependent Ramachandran sampling required approximately 5 GB of memory, making it impossible to run on any HPC resource. Before this study could be conducted, I redesigned the underlying data structure to include only the energy of each conformation and the cumulative probability across all conformations, representing all of the data in less than 160 MB.

The flags to run a standard NGK simulation are:

```
./loopmodel.macosclangrelease
  -native input_file.pdb
  -s input_file.pdb
  -nstruct 500
  -loops:loop_file h3.loops
  -loops:remodel perturb_kic
  -loops:refine refine_kic
  -loops:outer_cycles 5
  -kic_bump_overlap_factor 0.36
  -legacy_kic false
  -kic_min_after_repack true
  -corrections:score:use_bicubic_interpolation false
  -loops:kic_omega_sampling
  -loops:kic_rama2b
  -allow_omega_move
  -loops:ramp_fa_rep
  -loops:ramp_rama
  -ex1
  -ex2
  -extrachi_cutoff 0
```

where `h3.loops` contains

```
# FORMAT JSON
{"LoopSet" : [{
        "start" : { "resSeq" : 93, "iCode" : " ", "chainID" : "H" },
        "stop" : { "resSeq" : 102, "iCode" : " ", "chainID" : "H" },
        "extras" : { "extend" : true },
    }]
}
```

NGK simulations with constraints use the above command line with the addition of the following flags:

```
  -constraints:cst_file kink.constraint
  -constraints:cst_weight 1.0
  -constraints:cst_fa_file kink.constraint
  -constraints:cst_fa_weight 1.0
```

where the file `kink.constraint` contains the constraints as shown in Section 5.3.2.

Because of their object-oriented design, the loop modeling methods in Rosetta can be mixed-and-matched within a single simulation. To couple the more conservative CCD refinement with the aggressive NGK sampling, a combined NGK–CCD simulation can be run. The combined simulation uses the same command line as the constrained NGK simulation, but with `-loops:refine refine_kic` changed to `-loops:refine refine_ccd`.

5.3.4 Discrimination score

The discrimination score is used to measure how "funnel-like"—that is, lower RMSDs correspond to lower scores—a score vs. RMSD plot is, with a lower value being indicative of a more successful simulation. As defined by Conway *et al.*,[158] the discrimination score relies on scaling the scores of the decoys such that a value of 1.0 corresponds to the 95$^{\text{th}}$ percentile of scores and a value of 0.0 corresponds to the 5$^{\text{th}}$ percentile.

$$D = \sum_{r \in \{1,1.5,2,2.5,3,4,6\}} \min_{i,\mathrm{RMS}(i) \in [0,r]} S_i - \min_{i,\mathrm{RMS}(i) \in (r,\infty]} S_i, \tag{5.4}$$

where r is the RMSD cutoff in Å, S_i is the dimensionless scaled scored, and the discrimination score, D, is the sum of the score-differences of the best-scoring models above and below the seven RMSD cutoffs.

5.3.5 Preparation of input structures

The crystallographic coordinates of protein structure often do not score favorably within Rosetta. To address this, crystal structures must be relaxed, that is, optimized with respect to the Rosetta scoring function. Relaxation will result in small changes to the atomic

coordinates with significant improvements in the score; however, it is important that the coordinates do not vary much, especially in the case of loop modeling. The command line used for constrained relax is:

```
./relax.macosclangrelease
  -s input.pdb
  -nstruct 500
  -relax:constrain_relax_to_start_coords
  -relax:coord_constrain_sidechains
  -relax:ramp_constraints false
  -ex1
  -ex2
  -use_input_sc
```

Once the crystallographic coordinates have been optimized, the entire structure can be subjected to fixed-backbone side-chain optimization to further lower the score of the reference structure for loop modeling and to better approximate the side-chain conformations in the free, unbound conformation. The command line used for fixed-backbone side-chain optimization is:

```
./relax.macosclangrelease
  -s output_from_previous_simulation.pdb
  -nstruct 100
  -relax:bb_move false
  -ex1
  -ex2
  -extrachi_cutoff 0
```

The low-scoring model from each of these simulations is used as the input structure in the subsequent calculations.

5.4 Results

A set of 49 high-quality CDR H3 structures was constructed as described in section 5.3.1. Figure 5.2 shows an example loop from the data set with the electron density map for the

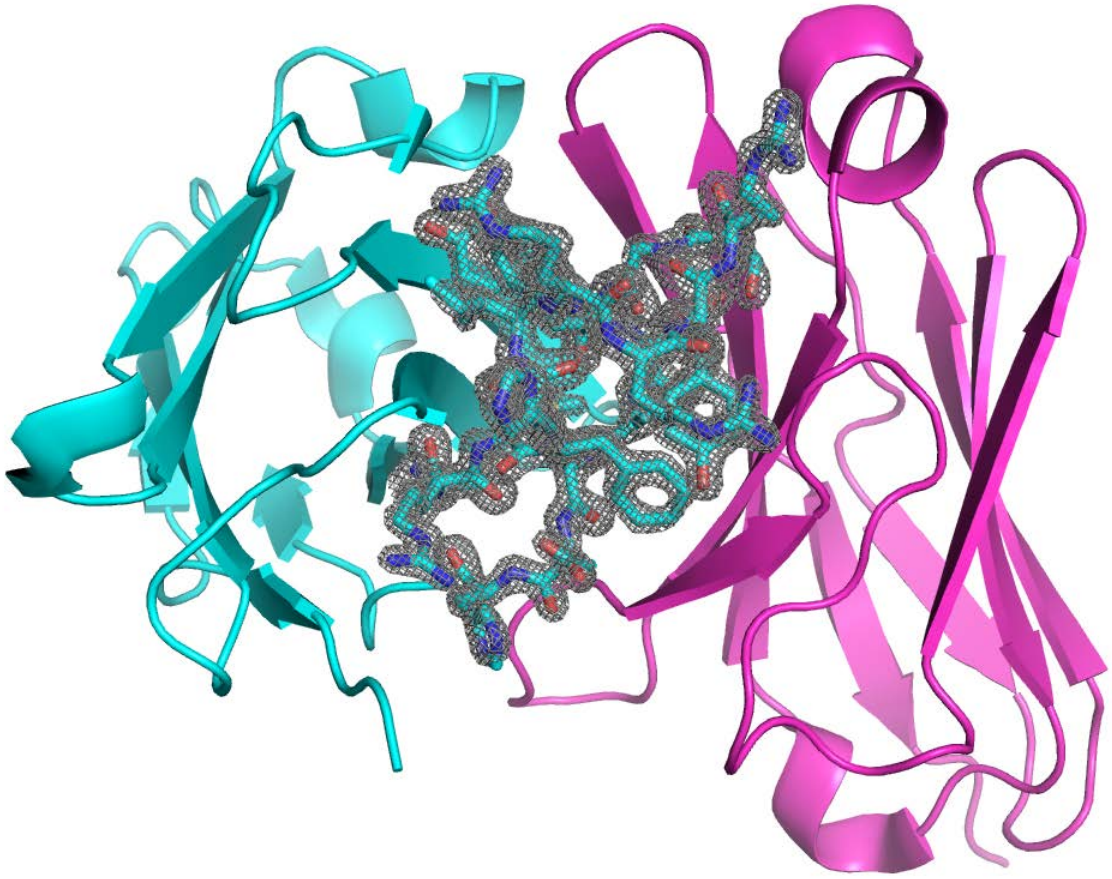

Figure 5.2: The F_V of an anti-peptidase S1 antibody (PDB accession code 3nps[159]) is shown with the V_H domain in cyan and the V_L domain in magenta. The electron density of the 19-residue CDR H3 loop is indicated with a mesh contour map within 1.6 Å of the coordinates in the PDB file. The crystal structure has an R-value of 0.190 and a resolution of 1.50 Å. The electron density is clearly resolved across the entire CDR H3 loop, indicating that both a high-quality crystal and a stable loop conformation among several symmetric copies of the antibody in the crystal. The crystal structure contains the full F_{ab} bound to the antigen, which may further stabilize the loop's conformation.

H3 residues shown in a gray mesh over the residues represented in sticks. At this level of detail all of the side-chain coordinates are well-defined, and the map even shows a hole in aromatic residues. The level of agreement between the electron density map and the coordinates and lack of ambiguity in the atomic coordinates suggests that this loop is in a stable conformation in the crystal, making it a prime candidate for loop modeling

experiments. The other loops in the set have similarly well-defined electron density.

Table 5.1 lists all of the loops in the set and includes information on the quality and content of the crystal structure, the species from which the antibody was derived, the length of the loop, the pH at which the crystallization experiment was conducted (if available) and the light chain isotype. In the set, 24 of the 49 structures are crystallized in the bound conformation with their antigen, 40 of the 49 structures are F_{ab}s, six are F_Vs and the remaining three are scF_Vs. Eighteen of the structures are of human antibodies, and 11 have λ light chains, making this a diverse set of structures.

The definition of the bounds of the CDR H3 loop differ from the Chothia-based definition[16] used within RosettaAntibody,[22] and instead are based on the Honegger–Plückthun-based definition[160] used by North *et al.*[19] and in Chapter 4.[41] Both definitions end on Chothia residue number 102, but the Chothia-based definition begins at residue 95 while the Honegger–Plückthun-based definition begins at residue 93, making the Honegger–Plückthun CDR H3 loops two residues longer than Chothia loops. In this set, the median and mode of the loop lengths are both 12 residues.

5.4.1 Unconstrained *de novo* modeling of CDR H3 loops

In Chapter 3[39] NGK was used to model CDR H3 loops on a crystallographic framework. Due to the limitations imposed by high memory requirements, only a small number of models could be produced using NGK, and the majority of modeling was done using KIC. In these cases, a filter was employed to favor kinked structures, using the ϑ_{base} (α_{101} in this work) definition developed by Shirai *et al.* and refined by Kuroda *et al.*[24–26] Without this penalty, very few kinked structures were produced.

Table 5.1: Structural information for the CDR H3 benchmark set.

PDB Code	R Value	Res. (Å)	Max. B factor (Å^2)	Max. H3 B factor (Å^2)	Species	CDR H3 length	Light chain isotype	Fragment	pH
1x9q*	0.193	1.50	32.57	21.22	Human	9	κ	scF$_V$–Ag	4.6
2d7t	0.191	1.70	49.16	47.73	Human	9	κ	F$_V$	9.3
3hc4	0.162	1.62	41.11	33.50	Human	9	κ	F$_{ab}$	8
1mlb	0.181	2.10	49.63	34.78	Mouse	9	κ	F$_{ab}$	—
2e27*	0.198	1.70	51.38	31.14	Mouse	9	κ	F$_V$–Ag	5.7
3g5y	0.199	1.59	40.08	35.90	Mouse	9	κ	F$_{ab}$–Ag	—
3m8o*	0.154	1.55	55.22	24.48	Human	10	κ	F$_{ab}$	7
1jpt	0.182	1.85	49.59	29.78	Humanized Mouse	10	κ	F$_{ab}$	4.6
3e8u	0.188	2.10	32.50	27.11	Mouse	10	κ	F$_{ab}$–Ag	—
1mqk	0.136	1.28	51.66	37.15	Mouse	11	κ	F$_V$	6.0
1nlb	0.197	1.60	44.78	41.47	Mouse	11	κ	F$_{ab}$	9
2adf	0.192	1.90	35.22	20.68	Mouse	11	κ	F$_{ab}$–Ag	4.6
2fbj	0.194	1.95	60.06	41.76	Mouse	11	κ	F$_{ab}$–Ag	—
2w60	0.171	1.50	54.60	36.38	Mouse	11	κ	F$_{ab}$	8
3gnm	0.189	2.10	46.42	43.99	Mouse	11	κ	F$_{ab}$	4.0
3hnt	0.199	1.80	54.27	28.18	Mouse	11	κ	F$_{ab}$–Ag	7.1
3v0w	0.184	1.73	60.85	60.81	Mouse	11	κ	F$_{ab}$–Ag	4.6
1mfa	0.166	1.70	69.57	41.64	Mouse	11	λ	F$_V$–Ag	—
3mxw	0.181	1.83	72.85	31.98	Mouse	12	κ	F$_{ab}$–Ag	—
2xwt	0.179	1.90	44.78	29.30	Human	12	λ	F$_{ab}$–Ag	5.0
1dlf	0.183	1.45	42.75	24.85	Mouse	12	κ	F$_V$	5.25
2ypv	0.183	1.80	75.36	37.38	Mouse	12	κ	F$_{ab}$–Ag	8.5
3ifl	0.180	1.50	34.75	22.57	Mouse	12	κ	F$_{ab}$–Ag	9.0
3liz*	0.178	1.80	52.33	43.85	Mouse	12	κ	F$_{ab}$–Ag	7.2
3oz9	0.192	1.60	55.29	31.28	Mouse	12	κ	F$_{ab}$	8.5
3umt	0.177	1.80	56.76	39.55	Mouse	12	κ	scF$_V$	9.5
4h0h	0.197	2.00	65.19	37.96	Mouse	12	κ	scF$_V$	6.5
4h20	0.197	1.90	45.61	20.26	Mouse	12	κ	F$_{ab}$	7.4
4hpy	0.171	1.50	55.43	42.96	Human	13	λ	F$_{ab}$–Ag	6.5
2v17	0.160	1.65	37.77	23.54	Mouse	13	κ	F$_{ab}$–Ag	7.5
3t65	0.194	1.45	63.85	34.13	Mouse	13	κ	F$_{ab}$–Ag	8.0
1oaq	0.160	1.50	44.45	43.75	Mouse	13	λ	F$_V$	5.00
2vxv	0.155	1.49	37.34	29.51	Human	14	κ	F$_{ab}$	10.5
3eo9	0.191	1.80	45.44	27.45	Humanized Mouse	14	κ	F$_{ab}$	5.0
3p0y	0.182	1.80	76.35	25.77	Human	14	κ	F$_{ab}$–Ag	8
1jfq	0.196	1.90	57.19	40.56	Mouse	14	κ	F$_{ab}$	7.5
2r8s	0.196	1.95	63.07	34.21	Human (library)	14	κ	F$_{ab}$–Ag	5.9

Continued on next page

Table 5.1 – *Continued from previous page*

PDB Code	R Value	Res. (Å)	Max. B factor (Å^2)	Max. H3 B factor (Å^2)	Species	CDR H3 length	Light chain isotype	Fragment	pH
3i9g	0.192	1.90	53.87	29.27	Human-ized	14	$\varkappa$	F_{ab}–Ag	6.0
3giz	0.198	2.20	52.12	52.12	Mouse Human	15	$\varkappa$	F_{ab}	6.3
3go1	0.192	1.89	37.78	31.35	Human	16	λ	F_{ab}–Ag	6.5
1fns	0.172	2.00	49.64	20.80	Mouse	16	$\varkappa$	F_{ab}–Ag	—
1seq†	0.194	1.78	68.08	68.08	Mouse	16	$\varkappa$	F_{ab}	7.5
1gig	0.195	2.30	53.69	34.01	Mouse	16	λ	F_{ab}	—
3mlr	0.181	1.80	42.10	37.70	Human	17	λ	F_{ab}–Ag	5.5
4nzu	1.370	1.20	69.10	32.80	Human	18	$\varkappa$	F_{ab}	4.0
3lmj	0.194	2.20	58.27	58.27	Human	18	λ	F_{ab}	7.5
4f57	0.188	1.70	64.70	40.56	Human	18	λ	F_{ab}	6.5
2fb4	0.189	1.90	39.92	39.20	Human	19	λ	F_{ab}	—
3nps	0.190	1.50	49.76	31.95	Human	19	λ	F_{ab}–Ag	—

*Extended base geometry

†Unclear base geometry

Table 5.1: Structural information for the CDR H3 benchmark set. All structures have R-values lower than 0.2, resolution better than 2.5 Å and maximum B factors lower than 80.0 Å^2. For the purposes of controlling which variables are being considered, the CDR H3 loops are restricted to range from 9–20 residues in length and to only be derived from human and mouse antibodies.

Now that the memory restrictions have been alleviated, NGK can be fully tested on the new benchmark set of structures. Figure 5.3 shows the results of a *de novo* CDR H3 modeling simulation on an anti-citrullinated collagen type II antibody (PDB accession code 2w60[161]). In Figure 5.3A, a funnel plot shows the models ranked by the scaled score and colored by their base geometry. The kinked models make up a small fraction of the structures produced; however, they have lower scores than extended structures at the same RMSD value. The top-ranked models have very low RMSDs, but only three such models were produced. Nonetheless, because the score function successfully separates the near-native and non-native conformations, the discrimination score is -0.5710.

Figure 5.3B shows the τ_{101} and α_{101} values for the models (black) and the crystal

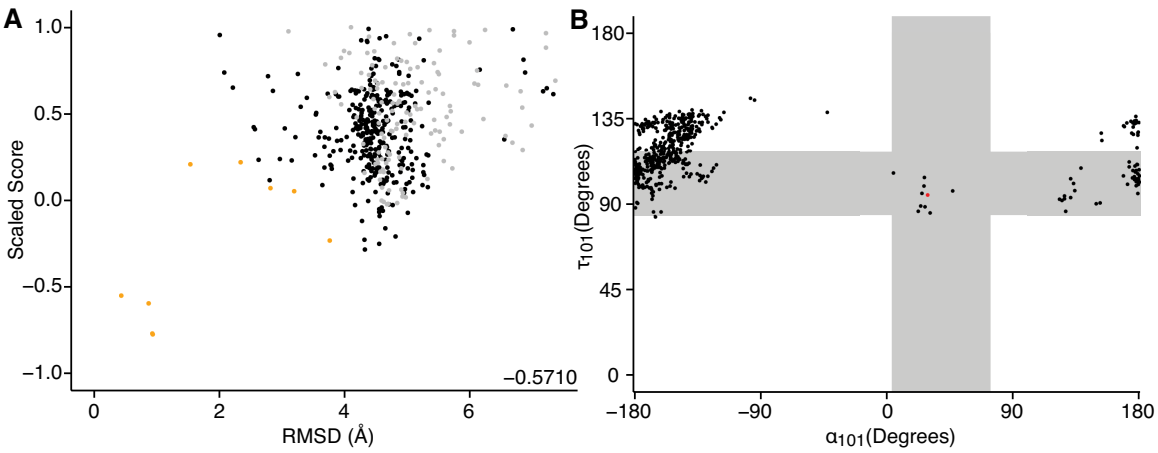

Figure 5.3: Results of unconstrained *de novo* NGK on an anti-citrullinated collagen type II antibody (2w60[161]). 2w60 is derived from a mouse, has an 11-residue H3 loop and a $\varkappa$ light chain. (A) Funnel plot showing scaled score vs. RMSD. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score of -0.5710 is shown in the lower right of the plot area. Very few kinked H3 models are produced, but the top-scoring models have sub-Ångström RMSDs. (B) τ_{101} vs. α_{101}. The • point is at the values of the native structure, and the • points correspond to the models. The vast majority of the points have τ_{101} and α_{101} values that correspond to extended conformations.

structure (red). The gray bars demarcate $\pm$ 3.0 σ from the mean of the distribution of each parameter in kinked antibodies as found in Chapter 4.[41] This plot shows a clear preference for NGK to produce H3 loops in the extended conformation, likely because it can form backbone–backbone hydrogen bonds in the low-resolution stage of modeling.

All of the CDR H3 loops in the benchmark set were modeled using the same flags. As shown in Table 5.2, the average scaled native score for kinked targets is -0.9480 with a standard deviation of 0.5033 (Table C.1). This means the native conformation has a significantly better score than the best decoys that are being produced by NGK.

Figures C.1 and C.2 in Appendix C show funnel plots and τ_{101} vs. α_{101} plots for the rest of the structures in the dataset. Across the whole set, kinked models represent a small fraction of the models that are produced, but those models tend to have lower RMSDs.

Simulation	Min. RMSD	Scaled Nat. Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
Unconstrained	1.3360	-0.9480	2.0180	3.6433	3.2174
Constrained	0.9332	-0.5264	1.2473	2.2179	2.0000
Combined	1.3789	-0.2396	1.8398	3.4148	2.7564

Table 5.2: Summary of *de novo* loop modeling simulations. The minimum RMSD, scaled native score, average of the top ten lowest RMSDs, average RMSD of the top 10 scoring models and the RMSD of the top-ranked model are shown for unconstrained NGK, constrained NGK and the combined NGK+CCD methods. Each value is the average of the values of the 44 kinked targets in the benchmark set. NGK with constraints proves to perform best over the whole set. Only the scaled native score in the combined simulations have superior values, however this affects the ability of the score function to discriminate between near-native and non-native conformations. Per-target values for each type of simulation can be found in Appendix C. All RMSDs are reported in Ångströms.

5.4.2 Constrained *de novo* modeling of CDR H3 loops

Because the kinked structures that are made have low RMSDs, biasing the simulation toward kinked conformations should increase the number of low-RMSD models produced in the course of the simulation. As described in methods, I used the parameters of the kink described in Chapter 4 to develop a kink constraint that can be employed during a simulation (Figure 5.1). Because the constraint potential is smooth and continuous, the conformation of a structure can be minimized with the constraint enabled.

Figure 5.4 shows the results of the constrained NGK simulation for anti-citrullinated collagen type II antibody (2w60[161]). From the τ_{101} vs. α_{101} plot shown in Figure 5.4B, it is clear that the constraint successfully biases the simulation to produce kinked structures. The τ_{101} and α_{101} values for the models (black) and the crystal structure (red). This plot shows that NGK with the kink constraint mostly produces kinked H3 loops. However, many

models are not kinked, which indicates that the simulations are not being over-constrained.

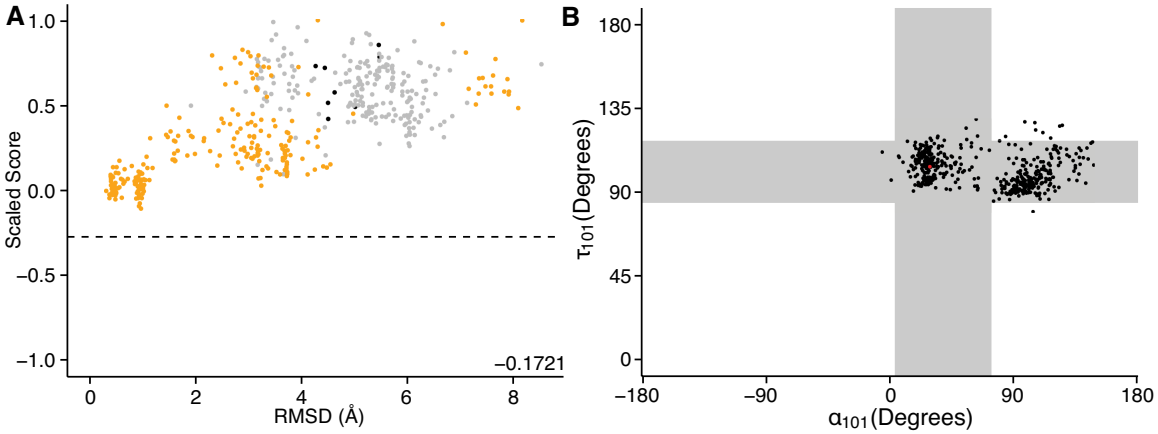

Figure 5.4: Results of constrained *de novo* NGK on an anti-citrullinated collagen type II antibody (2w60[161]). 2w60 is derived from a mouse, has an 11-residue H3 loop and a $\varkappa$ light chain. (A) Funnel plot showing scaled score vs. RMSD. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score of -0.1721 is shown in the lower right of the plot area. Many kinked H3 models are produced, and the top-scoring models have sub-Ångström RMSDs. The dashed horizontal line indicates the scaled score of the native structure, which is much higher compared to the unconstrained simulation. This indicates that the best-scoring models have a similar score to the native. (B) τ_{101} vs. α_{101}. The • point is at the values of the native structure, and the • points correspond to the models. While many of the points have τ_{101} and α_{101} values that correspond to kinked conformations, there are still many models that are not kinked. This suggests that the constraint has an appropriate penalty that can be overcome in cases with favorable interactions.

Figure 5.4A shows a funnel plot for the constrained NGK simulation of 2w60. The fraction of near-native structures has increased dramatically, demonstrating that generating more kinked structures is critical for successful CDR H3 predictions. The dashed horizontal line indicates the scaled score of the native structure, which was below the plotted bounds on the unconstrained plot. Therefore, the geometry of the models generated with constraints is more favorable than the geometry of the models generated by the unconstrained method.

However, because many more models between 1.0 and 3.0 Å RMSD are generated and those models score more favorably, the discrimination score for the constrained simula-

tion is -0.1721. Figure 5.4A shows that for many of the models with RMSDs < 2.0 Å, there is a model with RMSD 3.5 Å that scores as well. This result underscores the importance of producing many models even when using constrained NGK.

Figures C.3 and C.4 in Appendix C show funnel plots and τ_{101} vs. α_{101} plots for the rest of the structures in the dataset. Across the data set, with the exception of five targets (one of which has an unclear base geometry), the scaled native scores appear within the plot bounds, indicating that the scores of the models are close to the native structure. Figure C.4 shows that the four extended loops in the benchmark (1x9q, 2e27, 3liz, 3m8o) have sets of models that are predominantly kinked. While this is problematic, it appears that, with the exception of 1x9q, these particular targets were also not modeled successfully by unconstrained NGK, which confounds any analysis to determine if the constraint penalty could be overcome when appropriate.

The average RMSD of the 10 top-scoring models is lower with constraints in 40 of the 49 targets, and 30 targets have a lower RMSD of the top-scoring model. Without constraints, the average RMSD of the 10 top-scoring models is < 1.0 Å for three targets. This number increases to nine targets when using constraints. When considering only the top-scoring model, seven targets with and thirteen targets without constraint have RMSDs < 1.0 Å.

5.4.3 Generating low-RMSD models of CDR H3 loops

Although low-RMSD models can be produced by NGK with and without using constraints, the low-RMSD models produced by the constrained simulation have scores that are closer to the native score, even at very similar values of RMSD. This indicates that small deviations

in backbone geometry can have a substantial impact on the score of a model. To further probe the impact of small backbone perturbations, I used two methods: (1) CCD refine (small, shear moves followed by CCD closure); and (2) NGK refine (randomize non-pivot torsions, solve for the closed form). Both sets of simulations are performed starting with the native loop conformation and without using constraints. Figure 5.5 shows a comparison of CCD and NGK refinement for 2w60. The funnel plot uses the total score to allow the two methods to be directly compares with the orange points corresponding to models produced by CCD and the gray points being the NGK models, and the dashed horizontal line is the native score.

With few exceptions, CCD refinement produces models with better scores and lower RMSDs to the starting structure than NGK refinement. Both methods can produce models with a better score than the native structure. NGK does produce some models with lower scores than CCD, but at the expense of moving further from the native structure. The τ_{101} vs. α_{101} plot shown in Figure 5.5B shows a similar effect, with the CCD structures (red points) being more tightly clustered than the NGK structures (black points).

Figures C.5 and C.6 show the results of NGK refinement on the benchmark set, Figures C.7 and C.8 show the results of CCD refinement on the benchmark set, and Figures C.9 and C.10 show the comparison of CCD and NGK for the rest of the benchmark set. The aforementioned trends hold across the dataset.

These results show that the high-resolution stage of NGK allows more movement of the loop. It could be said that while CCD is refining the structure, NGK continues to perform more sampling. In the case of *de novo* CDR H3 modeling, it is unclear if the high-

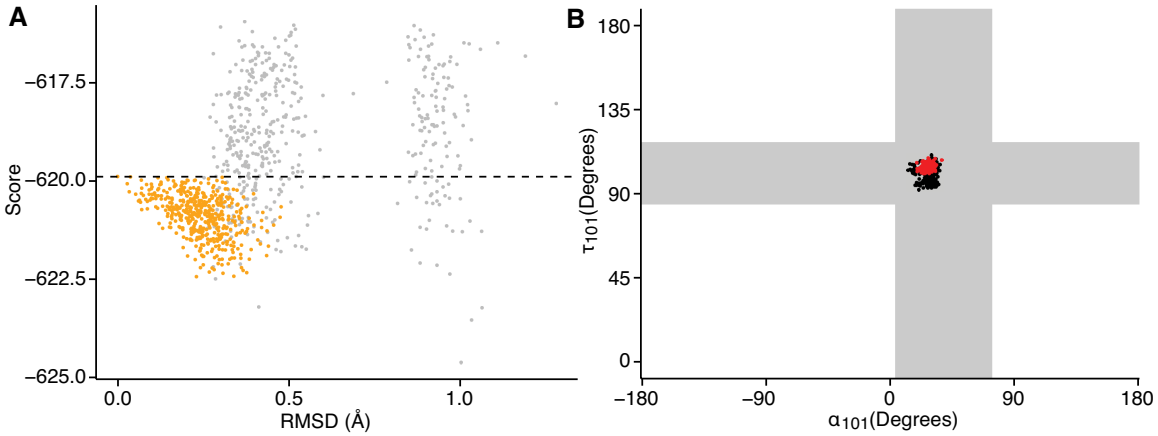

Figure 5.5: Results of NGK and CCD refinement on an anti-citrullinated collagen type II antibody (2w60[161]). (A) Funnel plot showing total score vs. RMSD. • points correspond to models produced by CCD, • points to models produced by NGK, and the dashed horizontal line indicates the score of the native structure. All of the CCD models have a more favorable score than the native structure and very low RMSDs ($\leq$ 0.5 Å), while many of the NGK models have higher scores than the native structure and and larger RMSDs than the CCD models. (B) τ_{101} vs. α_{101}. The • points correspond to models generated with CCD, and the • points correspond to models generated with NGK. The CCD models are distributed in a smaller region of the plot than the NGK models. Together these plots show that the high-resolution phase of NGK is performing more sampling, while the high-resolution phase of CCD is refining the input structure.

resolution stage should be doing more sampling or refining the structure that is produced in the low-resolution stage.

5.4.4 Combined NGK+CCD

Because the high-resolution stage of CCD focuses more on refining the structure of model developed in the low-resolution stage of modeling, combining it with the low-resolution stage of NGK with constraints may prove to be a useful strategy for generating high-quality models. To test this, I ran a combined simulation with constraints on the H3 loop benchmark set. Figure 5.6 shows the results of the combined method for 2w60. The funnel plot in Figure 5.6A shows that many low-RMSD models are being produced and their score is

approaching that of the native structure. However, many other structures have scores that approach the native score, even at an RMSD of 4.0 Å. Additionally, there are many more structures with unclear base geometries than in the constrained NGK simulation. In Figure 5.6B the increased diversity in τ_{101} and α_{101} values is even more pronounced. In addition to the kinked structures that are broadly distributed about the native values, there is a large cluster of models with unclear base geometric centered near τ_{101} and α_{101} of 90°.

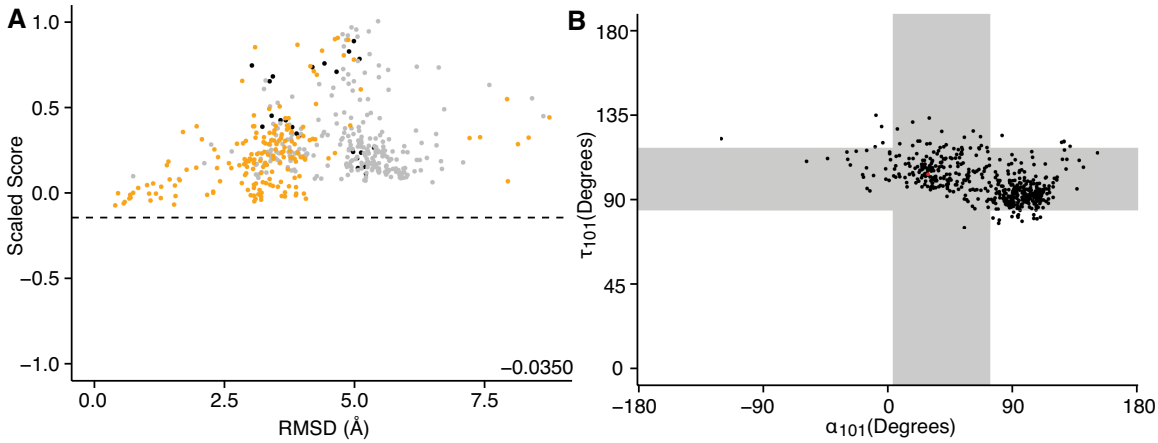

Figure 5.6: Results of constrained *de novo* NGK+CCD on an anti-citrullinated collagen type II antibody (2w60[161]). (A) Funnel plot showing scaled score vs. RMSD. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score of -0.0350 is shown in the lower right of the plot area. Compared to the results of the constrained NGK simulation shown in Figure 5.4, there are many more models with unclear base geometries and the discrimination score is much higher. The dashed horizontal line indicates the scaled score of the native structure, which is higher than in the unconstrained and constrained simulations. This shows that the score of the models is approaching that of the native structure, however many models across a wide range of RMSD values have similar scores. (B) τ_{101} vs. α_{101}. The • point is at the values of the native structure, and the • points correspond to the models. The points are much more widely distributed than in the constrained simulation and appear to cluster into two large groups, indicating that CCD was able to find a conformations that could offset the penalty imposed by the constraint. Because CCD focuses on refining the loop structure, the scores are better than when NGK is used, but it appears that the degree of refinement eliminates the ability of the score function to identify near-native conformations.

As shown in Figure C.11, the scores of the models are much closer to the native

score across the benchmark set. This translates to an increase in the scaled native score and in the discrimination score. If a discrimination score < 0.0 is used as the criterion for a successful simulation, the combined loop modeling mode has twenty-five successes, while constrained NGK has thirty-three. By this loose criterion, the combined method is less successful than constrained NGK.

That being said, evaluating sampling and scoring as separately as possible is useful to point out where deficiencies in the method lie. The minimum RMSDs for each target are lower with constrained NGK, with only four of the kinked targets achieving better RMSDs with the combined method. This likely means either (1) NGK is under-sampling conformation-space in centroid mode, so continuing to sample in all-atom (NGK) as opposed to refining that structure (CCD) yields better results, or (2) NGK may be attempting to move the loop too much, resulting in most of those moves being rejected except in cases where conformation moves substantially toward the native structure, resulting in better discrimination by score.

5.4.5 Considering pH effects

Table 5.1 includes the pH at which the crystallization experiment was performed if available. The effect of the pH on the structure of a protein, if any, is often impossible to discern due to a variety of factors, including the fact that the protein may not crystallize equally well at two significantly different pHs. The benchmark set contains thirteen structures that were crystallized at a pH below 6.0 and seven structures crystallized at a pH above 8.0. Among these more extreme-pH structures is 1dlf,[162] an anti-dansyl antibody, which has a histidine in its CDR H3 loop and was crystallized at pH 5.25. Interestingly, another

high-quality crystal structure for the same antibody but at a higher pH (6.75) was produced (PDB accession code 2dlf[162]) as part of the same study by Nakasako *et al.* for the purpose of studying pH-effects. Nakasako *et al.* found that the structure of the antibody remained the same except for the CDR H3 loop, which undergoes a pH-dependent conformational change, presumably controlled by the protonation state of the histidine within the loop. The rest of this section focuses on this anti-dansyl antibody because it has an ionizable residue within the loop and the conformations of the loop with the protonated and deprotonated histidine are known.

To test whether or not Rosetta and NGK can capture this pH-dependent change, I performed two calculations: (1) a constrained NGK simulation with the pH-aware packer and `e_pH` score term enabled;[163] and (2) recomputing the RMSDs of the models produced from the previously run constrained NGK simulation with the higher pH structure (2dlf) as the reference structure. As shown in Table 5.3, the lowest-RMSD structure produced by the pH-aware simulation has an RMSD of 0.94 Å, which is slightly worse than the the pH-naïve constrained simulation, which produces a model with RMSD 0.88 Å. The pH-aware packer's inability to correctly predict this structure is likely because most of the large structural perturbations occur in centroid mode, where the protonation state cannot be sampled. After this, the loop has an additional charge to accommodate, which then becomes a driving factor in the score. Using 2dlf as the reference structure results in twenty-six models with lower RMSDs than the lowest-RMSD model from constrained NGK with 1dlf as the reference. The average RMSD of the top-10 lowest RMSD structures is 0.66 Å vs. 0.95 Å when using 2dlf and 1dlf, respecively, as the reference.

Although the best-sampled structures would suggest that Rosetta successfully predicts the conformation at the pH closer to physiological conditions, the scores reveal a slightly more complicated story. While the average RMSD of the 10 top-scoring models is lower with 2dlf as the reference (0.9717 Å vs. 1.2657 Å), the RMSD of the top-ranked structure is higher (1.2125 Å vs. 0.8847 Å). Interestingly, the lowest-RMSD structure with 2dlf as the reference is the tenth best-scoring, and the lowest-RMSD structure with 1dlf as the reference is the best-scoring model. This scoring anomaly underscores the importance of considering multiple models simultaneously when making predictions based on the results of a Rosetta simulation.

Simulation	Min. RMSD	Scaled Nat. Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
NGK+cst	0.88	-0.52	0.95	1.27	0.88
pH Mode	0.94	-0.57	1.02	1.69	1.38
2dlf	0.53	-0.52	0.66	0.97	1.21

Table 5.3: The ability to capture pH-dependent conformational changes is shown by comparing the RMSDs of the models from constrained NGK simulations with (1) the low-pH structure as the reference for RMSD calculations; (2) using a pH-aware variant of the packer to sample protonated residues; and (3) the high-pH structure as the reference for RMSD calculations. The pH-aware simulation performs worse than the others, likely due to the amount of sampling that is performed in centroid mode. Using 2dlf as the reference structure yields the best results, which suggests that Rosetta favors structures that are present at physiological conditions. However, the top-ranked structure is closer to the low-pH conformation (1dlf) than to the higher-pH conformation (2dlf). All RMSDs are reported in Ångströms.

5.4.6 Utility of the new method

Accurately predicting CDR H3 conformations on a crystal framework is an important step toward the ultimate goal of predicting antibody structures. The incentive for improving

CDR H3 structure prediction in the context of the experimentally-determined framework is the promise of improving homology modeling and producing models of sufficient quality to be useful in downstream applications, namely antibody–antigen docking. In this section I present some proof-of-concept calculations to demonstrate the utility of using NGK with a kink constraint.

HOMOLOGY MODELING WITH CONSTRAINTS

In Chapter 3,[39] I presented the performance of RosettaAntibody in a blind prediction challenge. One of the takeaways from that study is the need to force Rosetta to generate kinked structures in the *de novo* loop modeling stage of the simulation. In that study, this was accomplished by using a filter based on Shirai *et al.*'s description of the kink.[24,25]

I modified RosettaAntibody to apply constraints to the *de novo* loop modeling phase and enabled the neighbor-dependent Ramachandran map sampling from NGK. This modification will enable the H3 loop to be constrained with the kink constraint developed in section 5.3.2 during the simulation. Figure 5.7 shows cumulative density estimates for RosettaAntibody with a kink filter (gray curve) and with the new kink constraint (orange curve) for 2w60. Both methods can generate low-RMSD models of the H3 loop, but with the kink constraint, 1106 of the 2000 models have H3 RMSD < 2.0 Å as opposed to only 796 with the filter.

This shows that the kink constraint leads to sampling improvements even in cases where RosettaAntibody is already successful. Comparable results can be achieved while generating fewer models, and additional simulation time can be spent performing other stages of modeling, *i.e.* V_L–V_H optimization.

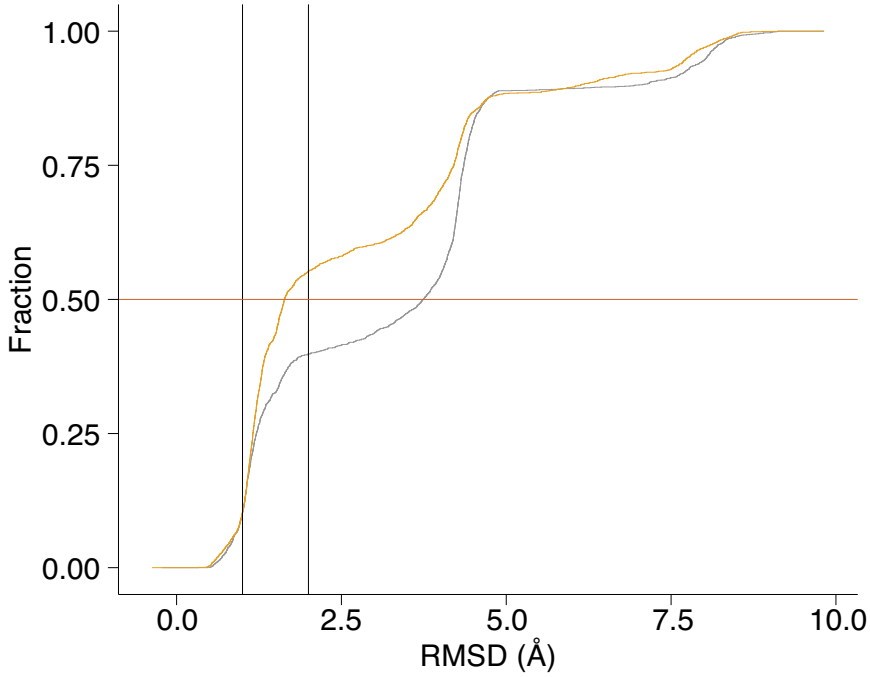

Figure 5.7: Modeling CDR H3 on a homology framework for 2w60. A cumulative density estimate of the RMSD of the backbone atoms in the CDR H3 loops of homology models built using the method described in Chapter 3 (gray) and with the new kink constraint (orange). Dashed vertical lines indicate the fraction of models with RMSD of 1.0 and 2.0 Å or better for each method. The red dashed line shows that 50% of the models produced by the standard method have an RMSD of 3.74 Å or lower, while 50% of the models from the method that exploits the kink constraint have an RMSD of 1.64 Å or lower. Although both methods are successful in producing some low-RMSD models, a significantly larger fraction are produced when using the kink constraint based on τ_{101} and α_{101} as opposed to the filter based solely on α_{101}.

DOCKING WITH MODELED H3 LOOPS

Successful docking is highly dependent on having accurate models of the bound conformation of each binding partner. To test whether or not the H3 loop conformations predicted using constrained NGK are accurate enough for binding, I focused on 2adf, which is crystallized with its antigen. The CDR H3 loop in 2adf is 11 residues and the constrained NGK simulation has a discrimination score of -0.1758, indicating a successful prediction.

I selected the top ten models by Rosetta score as an ensemble to dock to the bound form of the antigen using EnsembleDock.[31] EnsembleDock functions by cycling through a set of distinct backbone conformations after each rigid-body move during the low-resolution stage of docking. Each member of the ensemble is scored, and the best-scoring conformation observed in the low-resolution stage is the starting point for all-atom refinement.

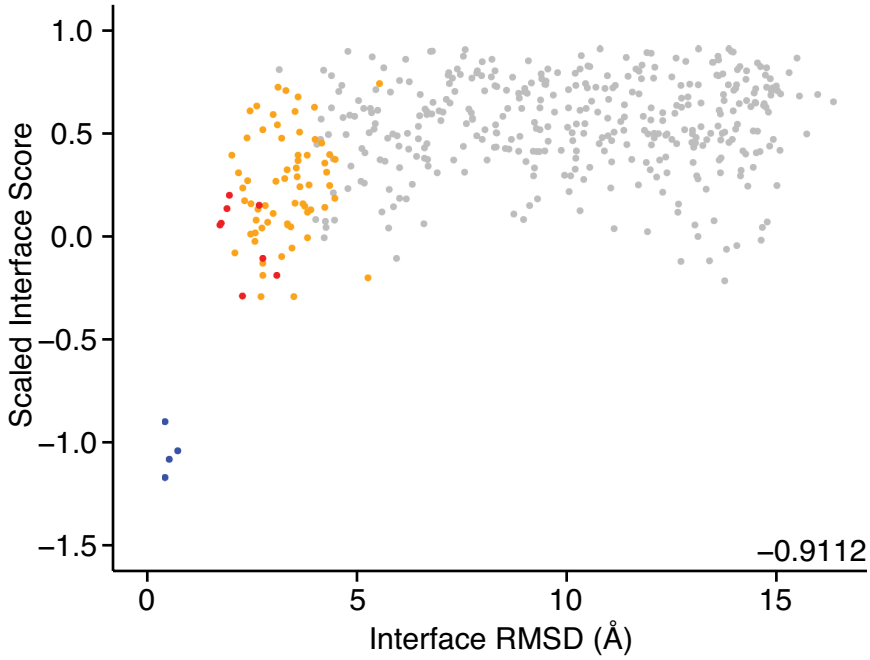

Figure 5.8: Results of docking an antithrombotic antibody to its antigen (2adf[164]). Funnel plots showing scaled Interface Score vs. Interface RMSD. Interface score is calculated as the score of the unbound partners subtracted from the score of the complex. Interface RMSD is the RMSD of the backbone atoms of the residues within 8.0 Å of a residue on the other docking partner. The points are colored using the CAPRI quality ratings,[165] where ● points correspond to incorrect structures, ● points to acceptable, ● points to medium, and ● points to high-quality models. EnsembleDock using CDR H3 loops modeled with constrained NGK results in 65 acceptable models, 8 medium-quality models and 4 high-quality models. The high-quality models are clearly separated from the other models by interface score as evidenced by the discrimination score of -0.9112 sown in the lower right corner of the plot area.

The EnsembleDock results are shown in Figure 5.8. Points are colored to indicate the Critical Assessment of PRedicted Interactions (CAPRI) quality rating[165] of each

model, with gray points corresponding to incorrect structures, orange to acceptable quality, red to medium-quality and blue to high-quality models. For a model to be considered a high-quality prediction by CAPRI metrics, the fraction of native residue–residue contacts recovered (f_{nat}) must be ≥ 0.5 and the Interface RMSD (I_RMSD) or Ligand RMSD (L_RMSD) must be ≤ 1.0 Å. Medium-quality predictions must have $f_{nat} \geq 0.3$ and L_RMSD ≤ 5.0 Å or I_RMSD ≤ 2.0 Å, while acceptable predictions have $f_{nat} \geq 0.1$ and L_RMSD $\leq$ 10.0 Å or I_RMSD ≤ 4.0 Å. Models that have $f_{nat} \leq 0.1$ or L_RMSD ≥ 10.0 Å and I_RMSD ≥ 4.0 Å are considered incorrect. The 10 top models by interface score encompass one incorrect model, four acceptable models, one medium-quality model and four high-quality models. The fraction of native contact recovered, ligand RMSD and interface RMSD for the 10 top-ranked models are shown in Table 5.4.

The ten models used in the ensemble have scores ranging from -594.19 to -586.90, and the average H3 RMSD is 1.48 Å. The top-ranked model has a loop RMSD of 1.53 Å, and the eighth structure in the set has an H3 RMSD of 0.75 Å. As shown in Table 5.4, the 10 top models do not converge on a single member of the ensemble, showing that considering several models simultaneously is a path forward.

5.5 Discussion and Conclusions

In this chapter I present the results of applying a constraint based on the kink parameters determined in Chapter 4 to *de novo* CDR H3 loop modeling simulations. Successful structure prediction required (1) developing a penalty that can be expressed as a differentiable function to enable energy minimization; and (2) redesigning the underlying data structures

Rank	Interface Score	Interface RMSD	Ligand RMSD	f_{nat}	H3 RMSD	CAPRI rating
1	-8.79	0.43	4.00	0.91	1.48	High
2	-8.51	0.54	1.78	0.86	0.75	High
3	-8.22	0.74	2.57	0.89	1.59	High
4	-7.67	0.43	1.60	0.91	1.53	High
5	-5.54	2.72	9.25	0.26	1.69	Acceptable
6	-5.55	3.50	9.49	0.29	0.75	Acceptable
7	-5.53	2.28	7.56	0.57	1.54	Medium
8	-5.27	13.75	25.24	0.11	1.54	Incorrect
9	-5.22	5.27	7.33	0.20	1.69	Acceptable
10	-5.18	2.76	9.38	0.34	1.53	Acceptable

Table 5.4: Summary of top 10 models produced by EnsembleDock. Interface score is calculated as the score of the unbound partners subtracted from the score of the complex. Interface RMSD is the RMSD of the backbone atoms of the residues within 8.0 Å of a residue on the other docking partner. Ligand RMSD is the backbone atom RMSD of the antigen after superposing the antibody to the native structure f_{nat} is the fraction of native residue–residue contacts recovered, where contacting residues are defined as residues on opposite binding partners within 5.0 Å of each other. The H3 RMSD column shows the RMSD of the CDR H3 loop of the model that was ultimately selected by EnsembleDock to generate the docked model. Interestingly, a different member of the ensemble was used in each of the high-quality docked models. All RMSDs are reported in Ångströms.

used by the neighbor-dependent Ramachandran sampling method of NGK. Although the CCD refinement protocol can achieve lower scores, the best structure results come from using NGK for both the low-resolution and all-atom stages of the simulation.

Part of this study required constructing a set of high-resolution CDR H3 loops from crystal structures. Not all CDR H3 loops meet the strict quality cutoffs that were used in this study. It is possible that some of the loops that meet these criteria are simply more stable or rigid than some other H3 loops. If that is the case, could that translate into these loops being easier modeling targets? While this is possible, it remains an unanswered question. However, since the atomic coordinates of these loops are well-defined, structural comparisons between models and the loops have real meaning

While there are not enough loops at each length to draw conclusions about prediction performance as a function of length, it is worth noting that the longest loop where the average RMSD of the 10 top-scoring models < 1.0 Å without constraints is 13 residues, and the longest with constraints is 14. At first glance this difference seems small, but the RMSD of the top-scoring model for some longer loops reveals the extent to which the kink constraint improves the performance of *de novo* loop modeling. For example, for the 19-residue CDR H3 loop in 2fb4, the RMSD without constraints is 14.67 Å, with constraints it's 3.63 Å. While this is a big improvement, it is unlikely that these models would lead to successful docking simulations. It is unclear if further improvements will come from using more cycles of NGK for longer loops, generating more models or incorporating knowledge of additional local structures (*e.g.* β turns) in addition to the kink into the simulation.

Some of these crystals were formed at pHs that deviate substantially from physio-

logical conditions, leading me to test the ability of the pH-aware side chain packer to model loops in these conditions. For the example I tested, the pH-unaware method produces results that are closer to the conformation of the loop in a crystal at a pH closer to physiological conditions. Currently the pH-effects cannot be captured, likely because the majority of the conformational sampling occurs in the low-resolution stage where side-chain protonation states are not represented. A possible path forward may be to update the "pair" score term by calculating statistics with separate counts based on predicted protonation states of ionizable residues. For the purposes of benchmarking, the crystallization conditions of the structures in the benchmark set should be considered.

The amount of sampling performed in the low-resolution stage raises another issue. Some important interactions are mediated by side-chain interactions, which may be lost entirely in a loop modeling simulation. Some modeling methods have been developed that operate in all-atom mode throughout the entire simulation. One such method available in Rosetta is step-wise assembly (SWA),[90] which builds the loop one residue at a time. While this method has shown promise, it is extremely computationally expensive and is therefore not well-suited to antibody homology modeling tasks that rebuild the H3 loop while simultaneously sampling V_L–V_H orientation. However, an all-atom loop-modeling routine may enable Rosetta to capture pH-effects as well as critical side-chain interactions.

While this study focuses on modeling CDR H3 loops on the crystallographic framework, the ultimate test of the utility of a new loop modeling method in the context of antibody modeling is predicting CDR H3 conformations on a homology framework. I tested the new method on a homology modeled framework by comparing the distribution

of CDR H3 RMSDs from the standard method, which uses a filter based on α_{101}, and the new constrained method, which evaluates a potential based on both τ_{101} and α_{101}. The constrained method produces a substantially larger fraction of low-RMSD models, which should enable the development of new protocols that focus more time on other aspects of antibody modeling, *e.g.* V_L–V_H orientation optimization. Additionally, the same performance as the current RosettaAntibody should be expected by generating fewer models.

Another goal for antibody structure prediction is to generate models of sufficient quality to be used in downstream applications, namely antibody–antigen docking. To assess the quality of the predicted H3 conformations, EnsembleDock was used with an ensemble of the 10 top-scoring models and the bound conformation of the antigen. The simulation correctly predicts the conformation of the complex. This is an idealized case because the CDR H3 loop was modeled on the crystal framework and the bound form of the antigen was used, but the successful simulation serves as important motivation moving forward. The success of both the homology modeling and docking simulations is encouraging and these simulations should be expanded to a larger set of structures to fully assess the implications of the kink constraint. In summary, a constraint developed by studying structures of H3 loops has enabled accurate *de novo* structure prediction of the most diverse region of antibodies.

CHAPTER VI

CONCLUSION

6.1 My Contributions

Antibodies are important immunological molecules that will have an ever larger role in pharmaceutical and biotechnological contexts. Developing the therapeutic antibodies, designed vaccines and biosensors of the future will be accelerated through an increased understanding of antibody structures. High-throughput sequencing methods coupled with computational tools will help paint the most complete picture to date of antibody diversity both in terms of sequence and structure. Atomically accurate structural models of antibodies, particularly of the paratope, will be critical inputs to downstream computational methods such as antibody–antigen docking algorithms.

My graduate research has focused on developing methods to improve protein structure prediction. Chapter 2 summarizes the development of several tools that assist in developing new methods by establishing a more structured organization of Rosetta, developing a new visualization technique[47] and controlling access to the degrees of freedom to which the simulation has access. Additionally, this chapter presented work that extended Rosetta to new classes of proteins (membrane proteins) and reorganizes a widely used

application (RosettaDock[42]) using object-oriented design principles.

In Chapter 3, I summarized my previously-published work[39] in predicting the structures of eleven unpublished crystal structures. I found that the existing method, RosettaAntibody, was adept at selecting low-RMSD structural templates when they are available, but does not produce models with CDR H3 conformations that are accurate enough to be used as inputs for another simulation. This prompted me to analyze CDR H3 structures to determine which factors contribute to the observed structural diversity of the loop.

My previously-published work studying CDR H3 structures[41] is presented in Chapter 4. I found that the previously observed C-terminal kink is actually a feature that is encoded into the immunoglobulin heavy chain and is not determined solely by the sequence of the CDR H3 loop. The kink serves to disrupt β-strand pairing at the base of the loop, which enables loops of the same length to adopt dramatically different conformations. I also found that other protein families, namely PDZ domain containing proteins, have evolved to use the same C-terminal kink for the same purpose of recognition and binding. Along the way, I developed the most detailed description of the geometry of the kink to date using two parameters (α_{101} and τ_{101}) that can be used to assess the quality of CDR H3 models.

Finally in Chapter 5, I use the description of the kink developed in Chapter 4 to develop a new kink constraint that can be applied during a *de novo* loop modeling simulation. Using this constraint with the next-generation KIC loop modeling method[89] dramatically increases the number of kinked and low-RMSD models of CDR H3 loops. Furthermore, this method produces models of sufficient quality for docking simulations

using EnsembleDock,[31] and leads to a substantial increase in the number of low-RMSD H3 loops produced with RosettaAntibody.[22,39]

While the focus of this dissertation is on identifying weaknesses in existing antibody homology modeling methods, collecting data from known structures to address those weaknesses, and ultimately developing an improved CDR H3 structure prediction method, the end goal of improving the accuracy of antibody homology modeling methods remains as potent as ever. To address this, we will incorporate the newly developed methods into an all-new version of RosettaAntibody, which will be available as a stand-alone application in the Rosetta software suite,[32] a web server powered by ROSIE[112] and through PyRosetta.[45]

6.2 Future Directions

In Chapter 3, I found that CDR H3 modeling was not the only problem plaguing antibody structure prediction methods. Specifically, predicting CDR L3 on λ light chains, and potential differences in canonical CDR loop conformations of antibodies from different species have proven difficult.

The failed prediction of an antibody from a rabbit was discussed in Chapter 3, and I pointed out that a complicating factor in predicting rabbit antibody structures is the paucity of structural models of rabbit antibodies developed from experimental data. One factor I did not mention, though, is that there is evidence that rabbits undergo a process called gene conversion,[166] wherein V(D)J recombination is predominantly restricted to a single V and J gene segment and one of several pseudogene segments may be incorporated into the cDNA as the major source of diversity in the antibody repertoire. While other mammalian

species, including mice[167] and humans,[168] have been shown to use gene conversion at low levels, rabbits appear to rely on it almost entirely.[166] Chickens are also known to use gene conversion in this way, which will likely complicate structure prediction of chicken antibodies.[169–172]

Gene conversion is not the only species-specific antibody oddity. Cows produce antibodies with extremely long CDR H3 regions that are highly enriched with cysteines that adopt, as Wang *et al.* deemed them, "stalk and knob" conformations.[173] Camelids[174] and cartilaginous fishes[175] produce heavy-chain antibodies, which are antibodies with two heavy chains and no light chains.

Accounting for these and other species-specific effects will continue to be necessary to improve the accuracy of antibody structure prediction methods. For example, North *et al.*[19] found that the cluster membership of an eleven-residue CDR L1 loops has a strong species dependence. This species preference can be explained by an explicit interaction between the loop and framework residue L71. Human antibodies have a phenylalanine at position 71, while mice have either phenylalanine or tyrosine, the latter of which forms a hydrogen bond with eleven-residue CDR L1 loops. Thus, I expect that incorporating this observation into the template section stage of antibody modeling will improve the accuracy of RosettaAntibody. In addition to performing analyses to uncover these kinds of interactions, one must pay careful attention to new studies conducted by other research groups.

In Chapter 3, I showed that the unavailability of template structures is the leading cause of modeling failures. To address this, we must communicate with crystallographers

and NMR spectroscopists to explain which structures will be most beneficial for our purposes and to gain access to unpublished structures. Blinded tests are the best simulation of real-world use cases and force method developers to grapple with a tool's true strengths and weaknesses.

In Chapter 4 I showed that different statistical analyses of the same data can lead to substantially different conclusions. Moving forward, we must continue to develop and make use of bioinformatics techniques on structures and sequences in order to determine what the data are really showing.

While this dissertation has focused on the C-terminal kink in CDR H3 loops, there are other local structures within loops that could be exploited to further improve sampling strategies. A common motif in loops is the β turn,[176–181] which is a four-residue segment where the backbone turns such that the C_α–C_α distance between residues i and $i+3$ of the motif is less than 7 Å. Preliminary calculations show that in an average 18-residue loop, there will be two to three turns (15.2% of the 15 overlapping four-residue intervals). Each turn restricts the torsion angles of the central two turn residues. Thus, the 36 degrees of freedom in the loop could be reduced (assuming non-overlapping turns) to as few as 24 degrees of freedom ($36 - 3 \times 4$), equivalent to a 12-residue loop prediction problem. This reduction in the number of degrees of freedom provides another path forward for accurate *de novo* modeling of increasingly long CDR H3 loops.

In order to continue to push the boundaries of what can be accurately modeled, we must continue to make efficient use of supercomputing resources. A principle known as Moore's law[182] states that number of transistors on an integrated circuit (IC), and thus the

computing power of that IC, doubles roughly every two years. In recent years, Moore's law has broken down for serial execution of instructions, which has led to the development of multi-core processors that can make up the difference through parallel execution of code. Unfortunately, taking advantage of parallelism requires writing code that can be split into several parallel tasks while maintaining internal integrity, which can require significant effort. Additionally, new massively parallel hardware resources, specifically GPUs, have become available for general purpose calculations through the advent of OpenCL[183] and CUDA.[184] A popular molecular dynamics package, NAMD,[185] has already been adapted to use GPU-acceleration for its Coulombic and non-bonded force calculations leading to a 100 fold speedup in those calculations and 10 fold overall increase in performance.[186] I believe that in order to solve increasingly complex problems, we will need to take full advantage of these massively parallel computational resources.

APPENDIX A

COMPLETE LIST OF LAT+KINK MATCHES

Accesion Code	ChainID	N anchor	C anchor	Pfam
3kt7	A	190	206	2OG-FeII_Oxy_3
1vgj	A	80	89	2_5_RNA_ligase2
2pzh	A	64	90	4HBT
2ov9	A	191	201	4HBT
1vh9	A	117	127	4HBT
3r87	A	93	104	4HBT
3e1e	A	117	130	4HBT
2xem	A	79	105	4HBT_2
2gf6	A	95	104	4HBT_2
2oiw	A	88	97	4HBT_2
1tbu	A	99	108	4HBT_3
2z1a	A	442	456	5_nucleotid_C
4jmd	A	120	137	ADC
3qwb	A	45	70	ADH_N
2wek	A	72	99	ADH_N
1g8m	A	435	444	AICARFT_IMPCHas
1zcz	A	359	368	AICARFT_IMPCHas
3nyq	A	283	293	AMP-binding
3etc	A	349	359	AMP-binding
2v7b	A	331	342	AMP-binding
3rix	A	339	350	AMP-binding
4eat	A	325	336	AMP-binding
1amu	A	322	333	AMP-binding
3fce	A	293	304	AMP-binding
2d1s	A	341	352	AMP-binding
3vnr	A	306	317	AMP-binding
3kxw	A	324	335	AMP-binding

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
1jdp	A	372	388	ANF_receptor
3ats	A	255	264	APH
2cu9	A	99	108	ASF1_hist_chap
1roc	A	98	107	ASF1_hist_chap
2dwc	A	235	244	ATP-grasp
1kjq	A	223	232	ATP-grasp
3k5i	A	209	218	ATP-grasp
3mwd	A	202	216	ATP-grasp_2
2fp4	B	205	220	ATP-grasp_2
2nu8	B	198	213	ATP-grasp_2
3t7a	A	308	321	ATP-grasp_4
3sty	A	140	152	Abhydrolase_6
3sk3	A	218	227	Acetate_kinase
4h0p	A	21	32	Acetate_kinase
3gy9	A	112	137	Acetyltransf_1
2q0y	A	56	65	Acetyltransf_1
1z4e	A	59	68	Acetyltransf_1
2dxq	A	55	64	Acetyltransf_1
3dsb	A	69	78	Acetyltransf_1
3f8k	A	58	67	Acetyltransf_1
3fnc	A	61	70	Acetyltransf_1
2pdo	A	47	56	Acetyltransf_1
2x7b	A	55	64	Acetyltransf_1
2gan	A	71	80	Acetyltransf_1
1cjw	A	83	92	Acetyltransf_1
4l8a	A	54	63	Acetyltransf_1
3c26	A	63	72	Acetyltransf_1
2qec	A	64	74	Acetyltransf_1
4kvx	A	44	54	Acetyltransf_1
2vez	A	98	108	Acetyltransf_1
3te4	A	76	86	Acetyltransf_1
4ag7	A	72	83	Acetyltransf_1
3t90	A	55	66	Acetyltransf_1
2o28	A	88	99	Acetyltransf_1
2b5g	A	56	73	Acetyltransf_1
3r8y	A	206	217	Acetyltransf_11
2ree	A	288	313	Acetyltransf_4
4jxr	A	57	67	Acetyltransf_4
3gkr	A	243	256	Acetyltransf_6
4ii9	A	243	256	Acetyltransf_6
2jdc	A	43	52	Acetyltransf_7
1xmt	A	26	41	Acetyltransf_CG

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
2rfq	A	153	162	Acyl-CoA_dh_M
2h30	A	144	154	AhpC-TSA
4eo3	A	106	116	AhpC-TSA
3gkn	A	126	136	AhpC-TSA
3lwa	A	175	185	AhpC-TSA
4grf	A	362	372	AhpC-TSA
4fo5	A	141	151	AhpC-TSA
3ros	A	5	16	Aldedh
3r31	A	33	44	Aldedh
4kna	A	19	30	Aldedh
3ju8	A	20	31	Aldedh
4h7n	A	6	17	Aldedh
3vz1	A	4	15	Aldedh
3u4j	A	39	51	Aldedh
4e3x	A	80	92	Aldedh
2v9l	A	91	106	Aldolase_II
1z45	A	499	526	Aldose_epim
1yga	A	138	165	Aldose_epim
2cir	A	216	225	Aldose_epim
1vav	A	102	117	Alginate_lyase2
3n40	P	151	174	Alpha_E2_glycop
3eyp	A	245	270	Alpha_L_fucos
4f0r	A	373	403	Amidohydro_1
3ooq	A	130	145	Amidohydro_4
4ig1	A	215	228	ApbE
4gve	A	387	396	Arena_nucleocap
3q7c	A	390	399	Arena_nucleocap
1vra	A	28	40	ArgJ
3it4	A	26	39	ArgJ
1vl2	A	347	360	Arginosuc_synth
1g4m	A	234	254	Arrestin_C
3ugu	A	240	260	Arrestin_C
3fz4	A	96	105	ArsC
3no2	A	265	275	Arylsulfotran_2
1b5f	A	80	106	Asp
1j71	A	90	116	Asp
1mpp	A	80	106	Asp
4i0e	A	137	165	Asp
3fv3	A	84	110	Asp
3vf3	A	76	104	Asp
4b78	A	76	104	Asp
3pvk	A	90	116	Asp

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3zkq	A	92	120	Asp
2-Apr	A	201	210	Asp
3k1w	A	207	216	Asp
3qp4	A	59	74	Autoind_bind
2rhs	B	311	320	B3_4
2bf6	A	621	644	BNR_2
4j9t	A	98	118	BNR_2
2bf6	A	298	318	BNR_2
2sli	A	679	700	BNR_2
4j9t	A	348	359	BNR_2
2jkb	A	625	640	BNR_2
1nnx	A	86	103	BOF
1vqz	A	123	132	BPL_LplA_LipB
3rkx	A	180	189	BPL_LplA_LipB
2dxu	A	104	113	BPL_LplA_LipB
2c8m	A	138	147	BPL_LplA_LipB
2eay	A	96	105	BPL_LplA_LipB
3dzw	A	32	41	B_lectin
3a0e	A	33	42	B_lectin
1b2p	A	43	52	B_lectin
1xd5	A	32	41	B_lectin
4gc1	A	47	56	B_lectin
1xd5	A	94	104	B_lectin
4h3o	A	64	74	B_lectin
3a0e	A	64	74	B_lectin
3dzw	A	63	73	B_lectin
1b2p	A	74	84	B_lectin
2okx	A	840	859	Bac_rhamnosid
4a8u	A	67	87	Bet_v_1
2qim	A	66	86	Bet_v_1
3i7j	A	79	96	Beta-lactamase
1m40	A	249	259	Beta-lactamase2
1jz8	A	841	850	Bgal_small_N
1yq2	A	934	943	Bgal_small_N
4dou	A	456	465	C1q
1o91	A	676	685	C1q
4dou	A	315	324	C1q
4dou	A	174	183	C1q
1gr3	A	613	622	C1q
2wnv	C	96	114	C1q
2fk9	A	46	55	C2
1tjx	A	315	324	C2

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3vvv	A	108	118	CALCOCO1
3p6b	A	122	134	CBM_4_9
1goi	A	473	496	CBM_5_12
1t4w	A	380	390	CEP1-DNA_bind
3a2z	A	144	163	CHAP
3eif	A	848	858	CHU_C
1hq0	A	993	1008	CNF1
2dyu	A	119	129	CN_hydrolase
2uxy	A	120	130	CN_hydrolase
1uf5	A	112	122	CN_hydrolase
3hkx	A	97	107	CN_hydrolase
3ouz	A	230	240	CPSase_L_D2
1ulz	A	228	238	CPSase_L_D2
2w70	A	230	240	CPSase_L_D2
4h3t	A	364	375	CRISPR_Cse1
3ulj	A	33	53	CSD
3by9	A	102	112	Cache_3
3k6d	A	78	88	Cadherin
3k6i	A	78	88	Cadherin
3pow	A	183	192	Calreticulin
3pow	A	111	130	Calreticulin
3gzk	A	35	45	CelD_N
1ut9	A	242	252	CelD_N
1clc	A	79	89	CelD_N
1eyq	A	35	48	Chalcone
4dok	A	30	43	Chalcone
1en2	A	17	37	Chitin_bind_1
2bem	A	120	145	Chitin_bind_3
1edq	A	78	87	ChitinaseA_N
1tt8	A	79	88	Chor_lyase
3km5	A	1250	1267	Cleaved_Adhesin
3m1h	A	1522	1540	Cleaved_Adhesin
1d2o	A	562	571	Cna_B
2x5p	A	45	55	Cna_B
1txn	A	138	148	Coprogen_oxidas
4b28	A	78	102	Creatinase_N
3hz2	A	7	37	Crystall
4iau	A	88	115	Crystall
3lwk	A	16	46	Crystall
4iau	A	8	31	Crystall
1kbv	A	173	200	Cu-oxidase
2dv6	A	298	325	Cu-oxidase

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
1aoz	A	143	169	Cu-oxidase
4e9x	A	1146	1167	Cu-oxidase
3zx1	A	220	245	Cu-oxidase
3aw5	A	175	200	Cu-oxidase
3gdc	A	182	193	Cu-oxidase_3
1yi9	A	110	139	Cu2_monooxygen
3sxx	A	377	385	Cu_amine_oxid
1oac	A	408	427	Cu_amine_oxid
1ksi	A	325	344	Cu_amine_oxid
3sxx	A	344	363	Cu_amine_oxid
3hi7	A	397	416	Cu_amine_oxid
3hi7	A	219	235	Cu_amine_oxidN3
1qks	A	247	257	Cytochrom_D1
1nir	A	229	239	Cytochrom_D1
1qks	A	269	280	Cytochrom_D1
1nir	A	251	262	Cytochrom_D1
2cfm	A	435	448	DNA_ligase_A_C
1x9m	A	8	17	DNA_pol_A_exo1
2hbj	A	239	253	DNA_pol_A_exo1
3cym	A	43	57	DNA_pol_A_exo1
4fj7	A	115	135	DNA_pol_B_exo1
3qex	A	115	135	DNA_pol_B_exo1
1noy	A	113	132	DNA_pol_B_exo1
3iay	A	322	341	DNA_pol_B_exo1
2xkj	E	1247	1260	DNA_topoisoIV
2xcs	B	1245	1263	DNA_topoisoIV
3npp	A	88	97	DUF1093
2qzb	A	160	186	DUF1131
3zxk	A	446	468	DUF1349
2q0x	A	267	277	DUF1749
3ia8	A	95	112	DUF1794
1qw2	A	73	91	DUF1805
1lmi	A	118	127	DUF1942
1uoy	A	34	47	DUF1962
2hqy	A	210	219	DUF2156
1wna	A	45	63	DUF3197
2hql	A	66	89	DUF3217
2owp	A	104	119	DUF3225
1vkd	A	243	266	DUF377
3taw	A	112	132	DUF377
3kzt	A	149	162	DUF3828
3u7z	A	72	85	DUF4430

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3rob	A	114	128	DUF4440
1yoc	A	125	134	DUF4442
2iml	A	101	118	DUF447
2if6	A	28	49	DUF830
4a9v	A	510	525	DUF839
1iwp	B	106	119	Dehydratase_MU
1ueb	A	112	121	EFP
1xb2	B	230	247	EF_TS
1cl8	A	94	108	EcoRI
1na6	A	122	133	EcoRII-N
3ju4	A	404	415	End_beta_propel
1dy2	A	142	152	Endostatin
1i5p	A	313	337	Endotoxin_mid
2c29	D	227	256	Epimerase
1axi	B	110	119	EpoR_lig-bind
2rcf	A	50	75	EutN_CcmL
4jg3	A	84	97	Exo_endo_phos
1ikp	A	118	128	Exotox-A_bind
2j1a	A	702	716	F5_F8_type_C
4a41	A	1435	1449	F5_F8_type_C
1k3i	A	86	101	F5_F8_type_C
1nkq	A	232	241	FAA_hydrolase
3lzk	A	144	164	FAA_hydrolase
2vfr	A	212	223	FAD-oxidase_C
2x3n	A	133	143	FAD_binding_3
2dkh	A	328	344	FAD_binding_3
2rj2	A	226	236	FBA
3r1m	A	204	215	FBPase_3
2he7	A	321	330	FERM_C
1mix	A	338	349	FERM_C
1h4r	A	242	254	FERM_C
1ef1	A	226	238	FERM_C
4bc3	A	23	33	FGGY_N
3gqs	A	28	50	FHA
2brf	A	32	51	FHA
3kt9	A	26	45	FHA
1sqh	A	213	224	FR47
4b0b	A	146	155	FabA
3tdq	A	72	85	Fimbrial_PilY2
1usc	A	118	140	Flavin_Reduct
2qck	A	117	139	Flavin_Reduct
3fge	A	134	158	Flavin_Reduct

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3d6r	A	134	143	Flu_NS1
3d6r	A	148	158	Flu_NS1
3rvc	A	148	158	Flu_NS1
4iib	A	842	852	Fn3-like
4i3g	A	811	821	Fn3-like
3obi	A	197	217	Formyl_trans_N
2d0o	A	17	31	FtsA
2isb	A	66	78	Fumerase_C
4agi	A	186	206	Fungal_lectin
3tip	A	595	606	G5
4g3v	A	46	59	GAF
3s7o	A	174	188	GAF
4glq	A	461	475	GAF
2ool	A	183	197	GAF
3w2z	A	57	71	GAF
4.00E+04	A	170	184	GAF
3zq5	A	174	188	GAF
3db2	A	203	215	GFO_IDH_MocA_C
3q2i	A	211	223	GFO_IDH_MocA_C
4koa	A	200	213	GFO_IDH_MocA_C
2glx	A	200	213	GFO_IDH_MocA_C
3dty	A	230	243	GFO_IDH_MocA_C
3ip3	A	203	217	GFO_IDH_MocA_C
2p2s	A	203	217	GFO_IDH_MocA_C
1h6d	A	289	304	GFO_IDH_MocA_C
4hkt	A	194	210	GFO_IDH_MocA_C
3ezy	A	194	210	GFO_IDH_MocA_C
2wsh	A	38	47	GIY-YIG
4jv8	B	79	88	GMP_PDE_delta
4jv8	B	21	32	GMP_PDE_delta
3g3s	A	33	41	GNAT_acetyltran
3g3s	A	163	172	GNAT_acetyltran
2ism	A	233	262	GSDH
3das	A	108	131	GSDH
1cru	A	115	135	GSDH
2ism	A	289	311	GSDH
2g8s	A	96	116	GSDH
2ism	A	106	124	GSDH
2c78	A	387	396	GTP_EFTU_D3
1d2e	A	423	432	GTP_EFTU_D3
1jny	A	363	374	GTP_EFTU_D3
2v36	B	470	479	G_glu_transpept

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3nv1	A	225	234	Gal-bind_lectin
2r0h	A	49	59	Gal-bind_lectin
2wkk	A	40	50	Gal-bind_lectin
4gxl	A	259	270	Gal-bind_lectin
3mbr	X	109	120	Glu_cyclase_2
2iwa	A	69	80	Glu_cyclase_2
2vgd	A	76	95	Glyco_hydro_11
3vgi	A	174	203	Glyco_hydro_12
1ks5	A	95	122	Glyco_hydro_12
3amn	A	112	140	Glyco_hydro_12
2nlr	A	100	126	Glyco_hydro_12
3vl9	A	107	118	Glyco_hydro_12
2nlr	A	106	119	Glyco_hydro_12
1olr	A	105	119	Glyco_hydro_12
1ks5	A	101	115	Glyco_hydro_12
2hyk	A	181	190	Glyco_hydro_16
3azy	A	191	200	Glyco_hydro_16
3i4i	A	159	168	Glyco_hydro_16
1o4y	A	216	225	Glyco_hydro_16
2uwa	A	154	163	Glyco_hydro_16
3dgt	A	207	216	Glyco_hydro_16
4asm	B	229	238	Glyco_hydro_16
4atf	A	258	267	Glyco_hydro_16
3rq0	A	201	214	Glyco_hydro_16
3cmg	A	282	291	Glyco_hydro_2
2je8	A	313	322	Glyco_hydro_2
3hn3	A	261	271	Glyco_hydro_2
4amw	A	808	817	Glyco_hydro_31
3lig	A	337	358	Glyco_hydro_32N
3lig	A	245	269	Glyco_hydro_32N
3kf3	A	81	100	Glyco_hydro_32N
1w2t	A	45	64	Glyco_hydro_32N
1w2t	A	162	181	Glyco_hydro_32N
3bvx	A	686	696	Glyco_hydro_38C
3bvx	A	1010	1021	Glyco_hydro_38C
3bmx	A	616	628	Glyco_hydro_3_C
3c7f	A	126	148	Glyco_hydro_43
3c2u	A	211	233	Glyco_hydro_43
3k1u	A	38	59	Glyco_hydro_43
2exh	A	212	234	Glyco_hydro_43
3p2n	A	200	221	Glyco_hydro_43
2x8s	A	127	147	Glyco_hydro_43

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3akh	A	42	63	Glyco_hydro_43
3qee	A	130	152	Glyco_hydro_43
1uv4	A	214	244	Glyco_hydro_43
3c7f	A	256	276	Glyco_hydro_43
1yif	A	211	233	Glyco_hydro_43
3nqh	A	184	195	Glyco_hydro_43
1uv4	A	201	214	Glyco_hydro_43
3nqh	A	71	90	Glyco_hydro_43
3fef	A	203	216	Glyco_hydro_4C
4b5q	A	81	106	Glyco_hydro_61
3vmn	A	343	352	Glyco_hydro_66
2yfr	A	491	514	Glyco_hydro_68
1oyg	A	191	218	Glyco_hydro_68
3vss	A	343	364	Glyco_hydro_68
2yfr	A	495	508	Glyco_hydro_68
3lm4	A	181	190	Glyoxalase
2zyq	A	192	202	Glyoxalase
2wl9	A	42	52	Glyoxalase_2
2ehz	A	45	55	Glyoxalase_2
3lm4	A	45	57	Glyoxalase_2
3k1t	A	47	60	GshA
4hxw	A	148	167	HATPase_c
3lnu	A	179	198	HATPase_c
1td4	A	39	62	HDPD
3k7i	B	132	150	HH_signal
1v3e	A	508	517	HN
2vsm	A	251	269	HN
1v3e	A	225	243	HN
2p08	A	21	31	HNOBA
4f01	A	398	412	HSP70
1u00	A	395	409	HSP70
2op6	A	425	439	HSP70
3dqg	A	423	437	HSP70
2ykf	A	41	55	H_kinase_N
1dk0	A	45	54	HasA
3oyo	A	84	99	Hemopexin
1eyb	A	167	187	HgmA
3tpd	A	159	178	HipA_N
2z1c	A	42	51	HupF_HypC
3esg	A	29	40	HutD
2b0t	A	432	442	IDH
2idr	A	195	213	IF4E

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3mfi	A	52	61	IMS
4ecq	A	35	49	IMS
1yoe	A	258	288	IU_nuc_hydro
2o99	A	132	142	IclR
1iam	A	122	131	Ig_2
3bn3	B	125	134	Ig_2
3ry4	A	122	131	Ig_2
1rhf	A	132	141	Ig_2
1jiw	I	69	79	Inh
3m86	A	101	114	Inhibitor_I42
2qfl	A	104	113	Inositol_P
2p3n	A	99	108	Inositol_P
3lv0	A	106	115	Inositol_P
1ka1	A	161	170	Inositol_P
2q7d	A	280	295	Ins134_P3_kin
2p26	A	65	80	Integrin_beta
1xs0	A	57	65	Ivy
3apa	A	99	107	Jacalin
3cz7	A	54	63	KAT11
3biy	A	1361	1374	KAT11
4fdw	A	128	137	LRR_5
1ljo	A	48	57	LSM
1mgq	A	55	64	LSM
4fp5	D	80	93	LT-IIB
3scy	A	259	272	Lactonase
3hfq	A	111	125	Lactonase
1ri6	A	203	218	Lactonase
2erf	A	88	99	Laminin_G_2
1t2d	A	278	287	Ldh_1_C
1a5z	A	291	301	Ldh_1_C
2i6t	A	431	440	Ldh_1_C
1pzg	A	291	301	Ldh_1_C
4bgt	A	269	278	Ldh_1_C
4ajj	A	292	301	Ldh_1_C
1mld	A	268	277	Ldh_1_C
1guz	A	268	277	Ldh_1_C
3vpg	A	291	300	Ldh_1_C
1z2i	A	209	225	Ldh_2
1ijq	A	476	491	Ldl_recept_b
3sov	A	118	133	Ldl_recept_b
4a0p	A	1082	1097	Ldl_recept_b
3sov	A	162	177	Ldl_recept_b

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
4a0p	A	728	743	Ldl_recept_b
4a0p	A	814	829	Ldl_recept_b
4a0p	A	1037	1052	Ldl_recept_b
4a0p	A	685	700	Ldl_recept_b
3sov	A	74	91	Ldl_recept_b
3zyr	A	157	167	Lectin_legB
1hql	A	146	156	Lectin_legB
1avb	A	130	140	Lectin_legB
2fmd	A	145	155	Lectin_legB
1fx5	A	145	155	Lectin_legB
1nls	A	26	36	Lectin_legB
1led	A	147	157	Lectin_legB
1gzc	A	144	154	Lectin_legB
2eig	A	137	147	Lectin_legB
2fsq	A	220	234	LigT_PEase
3o22	A	114	137	Lipocalin
2ypv	A	92	113	Lipoprot_C
2i5v	O	224	233	Lipoprotein_1
2i5v	O	124	134	Lipoprotein_1
1xs5	A	130	147	Lipoprotein_9
3fg1	A	699	713	Lipoxygenase
1iwm	A	128	138	LolB
3fka	A	100	113	Lumazine_bd_2
2hje	A	116	125	LuxQ-periplasm
2nyk	A	38	48	M157
2qyv	A	253	263	M20_dimer
4gwm	A	356	375	MAM
3gmo	A	117	126	MHC_I
1lqv	A	30	39	MHC_I
3fru	A	111	120	MHC_I
1u58	A	31	40	MHC_I
3jvg	A	111	120	MHC_I
4g43	A	25	34	MHC_I
1k5n	A	25	34	MHC_I
4l4v	A	111	120	MHC_I
1t7v	A	30	39	MHC_I
1nez	A	25	34	MHC_I
2pq8	A	302	316	MOZ_SAS
3k67	A	141	150	MaoC_dehydratas
4gr5	A	13	39	MbtH
2oiz	A	361	372	Me-amine-dh_H
2p0w	A	222	240	Mec-17

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
2qfp	A	403	416	Metallophos_C
1eu1	A	19	28	Molybdop_Fe4S4
1kqf	A	61	77	Molybdop_Fe4S4
1jcf	A	162	171	MreB_Mbl
2x5o	A	247	256	Mur_ligase_M
2odi	A	159	178	MvaI_BcnI
3pms	A	213	222	N-glycanase_C
3pms	A	64	79	N-glycanase_N
4art	A	60	89	NA
1vlm	A	175	199	NA
2zxr	A	505	526	NA
3zwl	B	123	145	NA
2z5w	A	89	109	NA
4b70	A	76	104	NA
4b72	A	76	104	NA
3f8t	A	128	136	NA
1lqt	A	278	286	NA
2gwn	A	21	30	NA
4asm	B	42	51	NA
1kve	B	192	201	NA
1p7t	A	219	228	NA
2xfg	B	600	609	NA
3ze9	B	35	44	NA
3fo8	D	117	126	NA
1e3d	B	31	40	NA
1h72	C	16	25	NA
1g87	A	601	610	NA
4h3o	A	32	41	NA
1cc1	L	30	39	NA
2wm1	A	38	47	NA
3hpa	A	54	63	NA
3myr	B	1021	1030	NA
3nzn	A	2	12	NA
2cy2	A	62	72	NA
1nv8	A	264	274	NA
1sqj	A	249	259	NA
2z8x	A	125	135	NA
4ii9	A	45	55	NA
3gkr	A	45	55	NA
4aio	A	100	110	NA
2qub	A	125	135	NA
3ayx	A	35	45	NA

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3uqy	L	35	45	NA
2w18	A	992	1002	NA
3fed	A	119	129	NA
4jgp	A	107	117	NA
2wfw	A	203	214	NA
2x5r	A	95	106	NA
2wlg	A	179	190	NA
3fot	A	449	460	NA
3jqy	A	204	215	NA
3sbq	A	156	167	NA
2cvb	A	136	147	NA
3nok	A	66	77	NA
3obf	A	150	163	NA
2w18	A	1138	1149	NA
2xgr	A	185	196	NA
1wzn	A	230	242	NA
3bjn	A	106	119	NA
2jbv	A	336	348	NA
3it5	A	22	34	NA
4eqa	C	144	157	NA
4f8l	A	111	124	NA
4b9g	A	137	150	NA
2ia2	A	112	125	NA
2py5	A	13	26	NA
3ts3	A	509	522	NA
4ha6	A	335	349	NA
4gt6	A	40	54	NA
3f6k	A	205	220	NA
1dp4	A	377	392	NA
3thr	A	268	283	NA
3qne	A	353	368	NA
1ijq	A	387	402	NA
4h2w	A	237	252	NA
3f6k	A	252	267	NA
1wle	A	388	403	NA
3lss	A	371	386	NA
3zn4	A	76	91	NA
3sov	A	30	47	NA
1r85	A	332	349	NA
4fh3	A	130	147	NA
3hjr	A	378	396	NA
3kya	A	424	438	NHL

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3hrp	A	174	189	NHL
3hrp	A	367	382	NHL
4g38	A	502	519	NIR_SIR
3b0g	A	512	530	NIR_SIR
3npf	A	248	263	NLPC_P60
2wuu	A	113	124	NMT
1iic	A	136	147	NMT
2wuu	A	312	321	NMT_C
3iu1	A	386	396	NMT_C
2vng	A	142	151	NPCBM
4k1v	A	105	115	NTF2
4k1u	A	105	115	NTF2
1jkg	A	113	130	NTF2
2xme	A	150	170	NTP_transferase
1yp2	A	184	194	NTP_transferase
2fvv	A	73	102	NUDIX
3ees	A	79	106	NUDIX
3bho	A	149	170	NUDIX_2
3hx6	A	849	860	Neisseria_PilC
4gdi	A	204	213	Neur
1yqw	Q	293	311	NiFeSe_Hases
1o13	A	6	26	Nitro_FeMo-Co
1x8q	A	87	97	Nitrophorin
1x7d	A	90	101	OCD_Mu_crystall
2x3l	A	125	136	OKR_DC_1
2x3l	A	398	409	OKR_DC_1_C
2x55	A	142	166	Omptin
3szv	A	129	151	OprD
4frx	A	137	159	OprD
3t0s	A	130	152	OprD
2y2x	A	132	154	OprD
2e8e	A	70	82	OsmC
2d7v	A	101	115	OsmC
4b2z	A	269	279	Oxysterol_BP
1zhx	A	249	265	Oxysterol_BP
3blj	A	629	637	PARP
3hkv	A	979	987	PARP
2pqf	A	652	660	PARP
3smj	A	1693	1701	PARP
2x5y	A	868	876	PARP
4ew7	A	35	45	PAS
2qkp	A	344	354	PAS_10

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3s7o	A	42	53	PAS_2
2r78	A	28	38	PAS_9
3olo	A	24	34	PAS_9
2gj3	A	39	49	PAS_9
3d72	A	74	87	PAS_9
4eet	B	392	405	PAS_9
2z6d	A	136	149	PAS_9
4hqa	A	30	43	PAS_9
1n9l	A	23	36	PAS_9
4hp4	A	30	43	PAS_9
3t50	A	35	48	PAS_9
4hoi	A	31	44	PAS_9
2wad	A	285	295	PBP_dimer
3chm	A	141	156	PCI
3sfj	A	32	60	PDZ
2h3l	A	1334	1358	PDZ
2fe5	A	238	259	PDZ
3qe1	A	53	81	PDZ
3egg	C	508	530	PDZ
2jik	A	25	46	PDZ
4h11	A	110	131	PDZ
1r6j	A	211	220	PDZ
3soe	A	615	628	PDZ
1kwa	A	503	517	PDZ
2qg1	A	1738	1753	PDZ
3o46	A	151	166	PDZ
1qau	A	30	45	PDZ
2uzc	A	16	32	PDZ
3nfk	A	530	546	PDZ
2pkt	A	16	32	PDZ
3hpk	A	35	51	PDZ
1qav	A	94	110	PDZ
2i04	A	464	480	PDZ
2q3g	A	16	32	PDZ
2v1w	A	15	31	PDZ
2pa1	A	15	31	PDZ
2v90	A	261	279	PDZ
3k50	A	98	118	PDZ_2
2wzb	A	285	314	PGK
16pk	A	288	316	PGK
1php	A	266	294	PGK
3v5w	A	589	599	PH

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3no8	A	423	434	PHR
1e8y	A	386	395	PI3K_C2
3d8d	A	432	443	PID
2ej8	A	69	80	PID
2wfp	A	8	35	PMI_typeI
1qwr	A	12	37	PMI_typeI
1jb7	A	147	173	POT1
3jqw	A	926	936	PPC
4guc	A	91	109	PPC
2nnu	A	111	136	PPV_E2_N
1tue	B	115	140	PPV_E2_N
2ad6	A	488	499	PQQ
3vgz	A	297	308	PQQ_2
3vgz	A	72	83	PQQ_2
2zuy	A	492	507	PQQ_2
2wjn	H	171	182	PRC
1rzh	H	167	178	PRC
3n0a	A	253	262	PTEN_C2
2xm5	A	110	120	PTase_Orf2
3d79	A	136	160	PUA
3r90	A	144	168	PUA
1q7h	A	118	142	PUA
4dmg	A	51	67	PUA
1wxx	A	45	61	PUA
2qjf	A	327	336	PUA_2
2f9w	A	15	25	Pan_kinase
3qu1	A	81	106	Pep_deformylase
3svj	P	117	146	Pep_deformylase
2os0	A	101	130	Pep_deformylase
2okl	A	97	126	Pep_deformylase
1rl4	A	145	171	Pep_deformylase
1xeo	A	80	105	Pep_deformylase
1lqy	A	97	126	Pep_deformylase
1m6d	A	133	159	Peptidase_C1
1iwd	A	132	158	Peptidase_C1
2bdz	A	133	159	Peptidase_C1
1cqd	A	135	161	Peptidase_C1
1s4v	A	136	162	Peptidase_C1
2fo5	A	141	167	Peptidase_C1
2wbf	X	735	762	Peptidase_C1
3qj3	A	240	267	Peptidase_C1
3ovx	A	83	108	Peptidase_C1

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3f75	A	141	167	Peptidase_C1
2p86	A	138	162	Peptidase_C1
3ioq	A	133	159	Peptidase_C1
2pns	A	131	157	Peptidase_C1
3kwz	A	134	162	Peptidase_C1
1ppo	A	133	159	Peptidase_C1
2cio	A	133	159	Peptidase_C1
1yal	A	133	159	Peptidase_C1
3i06	A	138	162	Peptidase_C1
3gq8	A	661	669	Peptidase_G2
1k7i	A	322	332	Peptidase_M10_C
3u1r	A	324	334	Peptidase_M10_C
1kap	P	310	320	Peptidase_M10_C
3u1r	A	454	465	Peptidase_M10_C
3rva	A	218	240	Peptidase_M24
3nqx	A	390	403	Peptidase_M4_C
3v39	A	334	354	Peptidase_S13
1k32	A	992	1011	Peptidase_S41
4hvt	A	96	116	Peptidase_S9_N
2xe4	A	218	228	Peptidase_S9_N
2xdw	A	362	373	Peptidase_S9_N
2bkl	A	95	110	Peptidase_S9_N
4hvt	A	201	217	Peptidase_S9_N
1yr2	A	244	261	Peptidase_S9_N
2xe4	A	228	245	Peptidase_S9_N
2xdw	A	406	423	Peptidase_S9_N
4hvt	A	217	234	Peptidase_S9_N
1yr2	A	134	152	Peptidase_S9_N
1pea	A	341	355	Peripla_BP_5
3n0w	A	358	379	Peripla_BP_6
4f06	A	361	378	Peripla_BP_6
4eyg	A	362	379	Peripla_BP_6
4evq	A	361	379	Peripla_BP_6
3w1e	A	183	192	Pfam-B_10290
3lwt	X	369	379	Pfam-B_11583
4ktb	A	45	55	Pfam-B_11859
1jz8	A	699	716	Pfam-B_12060
1pl3	A	144	165	Pfam-B_12144
2o14	A	102	111	Pfam-B_1453
4h5u	A	257	267	Pfam-B_15998
4hgd	A	257	267	Pfam-B_15998
4hg5	A	257	267	Pfam-B_15998

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
4i0o	A	107	118	Pfam-B_164
3un7	A	54	63	Pfam-B_17421
1oxx	K	212	221	Pfam-B_1850
4kl8	L	30	39	Pfam-B_19006
1wui	L	41	50	Pfam-B_19006
1yqw	Q	32	41	Pfam-B_19006
4kn9	L	30	39	Pfam-B_19006
4ko2	L	30	39	Pfam-B_19006
3gre	A	1303	1314	Pfam-B_2035
2zxq	A	484	514	Pfam-B_3371
2w18	A	935	943	Pfam-B_3828
3c9i	A	1160	1169	Pfam-B_475
2wfw	A	123	133	Pfam-B_475
3sbq	A	472	485	Pfam-B_475
1fwx	A	415	428	Pfam-B_475
1fwx	A	185	205	Pfam-B_5
1fwx	A	132	140	Pfam-B_5
1fwx	A	106	117	Pfam-B_5
1fwx	A	83	94	Pfam-B_5
3dwv	A	142	152	Pfam-B_512
2v1m	A	137	147	Pfam-B_512
3no2	A	36	47	Pfam-B_518
4ind	A	315	331	Pfam-B_518
3gre	A	1368	1384	Pfam-B_518
3sbq	A	237	257	Pfam-B_5281
3sbq	A	182	190	Pfam-B_5281
3s25	A	117	132	Pfam-B_5657
3e4g	A	37	64	Pfam-B_655
3w9a	A	156	165	Pfam-B_656
1fwx	A	379	391	Pfam-B_7
3sbq	A	436	448	Pfam-B_7
3h43	A	135	146	Pfam-B_7134
3te8	A	90	119	Pfam-B_7838
2.00E+11	A	218	228	Pfam-B_827
1uf5	A	249	259	Pfam-B_827
3hwp	A	198	210	Pfam-B_8359
3f6k	A	342	357	Pfam-B_89
2pkf	A	15	41	PfkB
2abq	A	15	30	PfkB
2v5i	A	639	649	PhageP22-tail
2vfo	A	594	604	PhageP22-tail
3d37	A	60	69	Phage_GPD

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3hxl	A	134	143	Phage_sheath_1
1ub0	A	40	49	Phos_pyr_kin
1l5w	A	180	195	Phosphorylase
3amr	A	156	174	Phytase
3lxl	A	889	899	Pkinase_Tyr
1vl4	A	33	42	PmbA_TldD
2vxq	A	63	82	Pollen_allerg_1
2zzj	A	91	110	Polysacc_lyase
1stm	A	75	86	Potex_coat
4k8l	A	129	138	Pro_racemase
1w61	A	46	63	Pro_racemase
3nec	A	22	33	Profilin
2jkg	A	25	36	Profilin
1ryp	2	19	28	Proteasome
3mfb	A	396	407	Pyocin_S
3ef6	A	211	219	Pyr_redox
1nhs	A	217	225	Pyr_redox
1q1r	A	217	228	Pyr_redox
2q9k	A	84	111	Pyridox_oxidase
3of7	A	373	383	RCC1
2ewf	A	438	448	RE_AlwI
3dd6	A	165	186	RNase_PH_C
1r6l	A	166	186	RNase_PH_C
2igi	A	12	26	RNase_T
3v9w	A	24	38	RNase_T
1wlj	A	12	27	RNase_T
2gui	A	13	32	RNase_T
2qm1	A	18	28	ROK
4htl	A	16	26	ROK
1xiy	A	140	149	Redoxin
4f82	A	136	145	Redoxin
2xhf	A	162	171	Redoxin
1kng	A	160	170	Redoxin
2b1k	A	147	157	Redoxin
1fn9	A	25	42	Reovirus_cap
2b3g	A	75	103	Rep-A_N
3i3f	A	17	32	Ribonuc_L-PSP
1jy5	A	168	177	Ribonuclease_T2
2pqx	A	195	206	Ribonuclease_T2
2gbw	A	105	115	Rieske
1uli	A	123	133	Rieske
2gbw	B	142	155	Ring_hydroxyl_B

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
1wql	B	152	167	Ring_hydroxyl_B
1gqv	A	109	129	RnaseA
1m07	A	94	103	RnaseA
3snf	A	88	97	RnaseA
2zpo	A	105	114	RnaseA
1oj8	A	88	97	RnaseA
2vq9	A	111	120	RnaseA
2p7s	A	98	107	RnaseA
1agi	A	106	115	RnaseA
2vq8	A	110	119	RnaseA
4aoh	A	105	114	RnaseA
3tsr	A	109	120	RnaseA
1rnf	A	105	116	RnaseA
1dy5	A	108	119	RnaseA
2e0j	A	108	119	RnaseA
4a2o	A	109	128	RnaseA
1gk8	I	108	119	RuBisCO_small
3s82	A	186	195	S-AdoMet_synt_M
2cw5	A	82	107	SAM_adeno_trans
2q6k	A	84	111	SAM_adeno_trans
2ece	A	250	273	SBP56
2ece	A	46	62	SBP56
2ece	A	281	298	SBP56
3g4e	A	174	194	SGL
2dg1	A	205	227	SGL
2ghs	A	242	252	SGL
3o4p	A	179	189	SGL
1pby	B	118	129	SGL
2ghs	A	150	163	SGL
2p4o	A	233	247	SGL
3o4p	A	47	62	SGL
3e5z	A	91	106	SGL
2p4o	A	189	208	SGL
2jk9	A	187	201	SPRY
1woc	A	74	96	SSB
2sic	I	15	29	SSI
3fss	A	105	114	SSrecog
2prv	A	111	126	SUKH_6
3hie	A	154	168	Sec3-PIP2_bind
2q0z	X	260	271	Sec63
2j3w	A	8	18	Sedlin_N
3cxk	A	52	63	SelR

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3hcg	A	419	430	SelR
3hcj	A	50	61	SelR
3mao	A	20	31	SelR
3al9	A	71	81	Sema
3ozq	A	372	381	Serpin
2arr	A	397	408	Serpin
1wz9	A	357	368	Serpin
2pef	A	404	415	Serpin
1sjw	A	109	118	SnoaL
1oh0	A	105	115	SnoaL_2
3ff2	A	31	41	SnoaL_2
1f1g	A	118	145	Sod_Cu
1xtm	A	164	185	Sod_Cu
1ej8	A	132	151	Sod_Cu
1p7g	A	143	173	Sod_Fe_C
4f2n	A	158	188	Sod_Fe_C
1coj	A	130	160	Sod_Fe_C
2rcv	A	133	161	Sod_Fe_C
1ids	A	127	157	Sod_Fe_C
1b06	A	137	167	Sod_Fe_C
4f6e	A	135	165	Sod_Fe_C
3dc5	A	124	152	Sod_Fe_C
4ffk	A	145	175	Sod_Fe_C
1mng	A	134	163	Sod_Fe_C
3h1s	A	124	153	Sod_Fe_C
1dt0	A	124	153	Sod_Fe_C
3g66	A	221	229	Sortase
4g1h	A	218	226	Sortase
3o0p	A	219	227	Sortase
4g1j	A	225	233	Sortase
1o6a	A	124	141	SpoA
1ryq	A	30	59	Spt4
3lpe	B	27	56	Spt4
4ia6	A	252	265	Strep_67kDa_ant
2okg	A	273	283	Sugar-bind
2qzu	A	425	434	Sulfatase_C
2p0a	A	351	365	Synapsin_C
3nje	A	113	136	T2SJ
1t6e	X	207	232	TAXi_C
4acj	A	752	762	TLD
1aol	A	16	26	TLV_coat
1uwv	A	50	68	TRAM

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
2ahn	A	54	69	Thaumatin
2yj7	A	78	87	Thioredoxin
2i1u	A	84	93	Thioredoxin
2voc	A	76	85	Thioredoxin
1nw2	A	76	85	Thioredoxin
2yzu	A	78	87	Thioredoxin
1v98	A	109	118	Thioredoxin
3die	A	76	85	Thioredoxin
2hls	A	199	208	Thioredoxin_3
2qgu	A	152	165	Tol_Tol_Ttg2
1xkw	A	630	638	TonB_dep_Rec
2h5f	A	36	50	Toxin_1
1epw	A	1014	1023	Toxin_R_bind_N
2vxr	A	1023	1032	Toxin_R_bind_N
3pme	A	1020	1029	Toxin_R_bind_N
3pbt	A	510	518	Transpeptidase
3un7	A	446	454	Transpeptidase
2wad	A	637	645	Transpeptidase
1vqq	A	543	559	Transpeptidase
3qva	A	23	32	Transthyretin
1fxy	A	32	42	Trypsin
1elv	A	620	636	Trypsin
1jke	A	12	21	Tyr_Deacylase
2okv	A	12	21	Tyr_Deacylase
1j7g	A	12	21	Tyr_Deacylase
2pkh	A	144	153	UTRA
2fa1	A	135	144	UTRA
2ikk	A	133	142	UTRA
1wdj	A	88	113	Uma2
3ijm	A	68	79	Uma2
1uwk	A	490	499	Urocanase
1hkf	A	92	103	V-set
2pnd	A	36	49	V-set
2jju	A	37	50	V-set
2yz1	A	96	110	V-set
1smo	A	54	68	V-set
3r0n	A	69	83	V-set
2qhl	A	33	47	V-set
2q87	A	85	101	V-set
3t3p	A	390	405	VCBS
4drr	A	175	188	VP4_haemagglut
1lsh	A	50	68	Vitellogenin_N

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
2r51	A	208	229	Vps26
3odt	A	163	172	WD40
3odt	A	122	131	WD40
4gqb	B	150	161	WD40
3zwl	B	221	232	WD40
4ery	A	112	123	WD40
4ery	A	154	165	WD40
4j73	A	253	264	WD40
1nr0	A	264	275	WD40
3frx	A	90	101	WD40
3vl1	A	161	172	WD40
4j0w	A	348	359	WD40
3vl1	A	203	214	WD40
1got	B	209	220	WD40
3zwl	B	77	88	WD40
1got	B	167	178	WD40
2pbi	B	264	275	WD40
4ery	A	239	250	WD40
4lg8	A	417	428	WD40
2pbi	B	222	233	WD40
4j87	A	120	131	WD40
3w15	A	86	98	WD40
3vl1	A	313	325	WD40
4i79	A	243	255	WD40
3i2n	A	142	154	WD40
4ggc	A	291	303	WD40
3fm0	A	175	188	WD40
2hes	X	178	191	WD40
3fm0	A	41	54	WD40
3fm0	A	130	144	WD40
2xyi	A	257	271	WD40
1gxr	A	508	522	WD40
3odt	A	35	51	WD40
3gre	A	1106	1123	WD40
2hes	X	224	243	WD40
2ygo	A	155	165	WIF
3hbz	A	263	287	Xylanase
2oc3	A	91	101	Y_phosphatase
1t82	A	124	133	YiiD_Cterm
3lmb	A	135	153	YiiD_Cterm
1t71	A	248	259	YmdB
3dn7	A	57	71	cNMP_binding

Continued on next page

Table A.1 – *Continued from previous page*

Accesion Code	ChainID	N anchor	C anchor	Pfam
3zpg	A	34	43	dCMP_cyt_deam_1
2hxv	A	29	38	dCMP_cyt_deam_1
2a8n	A	29	38	dCMP_cyt_deam_1
1wkq	A	29	38	dCMP_cyt_deam_1
2nx8	A	41	50	dCMP_cyt_deam_1
2yv5	A	50	64	eIF-1a
3i4o	A	54	70	eIF-1a
1khi	A	154	167	eIF-5a
2eif	A	117	130	eIF-5a
1x6o	A	146	159	eIF-5a
4af1	A	159	172	eRF1_2
3lpw	A	6	31	fn3
3f7q	A	1220	1245	fn3
2w1n	A	947	956	fn3
3b4n	A	48	64	fn3
1nkg	A	307	316	fn3_3
2wm5	A	409	422	p450
1q5d	A	395	408	p450
3r9b	A	396	409	p450
3ut2	A	721	732	peroxidase
3vlj	A	669	680	peroxidase
3u1f	A	487	497	tRNA-synt_1g
3nem	A	354	363	tRNA-synt_2
1e1o	A	414	423	tRNA-synt_2
3a74	A	404	413	tRNA-synt_2
2xgt	A	463	473	tRNA-synt_2
1nnh	A	206	217	tRNA-synt_2
1x54	A	348	359	tRNA-synt_2
1g5h	A	262	275	tRNA-synt_2b
3hy0	A	183	211	tRNA-synt_2c
2ztg	A	202	211	tRNA-synt_2c
1yfs	A	174	193	tRNA-synt_2c
3teg	A	240	253	tRNA-synt_2d
3nem	A	74	94	tRNA_anti-codon
3a74	A	114	132	tRNA_anti-codon
1r0v	A	244	260	tRNA_int_endo
2qkd	A	256	268	zf-ZPR1

APPENDIX B

PFAMS THAT OCCUR MORE THAN ONCE IN THE SET OF LAT+KINK MATCHES

PFam Architecture	Matches	Aligns	GO tags
WD40	34	4	protein binding
PDZ	23	1	protein binding
Acetyltransf_1	21	5	N-acetyltransferase activity
Peptidase_C1	18	2	cysteine-type peptidase activity; proteolysis
RnaseA	15	1	NA
Glyco_hydro_43	14	9	hydrolase activity, hydrolyzing O-glycosyl compounds; carbohydrate metabolic process
Sod_Fe_C	12	1	superoxide dismutase activity; metal ion binding; superoxide metabolic process; oxidation-reduction process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
PAS_9	11	2	signal transducer activity; signal transduction
Asp	11	4	aspartic-type endopeptidase activity; proteolysis
MHC_I	10	3	immune response; antigen processing and presentation
Peptidase_S9_N	10	5	serine-type endopeptidase activity; serine-type exopeptidase activity
B_lectin	10	3	NA
GFO_IDH_MocA_C	10	5	oxidoreductase activity; metabolic process; oxidation-reduction process
AMP-binding	10	2	catalytic activity; metabolic process
SGL	10	9	NA
Glyco_hydro_16	9	1	hydrolase activity, hydrolyzing O-glycosyl compounds; carbohydrate metabolic process
Lectin_legB	9	1	carbohydrate binding
Ldl_recept_b	9	2	NA
Glyco_hydro_12	8	4	cellulase activity; polysaccharide catabolic process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Ldh_1_C	8	1	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor; oxidation-reduction process
Aldedh	8	1	oxidoreductase activity; metabolic process; oxidation-reduction process
V-set	8	4	NA
Pep_deformylase	7	3	NA
GAF	7	4	protein binding
Thioredoxin	7	1	cell redox homeostasis
tRNA-synt_2	6	1	nucleotide binding; aminoacyl-tRNA ligase activity; ATP binding; tRNA aminoacylation for protein translation
AhpC-TSA	6	2	antioxidant activity; oxidoreductase activity; oxidation-reduction process
BNR_2	6	2	NA
GSDH	6	3	oxidoreductase activity, acting on the CH-OH group of donors, quinone or similar compound as acceptor; quinone binding; carbohydrate metabolic process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Cu-oxidase	6	1	oxidoreductase activity; oxidation-reduction process
C1q	6	2	NA
BPL_LplA_LipB	5	1	cellular protein modification process
4HBT	5	4	NA
Cu_amine_oxid	5	2	copper ion binding; primary amine oxidase activity; quinone binding; amine metabolic process; oxidation-reduction process
PUA	5	2	RNA binding
Glyco_hydro_32N	5	3	NA
dCMP_cyt_deam_1	5	1	zinc ion binding; hydrolase activity
Redoxin	5	2	oxidoreductase activity
PARP	5	2	NAD+ ADP-ribosyltransferase activity
Peripla_BP_6	4	3	NA
Inositol_P	4	1	phosphatidylinositol phosphorylation
SelR	4	1	peptide-methionine (R)-S-oxide reductase activity; oxidation-reduction process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Glyco_hydro_68	4	3	levansucrase activity; carbohydrate utilization
Transpeptidase	4	2	penicillin binding
HSP70	4	1	NA
Sortase	4	1	NA
Gal-bind_lectin	4	1	carbohydrate binding
Ig_2	4	1	NA
CN_hydrolase	4	1	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds; nitrogen compound metabolic process
FERM_C	4	2	NA
fn3	4	3	protein binding
Serpin	4	1	NA
Pfam-B_475	4	4	NA
Peptidase_M10_C	4	2	calcium ion binding; extracellular space
RNase_T	4	2	NA
Crystall	4	2	NA
DNA_pol_B_exo1	4	1	DNA-directed DNA polymerase activity
Pfam-B_5	4	4	NA

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Cytochrom_D1	4	2	NA
OprD	4	1	porin activity; transport; integral component of membrane
Flavin_Reduct	3	1	FMN binding; oxidoreductase activity; riboflavin reductase (NADPH) activity; oxidation-reduction process
Glyco_hydro_2	3	2	hydrolase activity, hydrolyzing O-glycosyl compounds; carbohydrate metabolic process
4HBT_2	3	2	NA
PQQ_2	3	3	NA
Pfam-B_518	3	3	NA
FHA	3	1	protein binding
Tyr_Deacylase	3	1	hydrolase activity, acting on ester bonds; D-amino acid catabolic process; cytoplasm
Pfam-B_15998	3	1	NA
ATP-grasp	3	1	NA
Pyr_redox	3	1	oxidoreductase activity; flavin adenine dinucleotide binding; oxidation-reduction process
Glyoxalase_2	3	2	NA

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
NHL	3	2	protein binding
PGK	3	1	phosphoglycerate kinase activity; glycolytic process
Lactonase	3	2	NA
eIF-5a	3	1	RNA binding; translation elongation factor activity; ribosome binding; translational frameshifting; positive regulation of translational elongation; positive regulation of translational termination
Aldose_epim	3	2	isomerase activity; carbohydrate metabolic process
ATP-grasp_2	3	2	NA
NTF2	3	2	transport; intracellular
p450	3	3	iron ion binding; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; heme binding; oxidation-reduction process
CelD_N	3	1	cellulase activity; carbohydrate metabolic process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Toxin_R_bind_N	3	1	metalloendopeptidase activity; toxin receptor binding; pathogenesis; inhibition of neurotransmitter uptake; extracellular region
Pfam-B_19006	3	1	NA
Sod_Cu	3	2	metal ion binding; superoxide metabolic process; oxidation-reduction process
UTRA	3	1	DNA binding; regulation of transcription, DNA-templated
SBP56	3	3	selenium binding
F5_F8_type_C	3	2	cell adhesion
DNA_pol_A_exo1	3	2	nucleic acid binding; 3'-5' exonuclease activity; nucleobase-containing compound metabolic process
HN	3	2	exo-alpha-sialidase activity; host cell surface receptor binding; viral life cycle; viral envelope
tRNA-synt_2c	3	2	nucleotide binding; alanine-tRNA ligase activity; ATP binding; alanyl-tRNA aminoacylation
Flu_NS1	3	2	RNA binding
GTP_EFTU_D3	3	2	GTP binding

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
CPSase_L_D2	3	1	ATP binding
ArgJ	2	1	glutamate N-acetyltransferase activity; arginine biosynthetic process
Glu_cyclase_2	2	2	NA
NTP_transferase	2	2	nucleotidyltransferase activity; biosynthetic process
Cadherin	2	1	calcium ion binding; homophilic cell adhesion; membrane
LSM	2	1	NA
eIF-1a	2	2	RNA binding; translation initiation factor activity; translational initiation
Chalcone	2	1	intramolecular lyase activity
SAM_adeno_trans	2	1	NA
OsmC	2	2	response to oxidative stress
NIR_SIR	2	1	oxidoreductase activity; heme binding; iron-sulfur cluster binding; oxidation-reduction process
PhageP22-tail	2	1	NA
Calreticulin	2	2	calcium ion binding; unfolded protein binding; protein folding; endoplasmic reticulum

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Ribonuclease_T2	2	1	RNA binding; ribonuclease T2 activity
Pfam-B_827	2	2	NA
Profilin	2	1	NA
Uma2	2	2	NA
Bet_v_1	2	1	defense response; response to biotic stimulus
Lipoprotein_1	2	2	cell outer membrane
Oxysterol_BP	2	1	NA
Cna_B	2	2	NA
PMI_typeI	2	1	mannose-6-phosphate isomerase activity; zinc ion binding; carbohydrate metabolic process
Fn3-like	2	1	NA
Acetate_kinase	2	2	kinase activity; phosphotransferase activity, carboxyl group as acceptor; metabolic process; phosphorylation; intracellular
ASF1_hist_chap	2	1	chromatin assembly or disassembly; nucleus
Ring_hydroxyl_B	2	1	catalytic activity; cellular aromatic compound metabolic process; oxidation-reduction process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
tRNA_anti-codon	2	1	nucleic acid binding
KAT11	2	2	NA
HATPase_c	2	1	NA
Trypsin	2	2	serine-type endopeptidase activity; proteolysis
Spt4	2	1	NA
IMS	2	1	damaged DNA binding; DNA-directed DNA polymerase activity; DNA repair
peroxidase	2	1	peroxidase activity; heme binding; response to oxidative stress; oxidation-reduction process
Pfam-B_7	2	1	NA
Pro_racemase	2	2	proline racemase activity
PfkB	2	2	NA
AICARFT_IMPCHas	2	1	IMP cyclohydrolase activity; phosphoribosylaminoimida-zolecarboxamide formyltransferase activity; purine nucleotide biosynthetic process
FAA_hydrolase	2	2	catalytic activity; metabolic process
YiiD_Cterm	2	1	NA

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
Molybdop_Fe4S4	2	1	oxidoreductase activity; oxidation-reduction process
Acetyltransf_6	2	1	NA
Acetyltransf_4	2	2	NA
PRC	2	1	NA
C2	2	2	protein binding
RNase_PH_C	2	1	NA
NMT	2	1	glycylpeptide N-tetradecanoyltransferase activity
Rieske	2	1	oxidoreductase activity; 2 iron, 2 sulfur cluster binding; oxidation-reduction process
PPC	2	1	NA
Bgal_small_N	2	2	beta-galactosidase activity; carbohydrate metabolic process; beta-galactosidase complex
Pfam-B_5281	2	2	NA
Cleaved_Adhesin	2	1	NA
PPV_E2_N	2	1	regulation of DNA replication; regulation of transcription, DNA-templated; viral process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
NMT_C	2	1	glycylpeptide N-tetradecanoyltransferase activity
PID	2	1	protein binding
GNAT_acetyltran	2	2	NA
SnoaL_2	2	2	NA
FAD_binding_3	2	2	NA
ROK	2	1	NA
Glyco_hydro_38C	2	2	mannosidase activity; mannose metabolic process
DNA_topoisoIV	2	1	DNA binding; DNA topoisomerase type II (ATP-hydrolyzing) activity; ATP binding; DNA topological change
Glyoxalase	2	2	NA
DUF377	2	2	NA
Arrestin_C	2	1	NA
GMP_PDE_delta	2	2	NA
Pfam-B_512	2	1	NA
Arena_nucleocap	2	1	viral nucleocapsid
ADH_N	2	1	oxidoreductase activity; oxidation-reduction process

Continued on next page

Table B.1 – *Continued from previous page*

PFam Architecture	Matches	Aligns	GO tags
NUDIX	2	2	hydrolase activity

APPENDIX C

DE NOVO CDR H3 MODELING IN ROSETTA

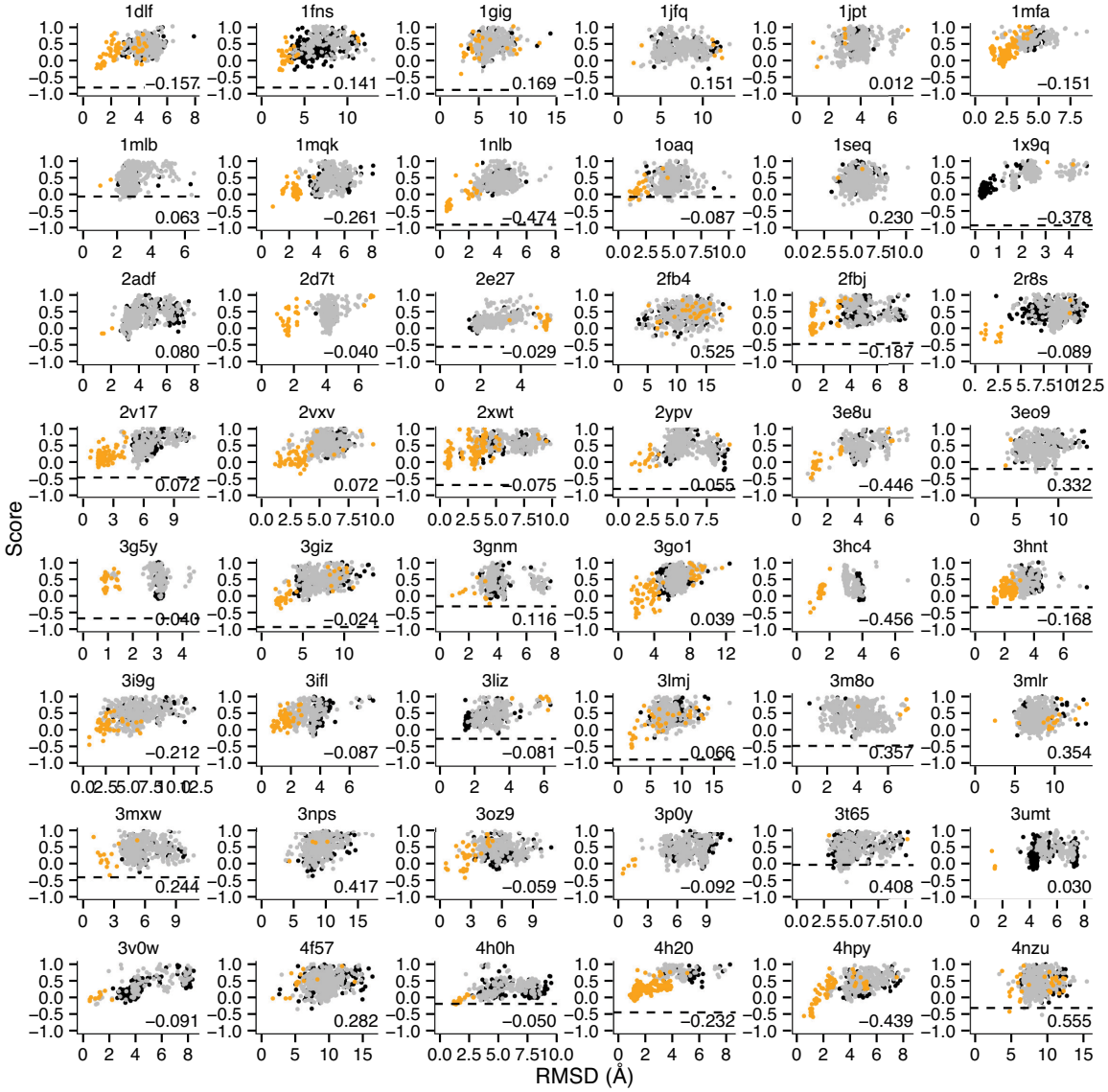

Figure C.1: Funnel plots showing scaled score vs. RMSD for unconstrained *de novo* NGK on the H3 loop benchmark set. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score is shown in the lower right of the plot area for each target. Very few kinked H3 models are produced, but, for kinked targets, they are frequently the lowest-RMSD models produced. The dashed horizontal line indicates the scaled score of the native structure. The scaled native score is too low to show up on the plot at the scale shown for 23 of the targets, indicating that there are inaccuracies even in the low-RMSD models.

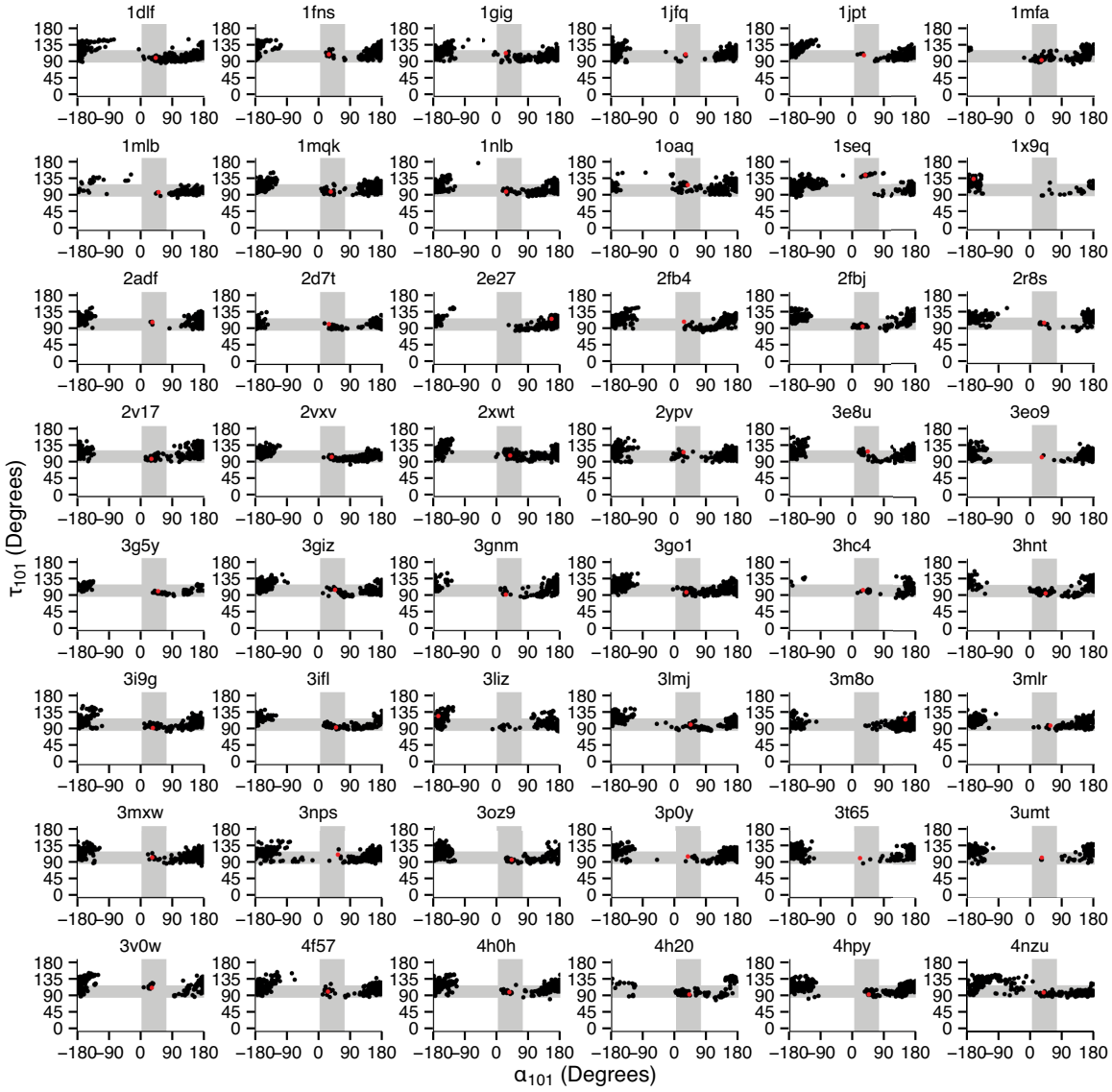

Figure C.2: τ_{101} vs. α_{101} plots for unconstrained *de novo* NGK on the H3 loop benchmark set. The • point is at the values of the native structure, and the • points correspond to the models. The vast majority of the points have τ_{101} and α_{101} values that correspond to extended conformations.

Table C.1: Quantitative results for unconstrained *de novo* NGK

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
1dlf	1.0057	-0.8328	1.2948	1.8636	1.0057
1fns	1.9961	-0.8346	2.2998	4.3725	2.3848
1gig	2.5965	-0.9055	3.1920	5.5097	2.5994
1jfq	1.8429	-1.0615	3.3965	5.1614	10.7125
1jpt	0.9972	-1.4458	1.7116	2.6516	1.2253
1mfa	1.2040	-1.7568	1.3367	2.1203	2.1916
1mlb	1.0512	-0.0720	1.9308	2.6627	2.2141
1mqk	0.8169	-1.0959	1.7024	3.3905	0.8169
1nlb	0.4001	-0.9162	0.4869	0.5303	0.4173
1oaq	1.0488	-0.0821	1.2219	2.3602	1.1193
2adf	1.4002	-1.0710	2.3663	2.8760	3.2794
2d7t	0.9517	-1.2204	1.4048	3.0854	1.6668
2fb4	3.0953	-1.2726	4.1053	10.2298	14.6668
2fbj	0.9700	-0.4629	1.0430	1.2078	1.1228
2r8s	0.6297	-1.2372	1.7502	3.2798	2.2758
2v17	0.7926	-0.4927	1.4139	2.6154	5.4328
2vxv	1.1711	-1.2601	1.5343	2.6953	3.3303
2w60	0.4450	-1.4061	1.5948	2.9602	0.9427
2xwt	0.3709	-0.7166	0.5898	1.9048	3.0002
2ypv	1.1427	-0.8333	2.0470	4.9384	1.8491
3e8u	0.6829	-1.6625	0.9632	1.1115	1.0002
3eo9	3.6195	-0.2207	4.1799	7.7170	9.6596
3g5y	0.7744	-0.6838	0.8740	3.0755	3.0403
3giz	1.7031	-0.9452	1.9713	2.5700	2.3243
3gnm	0.7956	-0.3262	1.9960	3.8477	3.5173
3go1	1.4950	-1.7705	1.8737	2.7053	1.9785
3hc4	0.7802	-1.0223	1.1963	2.4972	0.8056
3hnt	1.2357	-0.3490	1.2818	1.3508	1.4207
3i9g	0.7581	-1.1165	1.6035	3.4776	0.7707
3ifl	0.7096	-1.2403	0.8878	3.3150	3.5032
3lmj	1.8205	-0.8880	2.3313	3.3540	2.1587
3mlr	2.4555	-1.4629	4.9032	7.2062	8.2206
3mxw	1.1074	-0.4166	1.9946	4.9173	2.6573
3nps	3.8113	-1.1921	4.5338	8.3031	7.3586
3oz9	0.5730	-1.0583	1.4198	2.3506	2.4056
3p0y	0.4399	-1.5656	1.8978	5.9468	0.4399
3t65	2.7751	-0.0409	3.0369	4.1281	4.4965

Continued on next page

Table C.1 – *Continued from previous page*

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
3umt	1.2153	-1.0976	2.4331	3.4288	4.2158
3v0w	0.5647	-1.1450	0.9769	2.4002	1.0366
4f57	1.7381	-2.0141	3.7467	7.9585	5.3980
4h0h	1.1919	-0.1839	1.5712	1.9698	1.3644
4h20	0.8145	-0.4427	0.9513	1.2454	1.1819
4hpy	0.4970	-1.5788	0.9820	0.9909	1.1620
4nzu	3.2970	-0.3145	4.7625	8.0227	9.1961
MEAN	**1.3360**	**-0.9480**	**2.0180**	**3.6433**	**3.2174**
STD DEV	**0.8890**	**0.5033**	**1.1448**	**2.2258**	**3.0977**
1x9q	0.1933	-0.9542	0.2150	0.2953	0.2810
2e27	1.3989	-0.5466	1.5764	2.0258	1.7931
3m8o	0.8086	-0.4768	1.5090	3.7114	3.7824
3liz	1.3761	-0.2621	1.4268	3.3762	3.3939
MEAN	**0.9442**	**-0.5599**	**1.1818**	**2.3522**	**2.3126**
STD DEV	**0.5702**	**0.2894**	**0.6474**	**1.5528**	**1.6049**
1seq	3.1272	-2.0754	3.6206	5.6491	4.2490

Table C.1: Results of the unconstrained NGK simulation for each target in the benchmark set. The results are split up by the base geometry of the native structure to show the difference in performance across different base geometries. All RMSDs are reported in Ångströms.

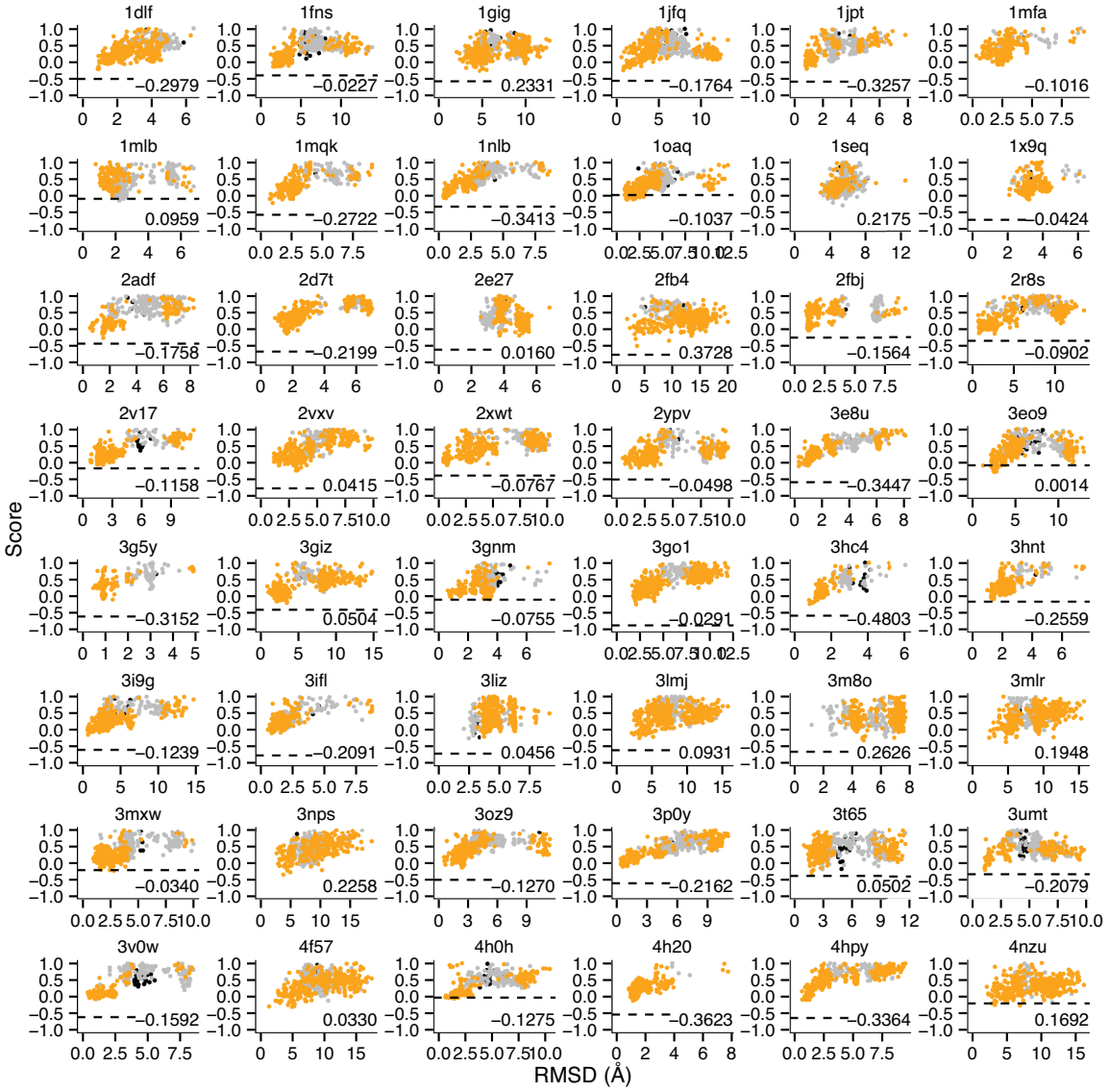

Figure C.3: Funnel plots showing scaled score vs. RMSD for constrained *de novo* NGK on the H3 loop benchmark set. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score is shown in the lower right of the plot area for each target. Many kinked H3 models are produced, and the top-scoring models often have sub-Ångström RMSDs. The dashed horizontal line indicates the scaled score of the native structure. The scaled native score is too low to show up on the plot at the scale shown for 5 of the targets, indicating that the constraint consistently enables Rosetta to produce models nearly as energetically favorable as the native structure.

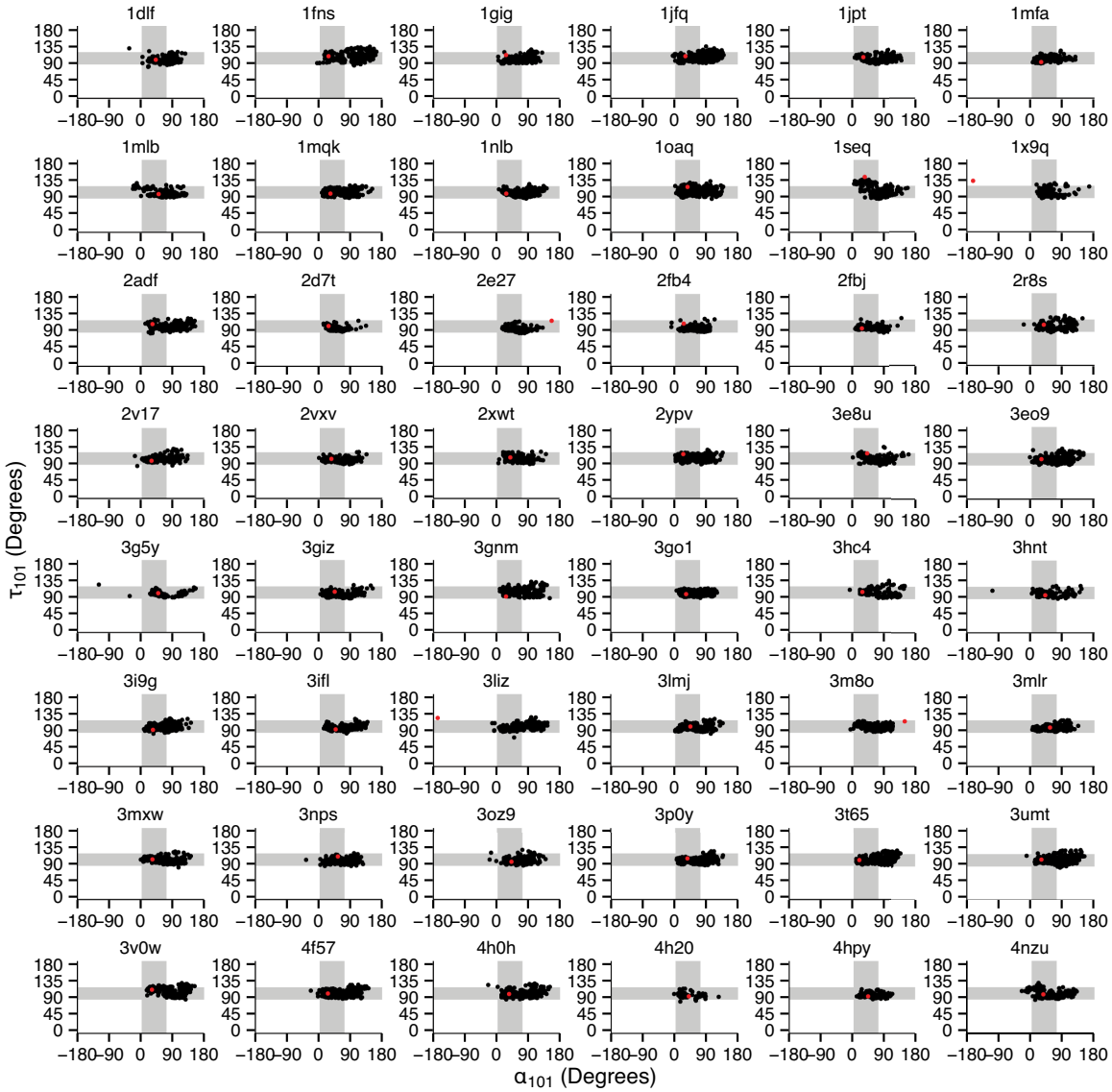

Figure C.4: τ_{101} vs. α_{101} plots for constrained *de novo* NGK on the H3 loop benchmark set. The • point is at the values of the native structure, and the • points correspond to the models. The vast majority of the points have τ_{101} and α_{101} values that correspond to kinked conformations. This is problematic for targets 1x9q, 2e27, 3liz, and 3m8o, which have extended base geometries, but 1seq, which has an unclear base geometry, is able to sample conformations at near-native τ_{101} and α_{101} values.

Table C.2: Quantitative results for constrained *de novo* NGK

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
1dlf	0.8847	-0.5242	0.9543	1.2657	0.8847
1fns	1.2479	-0.4160	1.5903	2.3255	1.6671
1gig	1.9806	-0.5957	2.7223	4.5488	4.6625
1jfq	0.5630	-0.5800	0.9476	1.3713	0.7621
1jpt	0.7319	-0.6053	0.7636	0.9550	0.8063
1mfa	0.3159	-1.1901	0.7432	2.2098	2.1495
1mlb	0.8055	-0.1031	1.0735	2.3342	2.2597
1mqk	0.7269	-0.5840	1.1203	1.9200	0.9632
1nlb	0.3155	-0.3380	0.3500	0.4191	0.5055
1oaq	0.8637	0.0092	0.9633	1.6457	1.1311
2adf	0.5459	-0.4331	0.9334	1.4835	1.5324
2d7t	0.8841	-0.6717	1.1722	1.6265	1.6384
2fb4	1.8331	-0.7694	3.1046	8.1272	3.6253
2fbj	0.8883	-0.2536	0.9904	1.1380	1.1326
2r8s	0.6658	-0.3648	0.7603	1.7915	2.6410
2v17	0.8039	-0.1895	0.9254	1.7289	2.0932
2vxv	1.1523	-0.7891	1.2648	2.9963	3.1094
2w60	0.3128	-0.2736	0.3864	0.7543	0.9680
2xwt	0.3651	-0.4075	0.4692	1.0994	0.7277
2ypv	0.5321	-0.5230	0.8209	2.3432	1.2450
3e8u	0.2630	-0.6061	0.5007	0.8308	0.2655
3eo9	1.2092	-0.0906	2.1158	2.5706	2.5314
3g5y	0.4713	-0.6198	0.6840	0.9594	1.3308
3giz	0.8620	-0.4162	1.3519	2.6305	3.4490
3gnm	0.5754	-0.1147	0.7056	3.1564	3.5112
3go1	1.6585	-0.8874	1.8769	2.6723	2.0606
3hc4	0.7545	-0.5954	0.9895	1.0121	0.8019
3hnt	0.9713	-0.1727	1.2096	1.3374	1.3806
3i9g	0.6414	-0.6051	0.7875	1.2535	0.7310
3ifl	0.5588	-0.7728	0.7645	0.9560	0.8367
3lmj	2.1115	-0.6141	2.5146	3.5080	3.2288
3mlr	2.2999	-1.0970	3.0873	5.5436	4.6692
3mxw	0.9767	-0.2255	1.1118	2.3990	2.8328
3nps	2.6624	-1.1616	3.1158	5.6184	3.7979
3oz9	0.5176	-0.5157	0.7455	2.2569	2.3685
3p0y	0.3675	-0.6203	0.4141	0.6348	0.5905
3t65	0.9007	-0.3983	1.3672	3.0506	1.7030

Continued on next page

Table C.2 – *Continued from previous page*

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
3umt	1.0396	-0.3304	1.1139	1.2401	1.0631
3v0w	0.4142	-0.6155	0.5060	0.9577	1.0368
4f57	1.3136	-1.6805	2.6350	4.1318	4.9528
4h0h	0.6141	-0.0298	0.8704	1.4223	1.4427
4h20	0.8209	-0.5384	0.9284	1.2520	1.1060
4hpy	0.5774	-0.6418	0.8136	0.9908	0.9373
4nzu	2.0598	-0.2073	2.6145	5.1172	6.8650
MEAN	**0.9332**	**-0.5264**	**1.2473**	**2.2179**	**2.0000**
STD DEV	**0.5808**	**0.3316**	**0.7847**	**1.5891**	**1.4331**
1x9q	2.0763	-0.7465	2.2300	3.1290	3.2012
2e27	2.5348	-0.6156	2.6465	4.4264	3.7690
3m8o	1.1371	-0.6624	2.2234	4.8490	4.7532
3liz	2.0070	-0.7202	2.5639	3.4758	2.4243
MEAN	**1.9388**	**-0.6862**	**2.4159**	**3.9700**	**3.5369**
STD DEV	**0.5835**	**0.0587**	**0.2211**	**0.8026**	**0.9805**
1seq	2.8814	-1.0669	3.2300	4.5532	5.3418

Table C.2: Results of the constrained NGK simulation for each target in the benchmark set. The results are split up by the base geometry of the native structure to prevent the constraint from confounding the results. All RMSDs are reported in Ångströms.

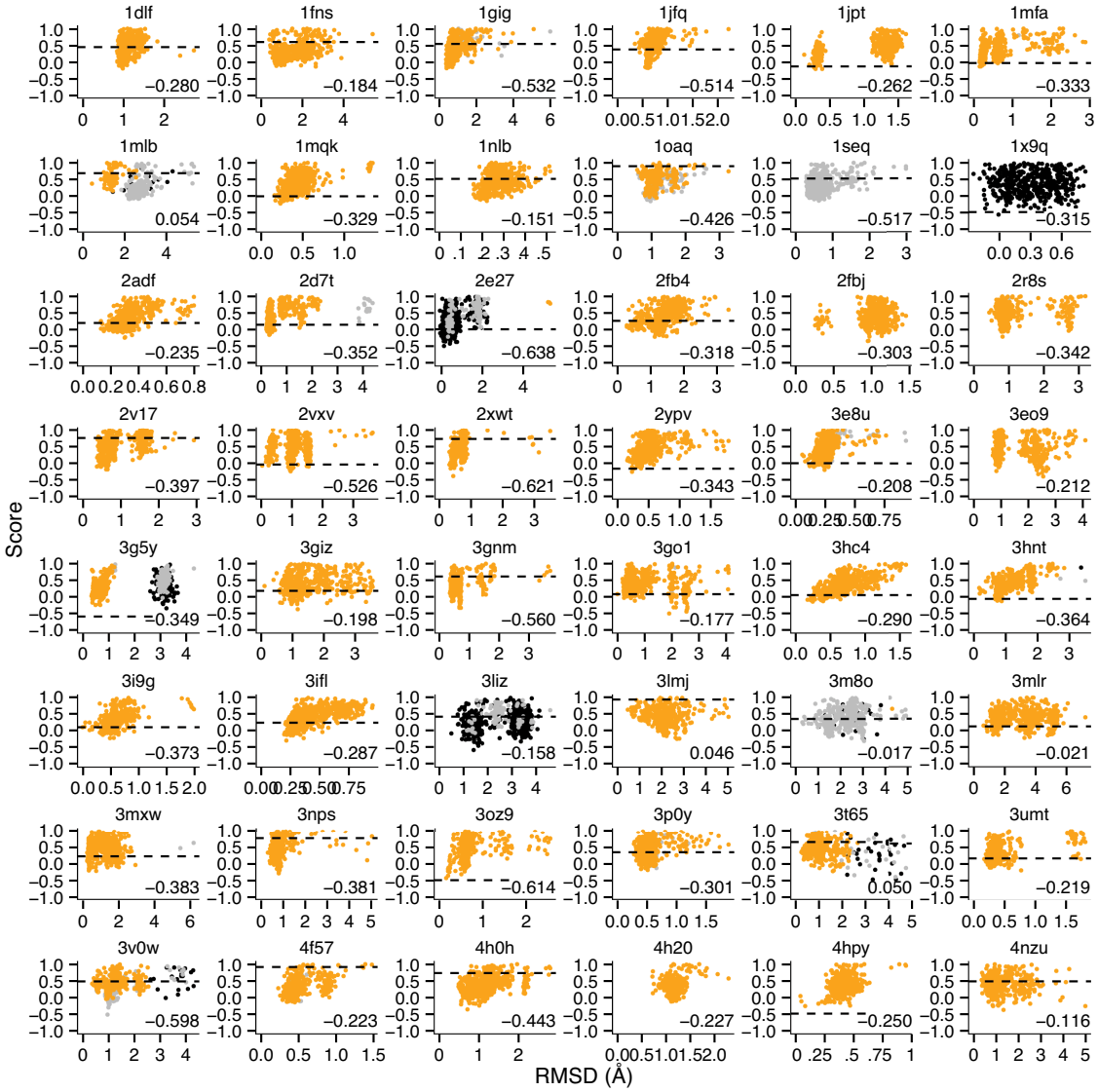

Figure C.5: Funnel plot showing discrimination score vs. RMSD. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score is shown in the lower right of the plot area for each target. The dashed horizontal line indicates the scaled score of the native structure. The NGK refinement simulations start with the native loop conformation and do not use constraints. Most of the scaled native scores are greater than zero, meaning the refinement protocol produces a significant number of models that score better than the native conformation.

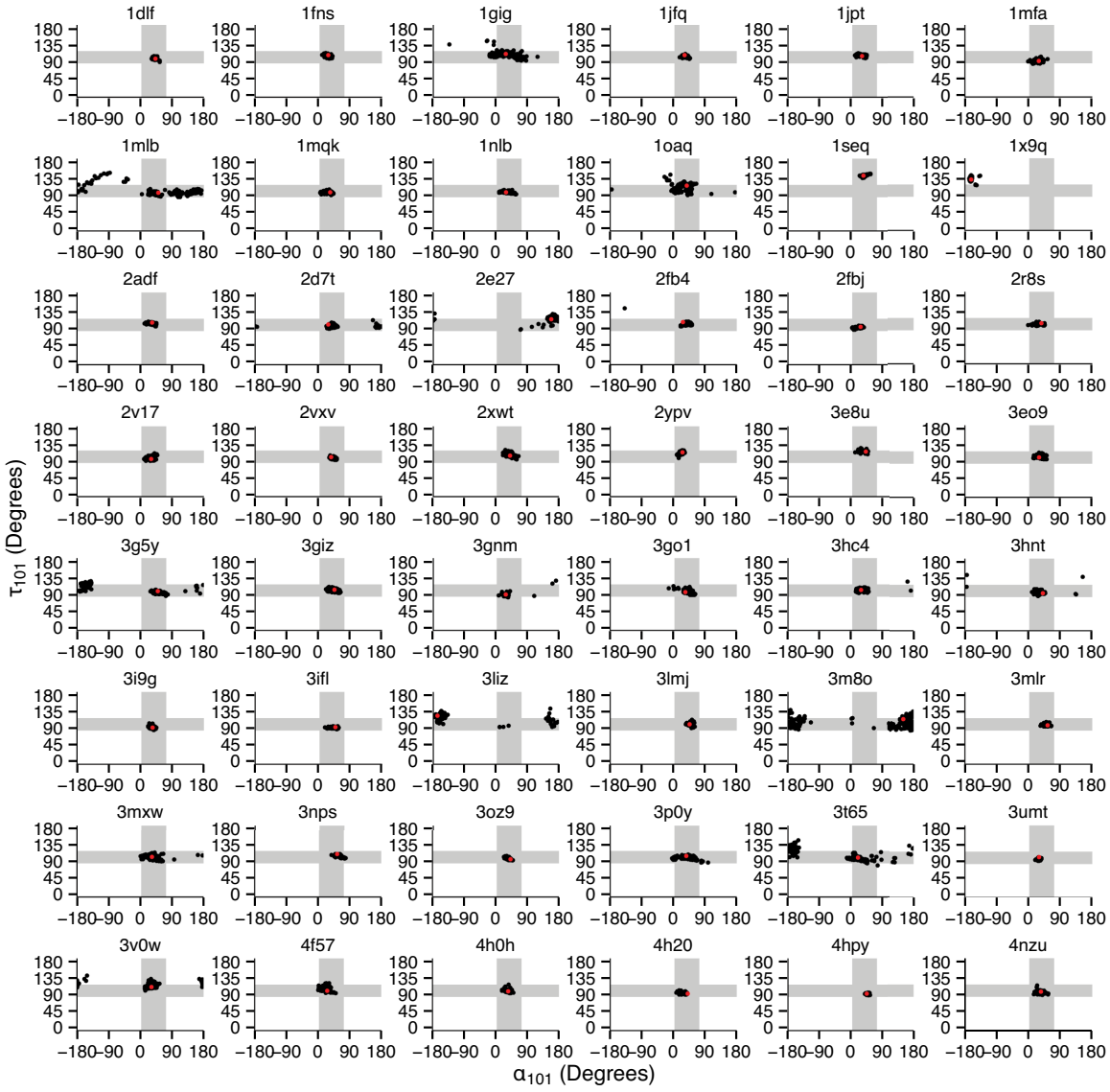

Figure C.6: τ_{101} vs. α_{101} plots for NGK refinement on the H3 loop benchmark set. The • point is at the values of the native structure, and the • points correspond to the models. The NGK refinement simulations start with the native loop conformation and do not use constraints. Most of the points remain tightly clustered around the native values, but some models transition from kinked to extended during refinement.

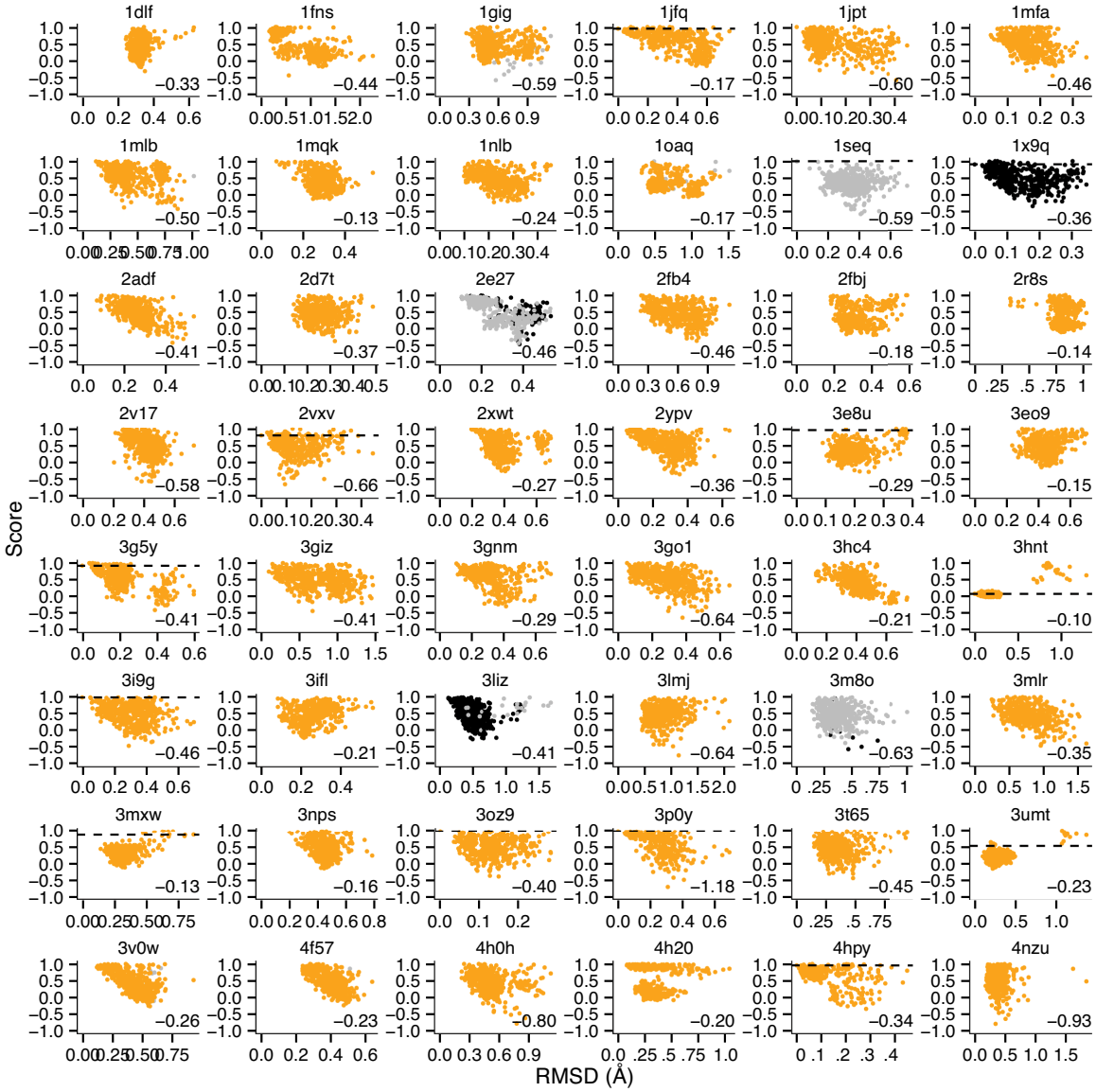

Figure C.7: Funnel plot showing discrimination score vs. RMSD. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score is shown in the lower right of the plot area for each target. The dashed horizontal line indicates the scaled score of the native structure. The CCD refinement simulations start with the native loop conformation and do not use constraints. Most of the scaled native scores are greater than zero, meaning the refinement protocol produces a significant number of models that score better than the native conformation. Unlike previous funnel plot figures, some scaled native scores do not appear due to them being too high to be included in the plot area.

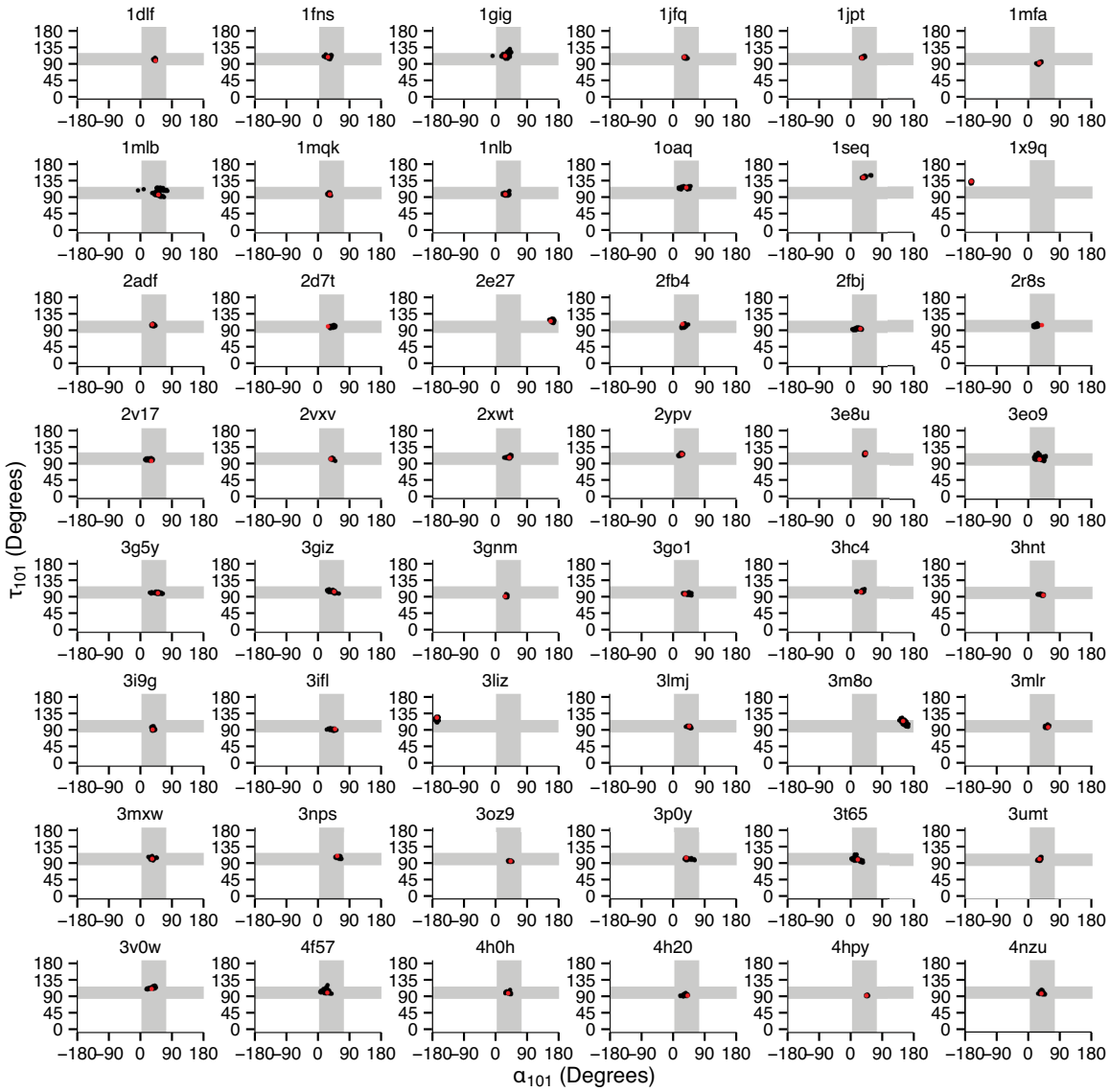

Figure C.8: τ_{101} vs. α_{101} plots for NGK refinement on the H3 loop benchmark set. The • point is at the values of the native structure, and the • points correspond to the models. The CCD refinement simulations start with the native loop conformation and do not use constraints. For all targets, the points remain tightly clustered around the native values and there are no transitions from kinked to extended during refinement. Less sampling of τ_{101} and α_{101}, shows that the CCD refinement process moves the loop less than NGK refinement.

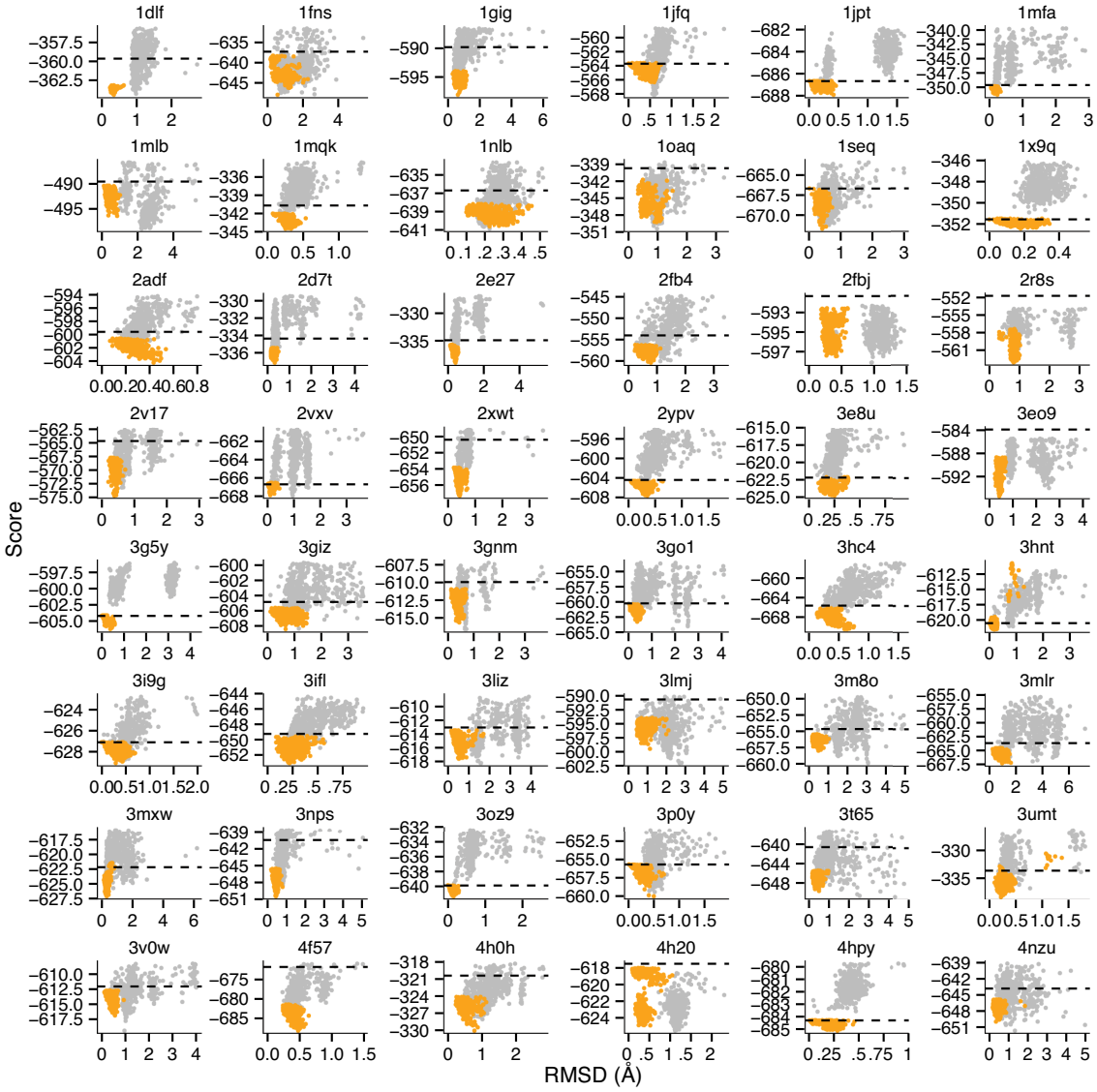

Figure C.9: Funnel plot showing total score vs. RMSD. • points correspond to models produced by CCD, • points to models produced by NGK, and the dashed horizontal line indicates the score of the native structure. Both sets of simulations start with the native loop conformation and do not use constraints. All of the CCD models have a more favorable score than the native structure and very low RMSDs ($\leq$ 0.5 Å), while many of the NGK models have higher scores than the native structure and and larger RMSDs than the CCD models. CCD samples less space than NGK, but achieves lower scores.

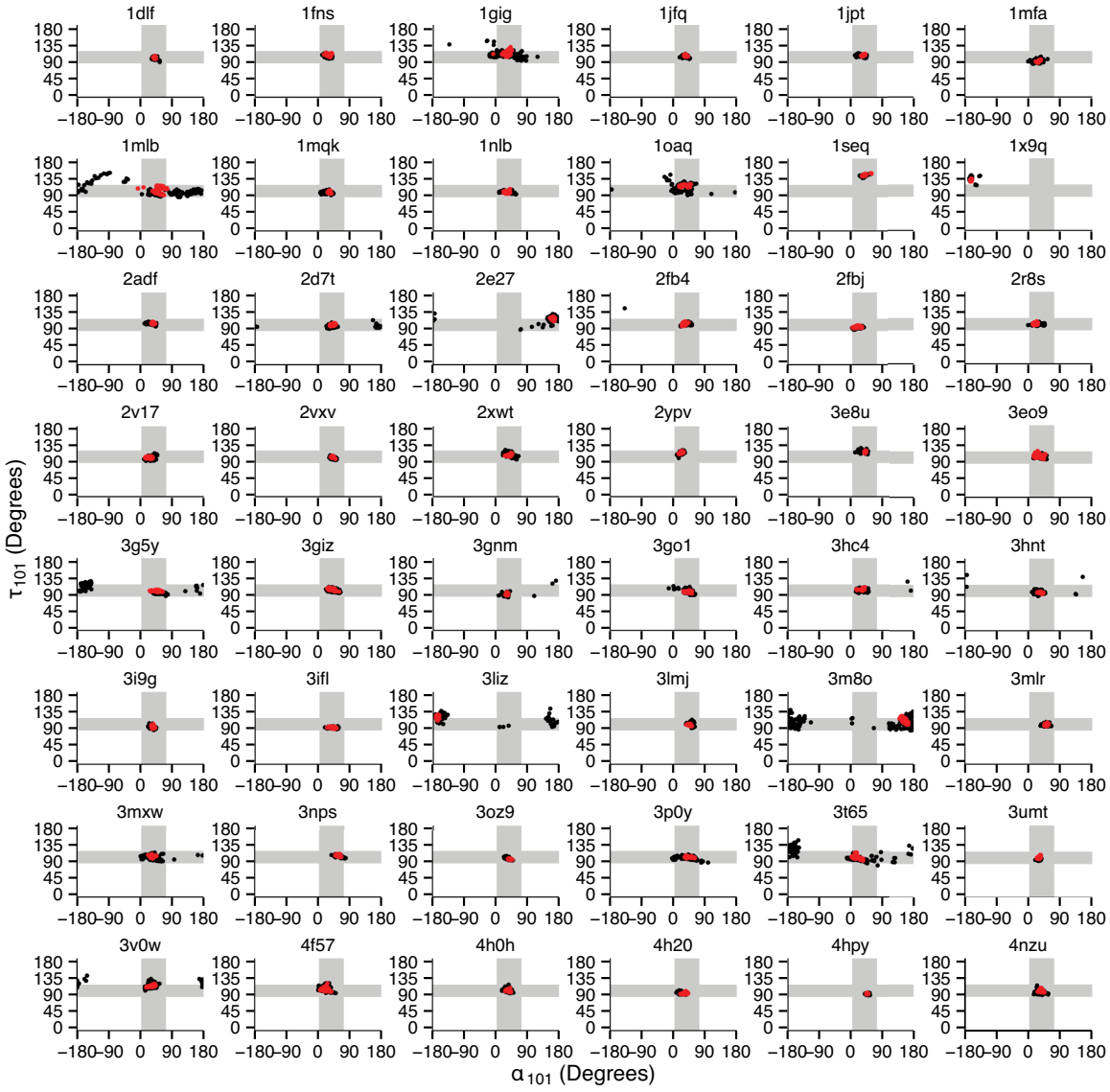

Figure C.10: τ_{101} vs. α_{101}. The • points correspond to models generated with CCD, and the • points correspond to models generated with NGK. Both sets of simulations start with the native loop conformation and do not use constraints. The CCD models are distributed in a smaller region of the plot than the NGK models. CCD does less sampling of τ_{101} and α_{101} than NGK, but achieves lower scores.

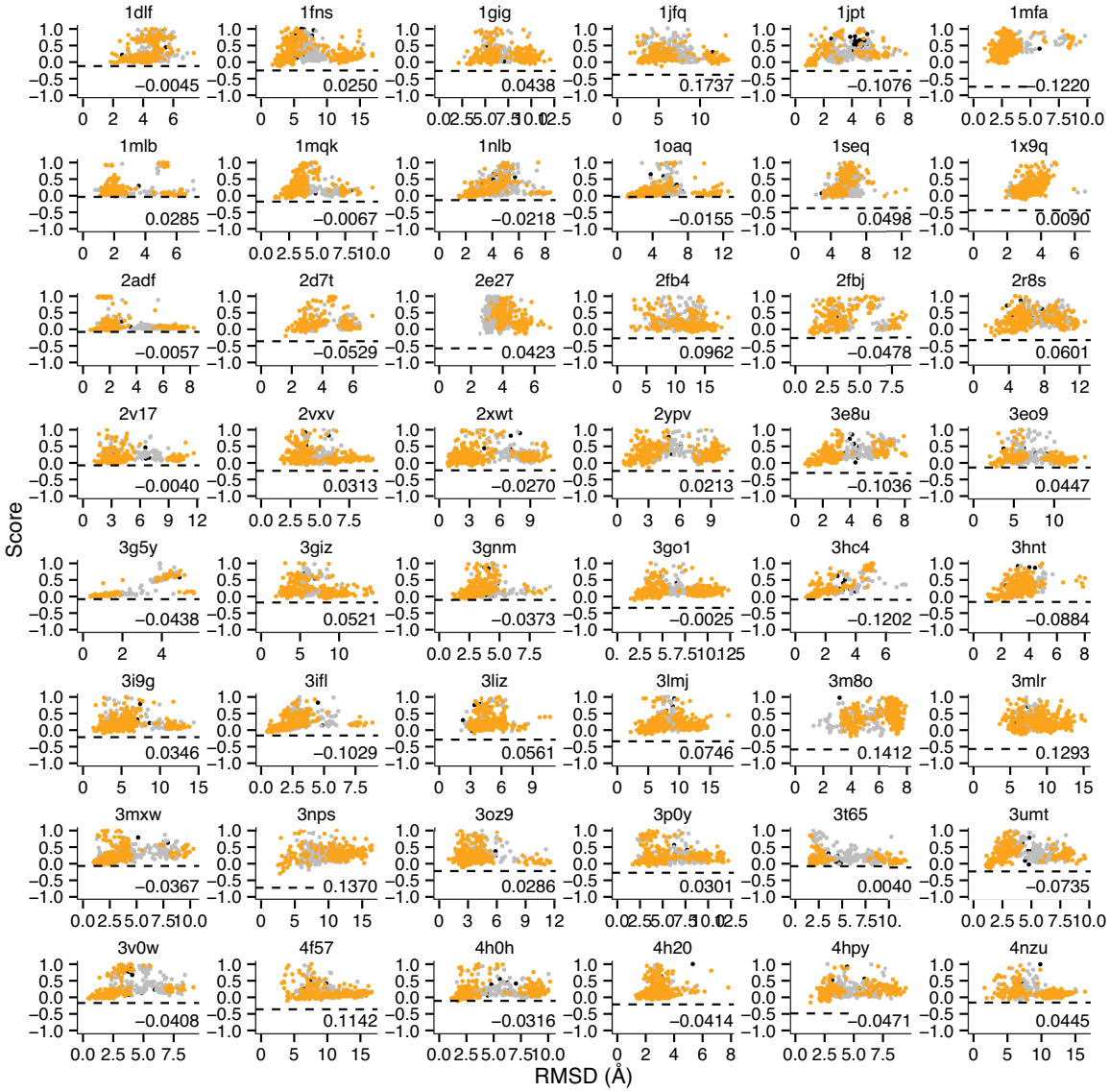

Figure C.11: Funnel plots showing scaled score vs. RMSD for constrained *de novo* NGK+CCD on the H3 loop benchmark set. • points correspond to a kinked base geometry, • points to an unclear base geometry, and • points to an extended base geometry. The discrimination score is shown in the lower right of the plot area for each target. The models are predominantly kinked, although with more models with unclear base geometry than in the constrained NGK simulation. The dashed horizontal line indicates the scaled score of the native structure, which is higher than in the unconstrained and constrained simulations, and is now visible for all targets. This shows that the score of the models is approaching that of the native structure, however many models across a wide range of RMSD values have similar scores. The discrimination score is higher for all of the targets, indicating that the degree of refinement reduces the ability of the score function to identify near-native conformations.

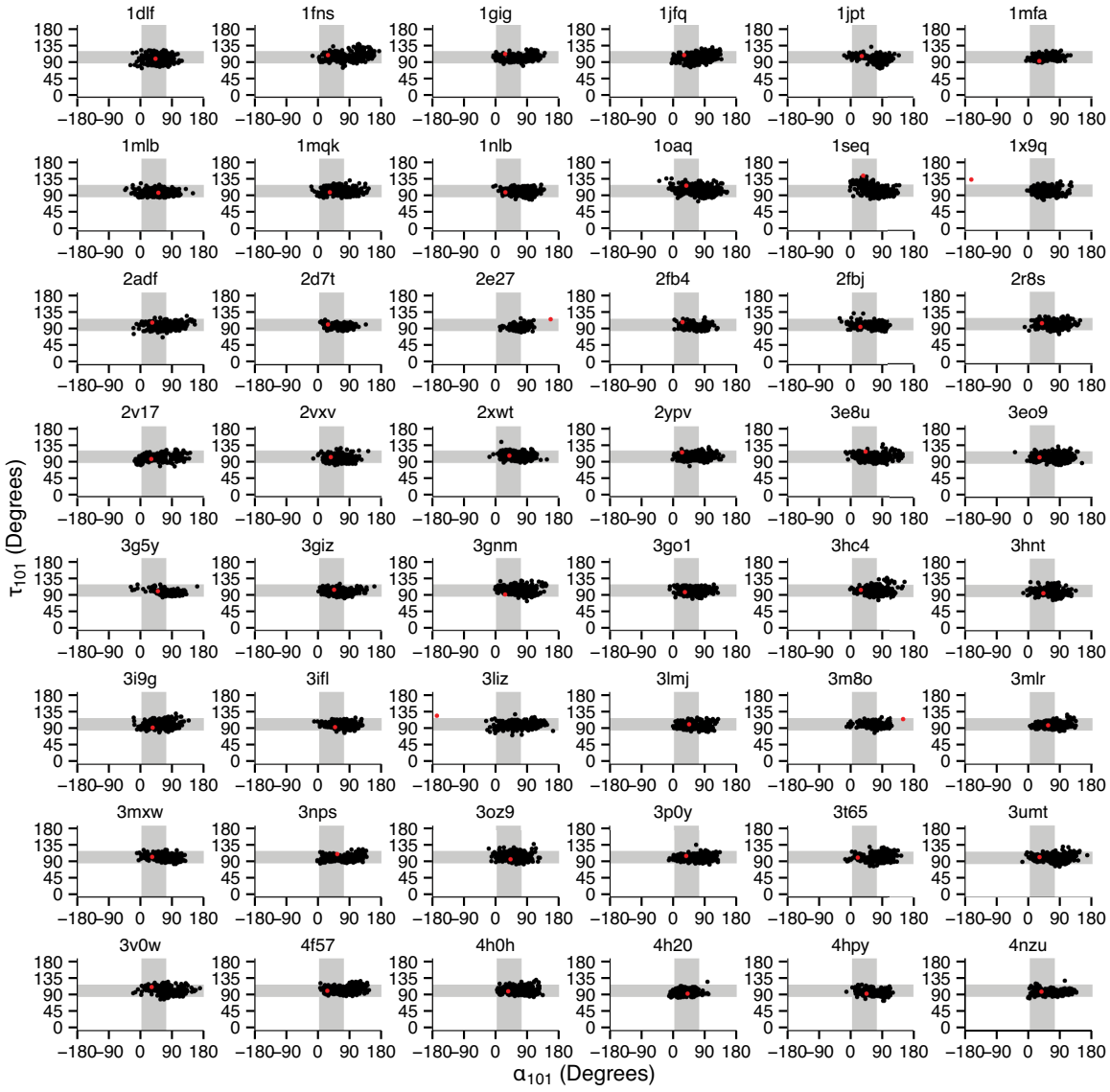

Figure C.12: τ_{101} vs. α_{101} plots for unconstrained *de novo* NGK+CCD on the H3 loop benchmark set. The • point is at the values of the native structure, and the • points correspond to the models. The points are much more widely distributed than in the constrained simulation and appear to cluster into two large groups, indicating that CCD was able to find a conformations that could offset the penalty imposed by the constraint.

Table C.3: Quantitative results for *de novo* NGK+CCD

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
1dlf	1.8731	-0.1371	2.2637	3.5617	3.3879
1fns	1.7367	-0.2661	2.2175	3.1091	2.8515
1gig	2.0493	-0.2807	3.1017	6.2559	5.6189
1jfq	1.5839	-0.3958	2.4753	7.1683	5.3348
1jpt	0.7126	-0.2825	0.8827	1.0769	0.7566
1mfa	1.0076	-0.7589	1.4710	2.4539	2.2864
1mlb	0.8296	-0.0401	1.1528	2.7493	2.1168
1mqk	1.2493	-0.1815	1.8356	2.7517	2.2314
1nlb	1.4574	-0.1354	1.8052	2.1917	1.5052
1oaq	1.1750	-0.0382	1.5320	2.1056	1.4066
2adf	0.5948	-0.0690	0.9090	2.0446	2.2096
2d7t	1.5838	-0.3509	2.0063	2.6899	1.5838
2fb4	2.4299	-0.2591	3.2692	12.4221	12.8322
2fbj	1.1181	-0.2544	1.2942	1.9516	1.3661
2r8s	1.3883	-0.3326	2.4096	3.3968	2.4095
2v17	1.2903	-0.0809	1.6221	2.5497	1.6868
2vxv	1.6912	-0.2455	2.0042	4.0897	3.7512
2w60	0.4083	-0.1444	0.6479	1.8338	0.4083
2xwt	0.7883	-0.2310	0.9380	2.4960	0.8652
2ypv	0.6068	-0.2493	0.9685	2.2491	3.0346
3e8u	0.7153	-0.3095	0.9608	1.2381	0.9581
3eo9	2.0257	-0.1378	2.7184	6.4946	2.8518
3g5y	0.4076	-0.0864	0.5173	0.7566	0.4076
3giz	1.8989	-0.1839	2.4836	7.0498	8.3471
3gnm	1.0890	-0.1088	1.5049	1.9327	1.6143
3go1	1.5085	-0.3463	2.1345	2.7728	1.9208
3hc4	0.6972	-0.0916	0.8092	0.9105	0.7954
3hnt	1.0599	-0.1567	1.3784	1.8165	1.3393
3i9g	1.4277	-0.2117	1.8726	4.4273	2.6732
3ifl	0.4775	-0.1634	0.6381	0.8673	0.8497
3lmj	2.1604	-0.3371	3.4260	4.8165	4.5231
3mlr	3.5985	-0.5553	4.3633	7.0475	6.5272
3mxw	1.0343	-0.0901	1.2049	1.9969	1.9518
3nps	2.5918	-0.7404	3.1771	4.9381	2.8715
3oz9	1.0039	-0.2430	1.6190	2.6541	2.5357
3p0y	1.7204	-0.2929	2.2138	3.8022	3.4332
3t65	1.4879	-0.0950	1.5587	3.1810	1.7673

Continued on next page

Table C.3 – *Continued from previous page*

Target	Minimum RMSD	Scaled Native Score	Top 10 RMSDs	RMSD of Top 10 Scored	RMSD of Top 1 Scored
3umt	1.0984	-0.2136	1.1949	1.3720	1.4351
3v0w	0.4941	-0.1642	0.7036	1.2600	0.6040
4f57	3.1961	-0.3528	3.7584	6.9276	4.0201
4h0h	1.1704	-0.1001	1.5051	2.0440	1.7095
4h20	1.3224	-0.2056	1.4601	2.4175	2.4093
4hpy	1.2855	-0.4757	1.9110	2.3319	1.2855
4nzu	1.6257	-0.1482	3.0332	8.0501	6.8078
MEAN	**1.3789**	**-0.2396**	**1.8398**	**3.4148**	**2.7564**
STD DEV	**0.6966**	**0.1603**	**0.9100**	**2.3858**	**2.3569**
1x9q	1.9131	-0.4531	2.0984	2.7843	3.0757
2e27	2.5599	-0.5630	2.7943	4.2942	5.4418
3m8o	1.2944	-0.5833	2.0798	5.5863	3.9549
3liz	2.0570	-0.2890	2.5411	4.4809	3.3138
MEAN	**1.9561**	**-0.4721**	**2.3784**	**4.2864**	**3.9466**
STD DEV	**0.5211**	**0.1348**	**0.3497**	**1.1524**	**1.0637**
1seq	2.3088	-0.3780	3.0652	4.5330	3.0050

Table C.3: Results of the NGK+CCD simulation for each target in the benchmark set. The results are split up by the base geometry of the native structure to prevent the constraint from confounding the results. All RMSDs are reported in Ångströms.

REFERENCES

1. Moja L, Tagliabue L, Balduzzi S, Parmelli E *et al.* (2012) Trastuzumab containing regimens for early breast cancer. *Cochrane Database Syst Rev* **4**(1), CD006243.

2. Balduzzi S, Mantarro S, Guarneri V, Tagliabue L *et al.* (2014) Trastuzumab-containing regimens for metastatic breast cancer. *Cochrane Database Syst Rev* **6**(1), CD006242.

3. BioPharma PDL Inc (2014). Historical Product Sales Revenue.

4. Xu J, Tack D, Hughes RA, Ellington AD & Gray JJ (2013) Structure-based non-canonical amino acid design to covalently crosslink an antibody–antigen complex. *J Struct Biol* **185**(2), 215–222.

5. Engvall E & Perlmann P (1971) Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **8**(9), 871–874.

6. Van Weemen B & Schuurs A (1971) Immunoassay using antigen–enzyme conjugates. *FEBS Lett* **15**(3), 232–236.

7. Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* **302**(5909), 575–581.

8. Alberts B, Johnson A, Lewis J, Raff M *et al.* (2002) Molecular Biology of the Cell. Garland Science. ISBN 0815332181, 1616 .

9. Georgiou G, Ippolito GC, Beausang J, Busse CE *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* **32**(2), 158–168.

10. Alzari PM, Lascombe MB & Poljak RJ (1988) Three-dimensional structure of antibodies. *Annu Rev Immunol* **6**(1), 555–580.

11. Davies DR, Padlan EA & Sheriff S (1990) Antibody–antigen complexes. *Annu Rev Biochem* **59**(1), 439–473.

12. Jones PT, Dear PH, Foote J, Neuberger MS & Winter G (1986) Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321**(6069), 522–525.

13. Weinstein JA, Jiang N, White RA, Fisher DS & Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**(5928), 807–810.

14. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ *et al.* (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* **31**(2), 166–169.

15. Berman H, Henrick K & Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Mol Biol* **10**(12), 980.

16. Chothia C & Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* **196**(4), 901–917.

17. Chothia C, Lesk AM, Tramontano A, Levitt M *et al.* (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342**(6252), 877–883.

18. Al-Lazikani B, Lesk AM & Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* **273**(4), 927–948.

19. North B, Lehmann A & Dunbrack, Jr RL (2011) A new clustering of antibody CDR loop conformations. *J Mol Biol* **406**(2), 228–256.

20. Chacko S, Silverton EW, Smith-Gill SJ, Davies DR *et al.* (1996) Refined structures of bobwhite quail lysozyme uncomplexed and complexed with the HyHEL-5 F_{ab} fragment. *Proteins* **26**(1), 55–65.

21. Sircar A (2010) Computational antibody structure prediction and antibody–antigen docking. Phd, Johns Hopkins University.

22. Sivasubramanian A, Sircar A, Chaudhury S & Gray JJ (2009) Toward high-resolution homology modeling of antibody F_V regions and application to antibody–antigen docking. *Proteins* **74**(2), 497–514.

23. Almagro JC, Beavers MP, Hernandez-Guzman F, Maier J *et al.* (2011) Antibody modeling assessment. *Proteins* **79**(11), 3050–3066.

24. Shirai H, Kidera A & Nakamura H (1996) Structural classification of CDR-H3 in antibodies. *FEBS Lett* **399**(1–2), 1–8.

25. Shirai H, Kidera A & Nakamura H (1999) H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett* **455**(1–2), 188–197.

26. Kuroda D, Shirai H, Kobori M & Nakamura H (2008) Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* **73**(3), 608–620.

27. Morea V, Tramontano A, Rustici M, Chothia C & Lesk AM (1997) Antibody structure, prediction and redesign. *Biophys Chem* **68**(1-3), 9–16.

28. Morea V, Tramontano A, Rustici M, Chothia C & Lesk AM (1998) Conformations of the third hypervariable region in the V_textH domain of immunoglobulins. *J Mol Biol* **275**(2), 269–294.

29. Oliva B, Bates PA, Querol E, Aviles FX & Sternberg MJ (1998) Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J Mol Biol* **279**(5), 1193–1210.

30. Kunik V & Ofran Y (2013) The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng Des Sel* **26**(10), 599–609.

31. Chaudhury S & Gray JJ (2008) Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J Mol Biol* **381**(4), 1068–1087.

32. Leaver-Fay A, Tyka M, Lewis SM, Lange OF *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545–574.

33. Levinthal C (1969) How to fold graciously. In Mössbaun Spectrosc Biol Syst Proc. University of Illinois Press, volume 24, 22–24.

34. Li Z & Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* **84**, 6611–6615.

35. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953) Equation of State Calculations by Fast Computing Machines. *J Chem Phys* **21**, 1087–1092.

36. Ratnaparkhi GS, Ramachandran S, Udgaonkar JB & Varadarajan R (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **37**(19), 6958–6966.

37. Bradley P, Misura KMS & Baker D (2005) Toward high-resolution *de novo* structure prediction for small proteins. *Science (80-)* **309**, 1868–1871.

38. Sagan C (1980) Cosmos. Random House, New York, 1st edition. ISBN 0-394-50294-9, 365 .

39. Weitzner BD, Kuroda D, Marze N, Xu J & Gray JJ (2014) Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins* **82**(8), 1611–1623.

40. Almagro JC, Teplyakov A, Luo J, Sweet RW *et al.* (2014) Second Antibody Modeling Assessment (AMA-II). *Proteins* **82**(8), 1553–1562.

41. Weitzner B, Dunbrack R & Gray J (2015) The Origin of CDR H3 Structural Diversity. *Structure* **23**(2), 302–311.

42. Chaudhury S, Berrondo M, Weitzner BD, Muthu P *et al.* (2011) Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLOS ONE* **6**(8), e22477.

43. Gray JJ, Moughon S, Wang C, Schueler-Furman O *et al.* (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* **331**(1), 281–299.

44. Sircar A & Gray JJ (2010) SnugDock: paratope structural optimization during antibody–antigen docking compensates for errors in antibody homology models. *PLOS Comput Biol* **6**(1), e1000644.

45. Chaudhury S, Lyskov S & Gray JJ (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**(5), 689–91.

46. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM *et al.* (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLOS ONE* **6**(6), e20161.

47. Baugh EH, Lyskov S, Weitzner BD & Gray JJ (2011) Real-time PyMOL visualization for Rosetta and PyRosetta. *PLOS ONE* **6**(8), e21931.

48. DeLano WL (2002) The PyMOL molecular graphics system. *Schrödinger, LLC* **Version 1.**, http://www.pymol.org.

49. Alford RF, Koehler Leman J, Weitzner BD, Duran AM *et al.* (2015) An integrated framework advancing membrane protein modeling and design. *sub judice* .

50. Porter JR, Weitzner BD & Lange OF (2015) A framework to simplify combined sampling modes in Rosetta. *sub judice* .

51. Bradley P & Baker D (2006) Improved β-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* **65**, 922–929.

52. Buss NA, Henderson SJ, McFarlane M, Shenton JM & de Haan L (2012) Monoclonal antibody therapeutics: history and future. *Curr Opin Pharmacol* **12**(5), 615–622.

53. Reichert JM (2014) Antibodies to watch in 2014. *MAbs* **6**(1), 5–14.

54. Lippow SM & Tidor B (2007) Progress in computational protein design. *Curr Opin Biotechnol* **18**(4), 305–311.

55. Farady CJ, Sellers BD, Jacobson MP & Craik CS (2009) Improving the species cross-reactivity of an antibody using computational design. *Bioorg Med Chem Lett* **19**(14), 3744–3747.

56. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C *et al.* (2006) Affinity enhancement of an *in vivo* matured therapeutic antibody using structure-based computational design. *Protein Sci* **15**(5), 949–960.

57. Barderas R, Desmet J, Timmerman P, Meloen R & Casal JI (2008) Affinity maturation of antibodies assisted by *in silico* modeling. *Proc Natl Acad Sci USA* **105**(26), 9029–9034.

58. Chennamsetty N, Helk B, Voynov V, Kayser V & Trout BL (2009) Aggregation-prone motifs in human immunoglobulin G. *J Mol Biol* **391**(2), 404–413.

59. Miklos AE, Kluwe C, Der BS, Pai S *et al.* (2012) Structure-based design of supercharged, highly thermoresistant antibodies. *Chem Biol* **19**(4), 449–455.

60. Voynov V, Chennamsetty N, Kayser V, Helk B *et al.* (2009) Dynamic fluctuations of protein–carbohydrate interactions promote protein aggregation. *PLOS ONE* **4**(12), e8425.

61. Chennamsetty N, Voynov V, Kayser V, Helk B & Trout BL (2010) Prediction of aggregation prone regions of therapeutic proteins. *J Phys Chem B* **114**(19), 6614–6624.

62. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K *et al.* (2012) Developability index: a rapid *in silico* tool for the screening of antibody aggregation propensity. *J Pharm Sci* **101**(1), 102–115.

63. Kuroda D, Shirai H, Jacobson MP & Nakamura H (2012) Computer-aided antibody design. *Protein Eng Des Sel* **25**(10), 507–521.

64. Chothia C, Gelfand I & Kister A (1998) Structural determinants in the sequences of immunoglobulin variable domain. *J Mol Biol* **278**(2), 457–479.

65. Tramontano A, Chothia C & Lesk AM (1990) Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the V_H domains of immunoglobulins. *J Mol Biol* **215**(1), 175–182.

66. Shirai H, Nakajima N, Higo J, Kidera A & Nakamura H (1998) Conformational sampling of CDR-H3 in antibodies by multicanonical molecular dynamics simulation. *J Mol Biol* **278**(2), 481–496.

67. Furukawa K, Akasako-Furukawa A, Shirai H, Nakamura H & Azuma T (1999) Junctional amino acids determine the maturation pathway of an antibody. *Immunity* **11**(3), 329–338.

68. Bond CJ, Marsters JC & Sidhu SS (2003) Contributions of CDR3 to V_HH domain stability and the design of monobody scaffolds for naive antibody libraries. *J Mol Biol* **332**(3), 643–655.

69. Zemlin M, Klinger M, Link J, Zemlin C *et al.* (2003) Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol* **334**(4), 733–749.

70. Kuroda D, Shirai H, Kobori M & Nakamura H (2009) Systematic classification of CDR-L3 in antibodies: implications of the light chain subtypes and the V_L–V_H interface. *Proteins* **75**(1), 139–146.

71. Sellers BD, Nilmeier JP & Jacobson MP (2010) Antibodies as a model system for comparative model refinement. *Proteins* **78**(11), 2490–2505.

72. Persson H, Ye W, Wernimont A, Adams JJ *et al.* (2013) CDR-H3 diversity is not required for antigen recognition by synthetic antibodies. *J Mol Biol* **425**(4), 803–811.

73. MacCallum RM, Martin AC & Thornton JM (1996) Antibody–antigen interactions: contact analysis and binding site topography. *J Mol Biol* **262**(5), 732–745.

74. Collis AV, Brouwer AP & Martin AC (2003) Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* **325**(2), 337–354.

75. Lee M, Lloyd P, Zhang X, Schallhorn JM *et al.* (2006) Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J Org Chem* **71**(14), 5082–5092.

76. Soga S, Kuroda D, Shirai H, Kobori M & Hirayama N (2010) Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Eng Des Sel* **23**(6), 441–448.

77. Sela-Culang I, Alon S & Ofran Y (2012) A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J Immunol* **189**(10), 4890–4899.

78. Raghunathan G, Smart J, Williams J & Almagro JC (2012) Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J Mol Recognit* **25**(3), 103–113.

79. Ramaraj T, Angel T, Dratz EA, Jesaitis AJ & Mumey B (2012) Antigen–antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim Biophys Acta* **1824**(3), 520–532.

80. Stave JW & Lindpaintner K (2013) Antibody and antigen contact residues define epitope and paratope size and structure. *J Immunol* **191**(3), 1428–1435.

81. Olimpieri PP, Chailyan A, Tramontano A & Marcatili P (2013) Prediction of site-specific interactions in antibody–antigen complexes: the proABC method and server. *Bioinformatics* **29**(18), 2285–2291.

82. Marcatili P, Rosi A & Tramontano A (2008) PIGS: automatic prediction of antibody structures. *Bioinformatics* **24**(17), 1953–1954.

83. Sircar A, Kim ET & Gray JJ (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* **37**(Web server issue), gkp387.

84. Whitelegg NR & Rees AR (2000) WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng* **13**(12), 819–824.

85. Dunbar J, Krawczyk K, Leem J, Baker T *et al.* (2013) SAbDab: the structural antibody database. *Nucleic Acids Res* **42**(D1), D1140–D1146.

86. Dunbar J, Fuchs A, Shi J & Deane CM (2013) ABangle: characterising the V_H–V_L orientation in antibodies. *Protein Eng Des Sel* **26**(10), 611–620.

87. Zhao S, Zhu K, Li J & Friesner Ra (2011) Progress in super long loop prediction. *Proteins* **79**(10), 2920–2935.

88. Mandell DJ, Coutsias EA & Kortemme T (2009) Sub-ångström accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* **6**(8), 551–552.

89. Stein A & Kortemme T (2013) Improvements to robotics-inspired conformational sampling in Rosetta. *PLOS ONE* **8**(5), e63090.

90. Das R (2013) Atomic-accuracy prediction of protein loop structures through an RNA-inspired ansatz. *PLOS ONE* **8**(10), e74830.

91. Chen VB, Arendall 3rd WB, Headd JJ, Keedy DA *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr, Sect D: Biol Crystallogr* **66**(Pt 1), 12–21.

92. Abhinandan KR & Martin AC (2010) Analysis and prediction of V_H/V_L packing in antibodies. *Protein Eng Des Sel* **23**(9), 689–697.

93. McLachlan A (1982) Rapid comparison of protein structures. *Acta Crystallogr, Sect A: Found Crystallogr* **38**(6), 871–873.

94. Martin ACR (2009). ProFit.

95. Whitelegg N & Rees AR (2004) Antibody variable regions: toward a unified modeling method. *Methods Mol Biol* **248**, 51–91.

96. Berman HM, Westbrook J, Feng Z, Gilliland G *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* **28**(1), 235–242.

97. Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**(3), 403–410.

98. Almagro JC & Fransson J (2008) Humanization of antibodies. *Front Biosci* **13**, 1619–1633.

99. Engh RA & Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr, Sect A: Found Crystallogr* **47**(4), 392–400.

100. Nivon LG, Moretti R & Baker D (2013) A pareto-optimal refinement method for protein design scaffolds. *PLOS ONE* **8**(4), e59004.

101. Pettersen EF, Goddard TD, Huang CC, Couch GS *et al.* (2004) UCSF ChimeraâĂŤA visualization system for exploratory research and analysis. *J Comput Chem* **25**(13), 1605–1612.

102. Canutescu AA & Dunbrack Jr RL (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci* **12**(5), 963–972.

103. Rohl CA, Strauss CE, Chivian D & Baker D (2004) Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* **55**(3), 656–677.

104. Ting D, Wang G, Shapovalov M, Mitra R *et al.* (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLOS Comput Biol* **6**(4), e1000763.

105. 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. *Proc Natl Acad Sci USA* **109**(2), 449–453.

106. Chailyan A, Marcatili P, Cirillo D & Tramontano A (2011) Structural repertoire of immunoglobulin λlight chains. *Proteins* **79**(5), 1513–1524.

107. Abhinandan KR & Martin AC (2007) Analyzing the "degree of humanness" of antibody sequences. *J Mol Biol* **369**(3), 852–862.

108. Sircar A, Chaudhury S, Kilambi KP, Berrondo M & Gray JJ (2010) A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13–19. *Proteins* **78**(15), 3115–3123.

109. Davey JA & Chica RA (2012) Multistate approaches in computational protein design. *Protein Sci* **21**(9), 1241–1252.

110. Totrov M & Abagyan R (2008) Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* **18**(2), 178–184.

111. B-Rao C, Subramanian J & Sharma SD (2009) Managing protein flexibility in docking and its applications. *Drug Discov Today* **14**(7–8), 394–400.

112. Lyskov S, Chou FC, Conchúir SO, Der BS *et al.* (2013) Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLOS ONE* **8**(5), e63906.

113. Padlan EA (1994) Anatomy of the antibody molecule. *Mol Immunol* **31**(3), 169–217.

114. Barbas CF, Bain JD, Hoekstra DM & Lerner RA (1992) Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc Natl Acad Sci USA* **89**(10), 4457–4461.

115. Sidhu SS & Fellouse FA (2006) Synthetic therapeutic antibodies. *Nat Chem Biol* **2**(12), 682–688.

116. Fellouse FA, Esaki K, Birtalan S, Raptis D *et al.* (2007) High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol* **373**(4), 924–940.

117. Lequin RM (2005) Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clin Chem* **51**(12), 2415–2418.

118. Lu D, Jimenez X, Zhang H, Bohlen P *et al.* (2002) Selection of high affinity human neutralizing antibodies to VEGFR2 from a large antibody phage display library for antiangiogenesis therapy. *Int J Cancer* **97**(3), 393–399.

119. Shirai H, Ikeda K, Yamashita K, Tsuchiya Y *et al.* (2014) High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. *Proteins* **82**(8), 1624–1635.

120. Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol* **104**(1), 59–107.

121. Mardia K & Jupp P (2000) Directional Statistics. John Wiley and Sons Ltd., 2nd edition. ISBN 0-471-95333-4.

122. Koliasnikov OV, Kiral MO, Grigorenko VG & Egorov AM (2006) Antibody CDR H3 modeling rules: extension for the case of absence of Arg H94 and Asp H101. *J Bioinform Comput Biol* **4**(2), 415–424.

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

124. Crooks GE, Hon G, Chandonia JM & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14**(6), 1188–1190.

125. Finn RD, Mistry J, Tate J, Coggill P *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res* **38**(Database issue), D211–D222.

126. 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**(S1), 147–160.

127. Xu Q & Dunbrack RL (2012) Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics* **28**(21), 2763–2772.

128. Ashburner M, Ball CA, Blake JA, Botstein D *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat Genet* **25**(1), 25–29.

129. Yan J, Pan L, Chen X, Wu L & Zhang M (2010) The structure of the harmonin/sans complex reveals an unexpected interaction mode of the two Usher syndrome proteins. *Proc Natl Acad Sci USA* **107**(9), 4040–4045.

130. Hillier BJ, Christopherson KS, Prehoda KE, Bredt DS & Lim WA (1999) Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science (80-)* **284**(5415), 812–815.

131. Kim S, Grant RA & Sauer RT (2011) Covalent linkage of distinct substrate degrons controls assembly and disassembly of DegP proteolytic cages. *Cell* **145**(1), 67–78.

132. Lee HJ & Zheng JJ (2010) PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun Signal* **8**(1).

133. Wei Q & Dunbrack Jr RL (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLOS ONE* **8**(7), e67863.

134. Sibanda BL & Thornton JM (1985) β-hairpin families in globular proteins. *Nature* **316**(6024), 170–174.

135. Sibanda BL, Blundell TL & Thornton JM (1989) Conformation of β-hairpins in protein structures. *J Mol Biol* **206**(4), 759–777.

136. Tramontano A & Lesk AM (1992) Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins* **13**(3), 231–245.

137. van Vlijmen HW & Karplus M (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol* **267**(4), 975–1001.

138. Michalsky E, Goede A & Preissner R (2003) Loops In Proteins (LIP)–a comprehensive loop database for homology modelling. *Protein Eng* **16**(12), 979–985.

139. Choi Y & Deane CM (2010) FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins* **78**(6), 1431–1440.

140. Tyka MD, Jung K & Baker D (2012) Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *J Comput Chem* **33**(31), 2483–2491.

141. Holtby D, Li SC & Li M (2013) LoopWeaver: loop modeling by the weighted scaling of verified proteins. *J Comput Biol* **20**(3), 212–223.

142. Wang G, Dunbrack R L J & Dunbrack Jr RL (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33**(Web Server issue), W94–W98.

143. Reczko M, Martin AC, Bohr H & Suhai S (1995) Prediction of hypervariable CDR-H3 loop structures in antibodies. *Protein Eng* **8**(4), 389–395.

144. Morea V, Lesk AM & Tramontano A (2000) Antibody modeling: implications for engineering and design. *Methods* **20**(3), 267–279.

145. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R *et al.* (2013) Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* **523**, 109–143.

146. Wickham H (2009) Ggplot2 : elegant graphics for data analysis. Springer, New York. ISBN 9780387981406 (softcover acid-free paper) 0387981403 (softcover acid-free paper), 212 .

147. R Development Core Team (2010). R: A Language and Environment for Statistical Computing.

148. Bujotzek A, Dunbar J, Lipsmeier F, Schäfer W *et al.* (2015) Prediction of V_H–V_L domain orientation for antibody variable domain modeling. *Proteins* , doi: 10.1002/prot.24756.

149. Breiman L (2001) Random Forests. *Mach Learn* **45**(1), 5–32.

150. Zhu K & Day T (2013) *Ab initio* structure prediction of the antibody hypervariable H3 loop. *Proteins* **81**(6), 1081–1089.

151. Zhu K, Day T, Warshaviak D, Murrett C *et al.* (2014) Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins* **82**(8), 1646–1655.

152. Chirikjian GS (2011) Modeling loop entropy. *Methods Enzymol* **487**, 99–132.

153. Adolf-Bryfogle J, Xu Q, North B, Lehmann A & Dunbrack RL (2014) PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res* **43**(D1), D432–D438.

154. Wang C, Bradley P & Baker D (2007) Protein–protein docking with backbone flexibility. *J Mol Biol* **373**(2), 503–519.

155. Ramachandran G, Ramakrishnan C & Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**(1), 95–99.

156. Coutsias EA, Seok C, Jacobson MP & Dill KA (2004) A kinematic view of loop closure. *J Comput Chem* **25**(4), 510–528.

157. Coutsias EA, Seok C, Wester MJ & Dill KA (2006) Resultants and loop closure. *Int J Quantum Chem* **106**(1), 176–189.

158. Conway P, Tyka MD, DiMaio F, Konerding DE & Baker D (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci* **23**(1), 47–55.

159. Schneider EL, Lee MS, Baharuddin A, Goetz DH *et al.* (2012) A reverse binding motif that contributes to specific protease inhibition by antibodies. *J Mol Biol* **415**(4), 699–715.

160. Honegger A & Pluckthun A (2001) Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* **309**(3), 657–670.

161. Uysal H, Bockermann R, Nandakumar KS, Sehnert B *et al.* (2009) Structure and pathogenicity of antibodies specific for citrullinated collagen type II in experimental arthritis. *J Exp Med* **206**(2), 449–462.

162. Nakasako M, Takahashi H, Shimba N, Shimada I & Arata Y (1999) The pH-dependent structural variation of complementarity-determining region H3 in the crystal structures of the F_V fragment from an anti-dansyl monoclonal antibody. *J Mol Biol* **291**(1), 117–134.

163. Kilambi KP & Gray JJ (2012) Rapid calculation of protein pKa values using Rosetta. *Biophys J* **103**(3), 587–595.

164. Staelens S, Hadders MA, Vauterin S, Platteau C *et al.* (2006) Paratope determination of the antithrombotic antibody 82D6A3 based on the crystal structure of its complex with the von Willebrand factor A3-domain. *J Biol Chem* **281**(4), 2225–2231.

165. Méndez R, Leplae R, Lensink MF & Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins* **60**(2), 150–169.

166. Becker RS & Knight KL (1990) Somatic diversification of immunoglobulin heavy chain VDJ genes: Evidence for somatic gene conversion in rabbits. *Cell* **63**(5), 987–997.

167. Maizels N (1989) Might gene conversion be the mechanism of somatic hypermutation of mammalian immunoglobulin genes? *Trends Genet* **5**, 4–8.

168. Darlow JM & Stott DI (2006) Gene conversion in human rearranged immunoglobulin genes. *Immunogenetics* **58**, 511–522.

169. Reynaud CA, Anquez V, Dahan A & Weill JC (1985) A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell* **40**(2), 283–291.

170. Reynaud CA, Anquez V, Grimal H & Weill JC (1987) A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**(3), 379–388.

171. Reynaud CA, Dahan A, Anquez V & Weill JC (1989) Somatic hyperconversion diversifies the single V_H gene of the chicken with a high incidence in the D region. *Cell* **59**(1), 171–183.

172. Thompson CB & Neiman PE (1987) Somatic diversification of the chicken immunoglobulin light chain gene is limited to the rearranged variable gene segment. *Cell* **48**(3), 369–378.

173. Wang F, Ekiert DC, Ahmad I, Yu W *et al.* (2013) Reshaping antibody diversity. *Cell* **153**(6), 1379–1393.

174. Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G *et al.* (1993) Naturally occurring antibodies devoid of light chains. *Nature* **363**(6428), 446–448.

175. Greenberg AS, Avila D, Hughes M, Hughes A *et al.* (1995) A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* **374**(6518), 168–173.

176. Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**(10), 1425–36.

177. Lewis PN, Momany FA & Scheraga HA (1973) Chain reversals in proteins. *Biochim Biophys Acta - Protein Struct* **303**(2), 211–229.

178. Richardson JS (1981) The anatomy and taxonomy of protein structure, volume 34 of *Advances in Protein Chemistry*. Academic Press. ISBN 9780120342341, 167–339 .

179. Rose GD, Young WB & Gierasch LM (1983) Interior turns in globular proteins. *Nature* **304**(5927), 654–657.

180. Wilmot C & Thornton J (1988) Analysis and prediction of the different types of β-turn in proteins. *J Mol Biol* **203**(1), 221–232.

181. Hutchinson EG & Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci* **3**(12), 2207–16.

182. Moore G (1965) Cramming More Components Onto Integrated Circuits. *Proc IEEE* **86**(1), 82–85.

183. Stone JE, Gohara D & Shi G (2010) OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems. *Comput Sci Eng* **12**(3), 66–72.

184. Nickolls J, Buck I, Garland M & Skadron K (2008) Scalable parallel programming with CUDA. *Queue* **6**(2), 40–53.

185. Phillips JC, Braun R, Wang W, Gumbart J *et al.* (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* **26**(16), 1781–1802.

186. Stone JE, Phillips JC, Freddolino PL, Hardy DJ *et al.* (2007) Accelerating molecular modeling applications with graphics processors. *J Comput Chem* **28**(16), 2618–2640.

CURRICULUM VITAE

BRIAN D. WEITZNER

DEPARTMENT OF CHEMICAL & BIOMOLECULAR ENGINEERING
JOHNS HOPKINS UNIVERSITY
3400 N. CHARLES STREET, BALTIMORE, MARYLAND 21218
BRIAN.WEITZNER@JHU.EDU

EDUCATION

The Johns Hopkins University — Baltimore, MD
Ph.D. Chemical & Biomolecular Eng. — *Spring 2015 (Expected)*

Cornell University — Ithaca, NY
B.S. Chemical & Biomolecular Eng. Minor Biomedical Eng. — *2009*

RESEARCH EXPERIENCE

Graduate Research Assistant, The Johns Hopkins University — 2009–2015
Advisor: Dr. Jeffrey J. Gray
Topic: Computational modeling of antibodies and antibody–antigen complexes

Undergraduate Research Assistant, Cornell University — 2006–2009
Advisors: Dr. Matthew P. DeLisa & Dr. Jeffrey D. Varner
Topic: Computational retargeting of an E3 ubiquitin ligase

Undergraduate Research Assistant, Fox Chase Cancer Center — 2005–2009
Advisor: Dr. Roland L. Dunbrack, Jr.
Topic: Dimerization motifs of cytosolic sulfotransferases

Howard Hughes Student Scientist, Fox Chase Cancer Center — 2004–2005
Advisor: Dr. Roland L. Dunbrack, Jr.
Topic: Agreement among automated quaternary structure prediction methods

PUBLICATIONS

9. Porter JR, **Weitzner BD**, Lange OF (2015) "A framework simplifying combined sampling modes in Rosetta," *sub judice*

8. Alford RF, Koehler Leman J, **Weitzner BD**, Duran AM, Tilley D, Elazar A, Gray JJ (2015) "An integrated framework for computational modeling and design of membrane proteins," *sub judice*

7. **Weitzner BD**, Dunbrack RL, Jr, Gray JJ (2015) "The origin of CDR H3 structural diversity," *Structure* 23(2), 302–11.

6. **Weitzner BD***, Kuroda D*, Marze N, Xu J, Gray JJ (2014) "Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization," *Proteins* 82(8), 1611–23. (* equal contribution authors)

5. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, Kuroda D, Xu J, **Weitzner BD**, Renfrew PD, Sripakdeevong P, Borgo B, Havranek JJ, Kuhlman B, Kortemme T, Bonneau R, Gray JJ, Das R (2013) "Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE)," *PLOS ONE* 8(5): e63906.

4. Baugh EH, Lyskov S, **Weitzner BD**, Gray JJ (2011) "Real-time PyMOL visualization for Rosetta and PyRosetta," *PLOS ONE* 6(8): e21931.

3. Chaudhury S, Berrondo M, **Weitzner BD**, Muthu P, Bergman H, Gray JJ (2011) "Benchmarking and analysis of protein docking performance in Rosetta v3.2," *PLOS ONE* 6(8): e22477.

2. Bourne PE, Beran B, Bi C, Bluhm W, Dunbrack R, Prlic A, Quinn G, Rose P, Shah R, Tao W, **Weitzner B**, Yukich, B (2010) "Will Widgets and Semantic Tagging Change Computational Biology?" *PLoS Comput. Biol.* 6(2): e1000673.

1. **Weitzner B**, Meehan T, Xu Q, Dunbrack R (2009) "An unusually small dimer interface is observed in all available crystal structures of cytosolic sulfotransferases," *Proteins.* 75(2), 1097–134.

SELECTED HONORS AND AWARDS

Rosetta Service Award: Instructor at inaugural Rosetta Boot Camp	2013
Rosetta Service Award: Leader of transition of Rosetta source code to a new version control system	2013
American Institute of Chemists Student Award	2009

1st place in national AIChE Car Competition; first team to ever perform perfectly	2008
Howard Hughes Medical Institute Student Scientist Program; Fox Chase Cancer Center	2004–2005
Eagle Scout	2003

INVITED SEMINARS AND TALKS

3. Weitzner BD, Gray JJ (2014) "Next-generation Antibody Modeling" *Seminar, Center for Biomolecular Structure and Dynamics, University of Montana*, Missoula, MT.

2. Weitzner BD, Dunbrack RL, Gray JJ (2014) "The origin of CDR H3 Structural Diversity" *Vortrag, Fakultät für Chemie, Technische Universität München*, Munich, Germany.

1. Weitzner BD, Gray JJ (2013) "Computational Structure Prediction, Docking and Design of Antibodies" *IBC Antibody Engineering and Therapeutics*, Huntington Beach, CA. (delivered on behalf of JJG during his paternity leave)

SCIENTIFIC TALKS

6. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ (2013) "Benchmarking RosettaAntibody: Antibody Modeling Assessment II" *Antibody Engineering and Therapeutics Conference*, Huntington Beach, CA.

5. Weitzner BD, Roland RL, Gray JJ (2013) "Kinked CDR H3-like loops are common" *AIChE Annual Conference*, San Francisco, CA

4. Weitzner BD, Dunbrack RL, Gray JJ (2013) "Antibodies are proteins too!" *Rosetta Conference*, Leavenworth, WA.

3. Lyskov S, Weitzner BD, Gray JJ (2011) "PyRosetta 2.0: I can make a new score term in 6 lines!" *Rosetta Conference*, Leavenworth, WA.

2. Weitzner BD, Leaver-Fay A, Kulp D, Lyskov S (2010) "Using PyRosetta for research" *Rosetta Conference*, Leavenworth, WA. [workshop]

1. Weitzner BD, Baugh EH, Gray JJ (2010) "PyMOL–PyRosetta Integration" *Rosetta Conference*, Leavenworth, WA.

SCIENTIFIC POSTERS

3. Weitzner, BD, Dunbrack RL, Gray JJ (2014) "CDR H3 loop prediction" *Rosetta Conference*, Leavenworth, WA.

2. Weitzner, BD, Dunbrack RL, Gray JJ (2012) "Are CDR H3 loops special?" *Rosetta Conference*, Leavenworth, WA.

1. Weitzner, BD, Dunbrack RL, Gray JJ (2011) "Accessing the conformation space of long CDR H3 loops through β-turn detection" *Rosetta Conference*, Leavenworth, WA.

TEACHING EXPERIENCE

Guest Lecturer, ChemBE 414/614 — Fall 2014
Computational Protein Structure Prediction and Design — JHU

Co-Instructor, Rosetta Boot Camp — Spring 2013
An intense week-long crash course to developing in Rosetta — Chapel Hill, NC

Co-Instructor, ChemBE 418 — Fall 2011
Projects in the Design of a Chemical Car — JHU

Teaching Assistant, ChemBE 409 — Fall 2010
Modeling, Dynamics & Control of Chem. & Biol. Systems — JHU

Teaching Assistant, ChemBE 414/614 — Spring 2010
Computational Protein Structure Prediction and Design — JHU

Teaching Assistant, ChemE 3900 — Spring 2009
Chemical Kinetics and Reactor Design — Cornell University

Teaching Assistant, ChemE 1120 — Fall 2008
Introduction to Chemical Engineering — Cornell University

SCIENTIFIC LEADERSHIP

Developed RosettaCon Code of Conduct — 2014
Led the development and implementation of Code of Conduct for RosettaCon to promote diversity. Served on the inaugural incident reporting panel.

Undergraduate research mentor — 2012–2014
Mentor to four undergraduate researchers in the Gray Lab

Organizer, Rosetta Developer Meeting — 2012
The annual Rosetta Developer Meeting was held in Baltimore. Arranged the program, travel, lodging and meals

ACTIVITIES AND OUTREACH

Student volunteer, STEM Achievement in Baltimore Elem. Schools — 2013–2014
Monthly visits to a Baltimore City elementary school to facilitate learning STEM skills

Runner, Baltimore Marathon — 2012

Member of the Extreme Rosetta Workshop (XRW) Team 2010–2011
A small team of developers gathered to overhaul the structure of the Rosetta source code

Volunteer at the Ricky Myers Day of Service Fall 2010
A city-wide day of service to clean parks, plant gardens, repair homes and more

Member of the JHU ChemBE Department STEM outreach group 2009–2013
Visits to a Baltimore Recreation Center to teach children about STEM through demonstrations and activities

Captain of the AIChE Car Team, Cornell University 2008–2009
Leader of the team, organized various sub-groups and kept the project on schedule

Member of the AIChE Car Team, Cornell University 2006–2008
A project team that builds a shoe box-sized car that is powered and stopped by chemical reactions