

**BUILDING COMPUTATIONAL TOOLS FOR ANTIBODY
MODELING AND PROTEIN-PROTEIN DOCKING**

by

Nicholas A. Marze

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

Baltimore, Maryland

June, 2017

© Nicholas A. Marze 2017

All rights reserved

Abstract

Protein-protein interactions underlie countless biological functions, the nature of which is determined by the structure of the protein complex. Computational modeling is an important resource for evaluating protein complexes, with tools like RosettaDock offering structural insights in a high-throughput and cost-efficient manner. Antibodies provide an interesting test case for computational protein-protein docking protocols; they are a highly homologous class of protein that naturally bind an enormous range of antigenic proteins. In this dissertation, I describe new computational methods I developed to model both antibodies and protein-protein complexes, as well as evaluations I made of their performance.

I begin with my additions to the RosettaAntibody protocol, which were motivated by community-wide shortcomings in antibody homology modeling revealed by the Second Antibody Modeling Assessment (AMA-II). I first built the Light-Heavy Orientational Coordinates (LHOC) framework to unambiguously describe the poorly defined antibody V_L - V_H orientation; I then developed the multiple-template grafting protocol, which leverages the LHOC framework to correctly model the V_L - V_H

ABSTRACT

orientation in a majority of antibody targets, tripling the accuracy of the previous RosettaAntibody version.

Seeing the guidance the AMA-II provided toward improving RosettaAntibody, I participated in several rounds of the Critical Assessment of Prediction of Interactions (CAPRI) to better understand the extant deficiencies of the RosettaDock protocol. CAPRI revealed a number of weaknesses in the protocol, including an inability to fully sample anisotropic proteins. I corrected this shortcoming in my novel Ellipsoidal Dock method, with which I correctly modeled two challenging CAPRI targets. More broadly, all protein-protein docking methods fared poorly on CAPRI targets with binding-induced conformational changes and/or large surface areas to search.

Addressing these difficult docking problems requires significantly more extensive conformational sampling protocols. So that such protocols remain computationally feasible, I developed Motif Dock Score (MDS) to rapidly evaluate the expanded pools of candidate structures. With no additional runtime, MDS provides three times the near-native enrichment and nine times the near-native discrimination as the low-resolution RosettaDock mode it replaces.

In summary, I built computational tools that improve the fidelity of antibody homology modeling and broaden the scope of protein-protein docking. Additionally, my contributions to the RosettaDock protocol set the stage for the next-generation of computational docking protocols.

ABSTRACT

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

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

External Reader: Professor Margaret E. Johnson (Department of Biophysics)

Acknowledgments

I owe thanks to many people who helped me along my journey through grad school, but none were more critical to my academic success than my advisor, Jeff Gray. I came to Johns Hopkins with no intention of working with Jeff, but I was so impressed by his enthusiasm for his research and by how much his students enjoyed working for him that I decided I needed to join his lab; that was one of the best decisions I ever made. While working with Jeff, I have learned to be a better writer, a better critical thinker, and a better scientist. Jeff has been outstanding in his role as a research advisor. Jeff always gave me plenty of freedom to learn new techniques and to approach the research problems in my own way, but whenever I wandered too far into the details, he helped me refocus my efforts; whenever my projects stagnated, he suggested new paths forward; and when I was floundering in my second year, he sat me down for a difficult talk that I needed to hear to get back on track. Most importantly, though, Jeff has created a nurturing environment which has made my almost six years in the lab a true pleasure. Through our weekly group meetings and lab lunches, our Rosetta conferences and our lab outings, there exists a culture of

ACKNOWLEDGMENTS

collaboration and mutual respect. Everyone is friendly; everyone is willing to help others; everyone wants everyone else in the lab to succeed. A culture like this is rare in academia or anywhere, but it is nothing more than a reflection of Jeff himself.

The research projects described in this thesis could not have been completed without the work of several of my labmates, as well as a number of external collaborators. Brian Weitzner, Daisuke Kuroda, and Jianqing Xu were instrumental to teaching me about antibodies and how to model them within Rosetta; without Brian's H3 modeling, Daisuke's databases and scripts, and Jianqing's protocol refactoring, my V_L - V_H orientation predictions would have no utility. I should also thank the organizers of the AMA-II, Juan Almagro and his colleagues, for helping to reveal the community-wide inability to model V_L - V_H orientation. Though most Gray Lab members have done some work on the CAPRI competitions, I am most indebted to Krishna Kilambi, from whom I inherited the project, and who showed me how to make and submit our predictions, Jeli azko Jeli azkov and Shourya Sonkar Roy Burman, who assisted me in running several CAPRI targets, and Justin Porter, who wrote the CAPRI scoring scripts and taught me how to use them. I also need to acknowledge Scott Boyken, who ran the interface water predictions for Targets 104 & 105, and the CAPRI organizers who make the competition possible, Marc Lensink, Sameer Velankar, and Shoshana Wodak. Shourya has also been critical to the Motif Dock Score project, building the Adaptive Ensemble Dock protocol and helping to prepare the benchmark. Additionally, MDS could not have been created without the ingenious underlying protocol,

ACKNOWLEDGMENTS

RPX score, designed by Will Sheffler. There are two other people without whom none of these projects could have been completed: Sergey Lyskov, a brilliant software engineer who has written key pieces of code all across Rosetta, and who maintains the benchmark server that preserves Rosetta's stability, and Matt Mulqueen, the Gray Lab sysadmin who makes sure our jazz cluster (and all of our other tech) keeps running smoothly. Lastly, none of my work would be possible without the Rosetta-Commons and their collective contributions to the Rosetta codebase that provide the foundation to all of my protocols.

I must also thank the rest of my thesis committee, Professors Rebecca Schulman, Marc Donohue, Maggie Johnson, and Feilim Mac Gabhann for taking the time to review my research and offer their unique perspectives. Furthermore, my work would not be possible without the funding provided by the National Institutes of Health grant R01 GM078221 awarded to my advisor, as well as the funding provided via my collaborator Alex Miklos at Excet, Inc. I would also be remiss if I did not thank the Department of Chemical & Biomolecular Engineering staff for their behind-the-scenes help with everything from booking conference rooms to processing my equipment orders, especially Caroline Qualls, Kailey Dille, Beth Rannie, Cynthia Fortner, and Lucy Raybon.

My six years in the Gray Lab were some of the best of my life, largely because I was always made to feel welcome by my amazing co-workers, who I also consider friends. They have helped me along from my very first day in the lab, when Brian spent an

ACKNOWLEDGMENTS

hour painstakingly walking me through the commands to navigate through a UNIX terminal. This would become something of a trend over the next five years; whenever I couldn't get Rosetta to download, or fix the compilation errors in XCode, or figure out the right git commands, et cetera, et cetera, Brian always had the solution. He would come to my office, solve my problem, and then we'd chat awhile about the last Planet Money podcast he'd listened to, or whatever other topic came to mind. Brian also always had a different take on my research projects, offering perspectives and approaches I would never have considered. In my first year, it was also Jianqing, my first mentor, and Daisuke who showed me the ropes, passing along the mountains of literature and the coding books I would need to read through to gain my first foothold in the lab.

Though Krishna was pretty much my sole office-mate for three years, I don't think I said more than a few words to him over the first two. A big part of this was the fact that I was still on a conventional work schedule, and Krishna was on what became known as "Krishna time," with work hours from about 4PM to God-Knows-When at night. Eventually, my hours drifted later and later, and Krishna and I got to talking more, often about differences between India and the US, be it political systems, education, or, especially, sports; we would compare cricket to baseball, discuss association vs. gridiron football, or we would talk about his favorite sport, tennis. Krishna was also the docking guru, and taught me everything I know about the protocol. In my third year, my other long-time office mate arrived: Jeliako, or Jeli for

ACKNOWLEDGMENTS

short. Contrary to my laconic early years with Krishna, Jeli was a frequent contributor to the office conversation (I had been warned by the previous Biophysics rotation student, Henry Lessen, that Jeli was "gregarious"). Many of our conversations were scientific in nature, both in regards to our collaborations on the CAPRI and antibody projects, and to our solo projects, me asking for feedback on MDS, and him showing me iteration after iteration of his Hfq FloppyTail models. Just as important to the research help Jeli and I have given each other is the time we've spent distracting one another. Countless Fridays were wasted competing to solve the FiveThirtyEight Riddlers. Many hours were spent trash talking each other over our fantasy football teams (I won the league, by the way, in a resounding victory over Jeli in the finals). Many more were lost talking about obscure music, or Liverpool, or any number of other topics. Aside from Jeli and Krishna, my time in Maryland 303 was marked by the endless parade of rotation students, undergrads, and other temporary residents taking their turn at the empty desks, never staying more than a few months, but always adding their own character to the office: Xiyao Long (appropriately, the longest-tenured), Henry, Morgan Nance, Will Proffitt, Max Klein, Nikhil Shah, Sofia Bali, Kavan Reddy, Naireeta Biswas, Rahel Frick, and several others.

A few years into my graduate work, two new members joined the lab: Rebecca Alford and Julia Koehler Leman, both of them brilliant scientists who always had constructive critiques on my projects. Though they only sat two rooms away, I didn't really get to know Rebecca and Julia until our MiniCon and RosettaCon trips. It

ACKNOWLEDGMENTS

seemed like I was always traveling with one or both of them, to San Francisco, to Boulder, to Burlington, to San Diego, and to Seattle several times, and we had plenty of time waiting for those flights to sit and chat, often about the internal politics of the RosettaCommons or the ChemBE department. It was also at these conferences that I was able to catch up with Sergey after he moved to Montana, and to hear his opinions on Russia or the latest video games.

What best epitomizes my time in the Gray Lab, however, is our lab Dungeons & Dragons group. Biweekly, for almost five years, we have been undertaking a campaign brilliantly designed by our DM: glycobiochemist-extraordinaire, extreme hiker, and amateur-linguist Jason Labonte. Six months into the campaign, my character, the ever-lawful cleric Hakam, joined a ragtag band comprised of Belvin, a wild elf druid with a penchant for fire and beheadings (played by Brian), Mythlos, a treasure-obsessed moon elf with a magical heirloom sword, Leokas, an elf ranger out to avenge his parents' deaths (played by my long-time colleague Mike Pacella), and Jayce, an impossibly charismatic bard (played by lab undergrad Craig Bohrsen). Over the years, as this party has pursued the thaumaturge Samber all across Faerûn, old characters have left and new ones have joined, including Szordrin, a devious wizard and rogue (played by Shourya), Kytharrah, a playful and naïve minotaur (played by Master's student Joey Lubin), and Solisar, an intellectual elf wizard with an interest in giants (played by Jason's brother Jared). Our D&D campaign has not only been a longstanding source of entertainment, but it has also created a true camaraderie

ACKNOWLEDGMENTS

within the Gray Lab.

During my thesis, I have also been impacted by my peers within the ChemBE department. The other students in my cohort provided a great support system early in my graduate work: Abdul Majeed Mohammed, Brad Rupp, Smitha Krishnan, Ivie Aifuwa, Qianru Jin, Markela Ibo, Ran Lin, Hye Rin Kwag, and the rest. My colleagues on the GSLC helped maintain a friendly and fun environment within the department, and I am particularly thankful for my co-chairs on the recruitment committee: Sebastian Barreto, Charles Dhong, and Allison Chambliss.

Part of my graduate success is due to the fantastic living situation I had for most of my Ph.D. From when I moved to Baltimore in 2011 through a few months ago, I lived in a quaint townhouse at 832 West 33rd Street, just a 15 minute walk from my lab. I was blessed with fantastic roommates, first Abdul and Brad, then Abdul and John Zenk (and for a month, Angelo Cangialosi: a Schulman Lab trio!), and later Abdul and Nico Perez. In particular, living with Abdul for nearly six years as we were both going through the Ph.D. program allowed us to forge a lasting friendship.

I also owe credit to my friends from the University of Delaware who kept me from becoming a hermit. First and foremost, I need to thank my Double T Diner crew: Jenny Silbaugh, Sharon Bennett, and Malcolm White. Hearing you complain about your jobs at our dinners helped me realize how great my job was. Next, I need to thank all of my board gaming friends: Jenny & Sean Silbaugh, Sharon & Andy Bennett, Malcolm, Kevin Graney, Jeff Shulenburg, Karen & Vadim Rodin, and

ACKNOWLEDGMENTS

everyone else. Our game nights are tons of fun, and I enjoy winning pretty much every game we play. Lastly, I need to thank my fellow UD season ticket holders: Kevin, Jenny & Sean, and Katie Eiermann & Matt "Dish" Disharoon. I love our tailgates and I'm looking forward to this season; it's always great to watch the marching band, and maybe the football team will finally be good this year.

Finally, I want to thank my family for all of the love and support they've given me throughout my life. I thank my Grandma, Frances Marze, for her incredible kindness and devotion to family. I thank my Oma & Pake, Itie and Bill Ytsma, for modeling the integrity and work ethic that helped make me successful. I thank my siblings, Paige, Anthony, and Hannah, for all the life lessons I learned growing up with you, and for the strong relationships we maintain as adults. I thank my father, Andrew, for providing me with a wonderful life, and for his sage advice, and also for our shared interest in Pittsburgh sports. Most of all, I thank my mother, Donita, for always looking out for me, for her constant willingness to help, and for our daily conversations that have helped get me through the last six years.

To Mom & Dad

Contents

Abstract	ii
Acknowledgments	v
List of Tables	xviii
List of Figures	xix
1 Introduction	1
1.1 Rosetta	3
1.1.1 RosettaDock	4
1.1.2 Challenges in Protein–Protein Docking	5
1.2 Antibodies	6
1.2.1 RosettaAntibody	9
1.2.2 Challenges in Antibody Homology Modeling	9
1.3 Goals and Outline of Thesis	10

2	Improved Prediction of Antibody V_L-V_H Orientation	13
2.1	Overview	13
2.2	Introduction	14
2.3	Materials and Methods	18
2.3.1	Orientational Coordinates Framework Calculation	18
2.3.2	Orientational Coordinate Distance Measurement	19
2.3.3	RosettaAntibody Command Lines	19
2.3.4	Preparation of Antibody Database Set	20
2.3.5	Preparation of Antibody Benchmark Sets	21
2.4	Results	22
2.4.1	A New V_L - V_H Coordinate Frame	22
2.4.2	V_L - V_H Orientation Prediction in Rosetta	31
2.5	Discussion	44
3	Modeling Oblong Proteins and Water-Mediated Interfaces with RosettaDock in CAPRI Rounds 28–35	47
3.1	Overview	47
3.2	Introduction	49
3.3	Methods and Results	51
3.3.1	Successes	53
3.3.1.1	Targets 96/97: α rep-GFP	53
3.3.1.2	Targets 104/105: DNase-Immunity protein	56

3.3.2	Failures	61
3.3.2.1	Target 95: nucleosome-Bmi1/Ring1b-UbcH5c	61
3.3.2.2	Targets 98–101: UCH-L5(\pm Ub)–[RPN13 or INO80G]	62
3.3.2.3	Target 102 (107): HxuA–Hemopexin	63
3.3.2.4	Target 103: UBE2Z–FAT10	65
3.3.2.5	Target 59: Rps28b–Edc3	65
3.4	Discussion	66
4	Motif Dock Score (MDS): A fast, accurate coarse-grained score function for flexible-backbone docking	69
4.1	Overview	69
4.2	Introduction	70
4.3	Methods	73
4.3.1	PDB Curation	73
4.3.2	Motif Querying	73
4.3.3	Score Grid Generation	74
4.3.4	Motif Dock Score	74
4.3.5	Benchmark Set Generation	75
4.3.6	Generation of Backbone Ensembles	76
4.3.7	Benchmark Evaluation Metrics	76
4.4	Results	77
4.4.1	Optimization of Motif Dock Score	77

4.4.2	Benchmarking Motif Dock Score's Performance	90
4.5	Discussion	100
5	Conclusion	103
5.1	My Contributions	103
5.2	Future Directions	106
5.2.1	Antibody Homology Modeling	106
5.2.1.1	V_L - V_H Orientation	106
5.2.1.2	CDR Prediction	109
5.2.1.3	Antibody-Antigen Docking	110
5.2.2	Protein-Protein Docking	111
5.2.2.1	Multi-Body Docking	111
5.2.2.2	Flexible Docking	114
5.2.2.3	Global Docking	115
5.2.3	External Limitations to Protein Structure Prediction	116
	Bibliography	119
	Vita	134

List of Tables

2.1	Performance of ST, MT, bMT, and rMT RosettaAntibody	38
2.2	Performance of ST and MT RosettaAntibody vs. ABangle	44
3.1	Summary of my results in CAPRI rounds 28-35	52
4.1	Enrichment performance of MDS clash terms	89
4.2	Bootstrapped docking metrics	90
4.3	Comparison of docking methods	102

List of Figures

1.1	Structure of an IgG antibody	8
2.1	Orientational Coordinate (LHOC) definition	23
2.2	Two antibodies with identical packing angles	24
2.3	Database distribution of four LHOC metrics	26
2.4	Correlation between LHOC metrics	27
2.5	Visualization of 1.0 OCD and 2.0 OCD	29
2.6	Correlation of OCD and RMSD	30
2.7	Biasing by grafted model V_L - V_H orientation	32
2.8	Flow chart for the RosettaAntibody protocol	33
2.9	Relative performance of MT and ST protocols	35
2.10	Benchmark performance of ST, MT, bMT, and rMT protocols	36
2.11	Relative performance of RosettaAntibody and ABangle	43
3.1	Description of Ellipsoidal Dock	54
3.2	Results of CAPRI rounds 28–35	57
4.1	Effect of smoothing factor and population method on enrichment	80
4.2	Score stratification at small bin size	81
4.3	Effect of bin size on enrichment	82
4.4	Effect of motif score threshold, single-body score terms on enrichment	85
4.5	Effect of motif-generating score function on enrichment	87
4.6	Motif Dock Score lacks an inherent clash penalty	88
4.7	MDS performance vs. centroid mode, target 1LFD	91
4.8	MDS performance vs. hybrid method, target 3F1P	94
4.9	Results of four docking methods, rigid-body targets	95
4.10	Results of four docking methods, medium targets	96
4.11	Results of four docking methods, difficult targets	97
4.12	MDS global docking performance	99
5.1	Description of Universal FoldTree	113

Chapter 1

Introduction

Proteins are an incredibly important class of molecules, integral to nearly every biological function. Their ubiquity and utility are results of their extreme diversity, itself a result of the combinatorial polymeric nature of proteins. Fundamentally, a protein is a linear chain of amino acid residues, the overwhelming majority belonging to a set of twenty "canonical" L-amino acids, linked together by peptide bonds; this sequence of amino acids defines the primary structure of the protein. Local hydrogen-bonding and other non-covalent interactions between the "backbone" atoms of the amino acids ($\text{NH-C}_\alpha\text{H-CO}$) contort the protein into consistent, repeating structural motifs, the most common two being α -helices and β -sheets; these local motifs define the secondary structure of the protein. The protein is further folded into a more condensed three-dimensional structure, mediated largely by non-covalent interactions between the variant "side chain" atoms of the amino acids, as well as

CHAPTER 1. INTRODUCTION

by the entropic gains by packing the hydrophobic regions of the protein away from the aqueous solvent; these higher-order topologies define the tertiary structure of the protein. The three levels of protein structure, taken together, form a highly complex three-dimensional conformation. A protein's structure determines the interactions it can and will make; it also defines the nature of those interactions, including their geometry, strength, and persistence. It is these structure-based interactions that determine every protein function. As such, methods that can predict protein structure or protein interactions are critically important to understanding biological processes.

A number of experimental methods exist for protein structure determination, including x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy. While experimental methods can produce accurate protein structures, they also require sizable pure samples of the protein of interest, utilize expensive instruments and reagents, need to be reoptimized for each protein target, and have a high potential for failure. Computational modeling protocols provide an alternative to experimental methods in the determination of protein structures. Though computational methods have their own limitations, particularly in the accuracy of their structural models, they are often faster, cheaper, and more general than experimental methods.

1.1 Rosetta

Rosetta is a leading computational tool for protein structure prediction and design.^{12,16,24,28,40,63} It contains hundreds of protocols, but most follow the same guiding principles. Rosetta is fundamentally based on a two-stage algorithm: a Monte Carlo plus minimization (MCM) conformational and/or sequence search to generate candidate structures, followed by energy evaluation with a multi-term score function to discriminate the most stable candidate structures.^{5,50}

The MCM search proceeds as a trajectory of consecutive perturbations of the system degrees of freedom (translation, rotation, amino acid mutation, etc.), known as "moves." The energy of the conformation is minimized over the system degrees of freedom, and the pre-move conformation is compared to that of the post-move conformation. If the move lowers the energy, the move is accepted, and the trajectory proceeds. If the move increases the energy, the Metropolis criterion is applied as follows: The energy gap between the two conformations, ΔE , is converted to a Boltzmann probability by the equation $P = \exp(-\Delta E/kT)$, where k is the Boltzmann constant and T is the simulated temperature of the system. The probability is compared to a random number between 0 and 1, with the move being accepted if the probability exceeds the random number, and rejected otherwise, reverting the state to the pre-move conformation. This criterion has the effect of accepting a large fraction of moves which increase the conformational energy only slightly, and accepting a small fraction of moves which increase the conformational energy greatly, allowing

CHAPTER 1. INTRODUCTION

the MCM search to climb out of deeper local minima in the energy landscape and more thoroughly sample the conformational space.

The Rosetta energy function consists of several score terms, broadly divided into two types: physics-based energies and statistical potentials.⁵ The physics-based energies derive their contributions to the score from approximations of physical laws; these include a Lennard-Jones Van der Waals energy, a Lazaridus-Karplus implicit solvation energy,⁴⁸ and a Coulombic electrostatic potential with a distance-dependent dielectric constant. The statistical potentials are instead derived from feature distributions in known crystal structures, with the energy contributions optimized for recapitulation of these feature distributions. The Rosetta energy function contains statistical potential terms that capture features like hydrogen bonding, backbone and side chain torsional angles, and disulfide bonds. The score terms are combined using empirically derived weight sets to give a single score for each candidate structure evaluated. Lower Rosetta scores correspond to lower free energies, and thus, the lowest-scoring candidate structure corresponds to the most stable (and likely the native) conformation.

1.1.1 RosettaDock

The RosettaDock algorithm is one of the most fundamental protocols in the Rosetta suite, used for the prediction of protein complexes.³³ It consists of three phases: initial placement, low-resolution search, and high-resolution refinement. In

CHAPTER 1. INTRODUCTION

the initial placement phase, the two docking partners are placed randomly in a putative conformation. In the low-resolution search, the protein side chains are modeled with a ball-and-stick "centroid" representation, and the complex is sampled using moderately aggressive rigid-body moves. In the high-resolution phase, the protein side chains are fully represented, and the best candidate structures generated by the low-resolution phase are refined using small rigid-body moves, and repacking of the protein side chains. RosettaDock also contains alternate modes to capture more degrees of freedom, including EnsembleDock,²¹ which simulates the conformer selection model of complex formation, and SnugDock,⁷¹ which simulates the induced-fit model.

1.1.2 Challenges in Protein–Protein Docking

The basic protein–protein docking case, in which two known rigid protein partners bind at an interface with high shape complementarity, is a tractable problem that many computational protocols can solve. When complications are added to this base case, however, the docking problem quickly becomes difficult.⁵³ Protein flexibility is a particularly difficult complication, with motions as small as 1-2 Å RMSD or as simple as a single loop remodeling confounding most docking protocols.^{47,53,80} The inability for docking protocols to effectively handle different conformations during docking also increases the difficulty when the docking partners are not structurally characterized or homologous to known protein structures.⁵¹ Another common complication arises when the binding site cannot be identified by bioinformatics or characterization data,

requiring the use of a global docking search. Though some methods, particularly those that are fast-Fourier-transform-based (FFT-based),^{26,65} can discriminate binding sites with high shape complementarity in a global search, all methods perform poorly when the shape complementarity is low, such as in many antibody–antigen complexes, or when a compounding factor is added, such as protein flexibility. Global docking also becomes more difficult as the protein partners grow larger due to the increased conformational space that needs to be sampling.

1.2 Antibodies

Antibodies are a biologically important subclass of proteins that play a key role in the immune system of higher animals. Antibodies possess hypervariable antigen-binding sites that allow them to bind with high specificity to nearly any foreign "antigenic" protein, peptide, or small molecule, as well as signaling domains that allow the antibody–antigen complex to be sequestered and destroyed, neutralizing the antigenic molecule. The hypervariability of antibodies arises via a few methods. First, the three genes that comprise the antigen-binding site exist in multiple variants in the genome; through several recombination events, collectively known as V(D)J recombination, random variants of each of these genes are spliced together, allowing for approximately 10^6 antigen-binding sites through gene combinatorics alone.⁴ The actual diversity created by splicing is a few orders of magnitude larger due to vari-

CHAPTER 1. INTRODUCTION

able splicing at the junctional sites, causing the deletion of an imprecise number of nucleotides from the VDJ genes and potentially changing the reading frame of the D gene.⁴ These germline antibodies grow even more diverse during the maturation phase, where error-prone polymerases target regions of the antigen-binding site, causing hypermutation of key binding residues during expansion of antibody-producing cells.⁴ In total, the diversity of antibodies is extraordinarily large, estimated at approximately 10^{11} structures in a typical human, although the theoretical diversity drastically exceeds even this figure.⁴

Antibodies have highly homologous structures; an example is shown in Figure 1.1. In nature, they consist of two identical heavy chains, usually made up of four immunoglobulin domains, and two identical light chains, made up of two immunoglobulin domains, linked together with disulfide bonds in a "Y" shape. The largest structural variations occur in the two variable (F_V) regions, which are located at the end of the arms of the Y. The F_V consists of a heavy chain variable domain (V_H) and a light chain variable domain (V_L), each with a well-conserved immunoglobulin beta-sandwich topology, or framework. The F_V also contains the antigen-binding site, comprised primarily of six loops, three on the V_H and three on the V_L , known as complementarity-determining regions (CDRs). Five of these loops (H1, H2, L1, L2, and L3) consistently form a small set of canonical structures strongly determined by their sequence, while the sixth (H3) has a highly variable structure. The H3 loop includes the entire VDJ junction region, and it is also the region most susceptible to

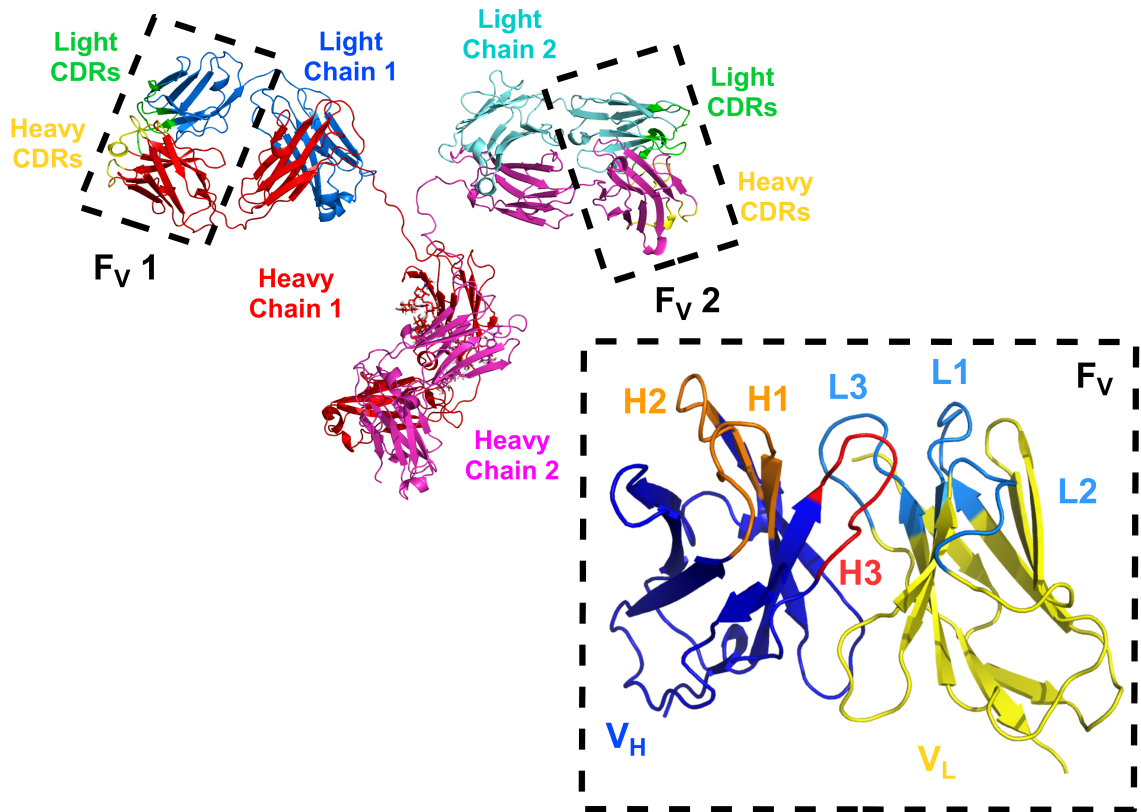


Figure 1.1: Structure of an IgG antibody. Complete structure (top left) is comprised of two heavy chains (red & pink) and two light chains (blue & cyan). The antigen-binding sites are located on the two F_V regions, boxed in black and shown in detail in the bottom right. The F_V is comprised of the V_H (blue) and V_L (yellow) domains. The antigen-contacting residues largely reside in the six CDR loops: L1, L2, & L3 on the V_L (cyan), and the H1 & H2 (orange) and the hypervariable H3 (red) on the V_H.

somatic hypermutation, making it by far the most diverse antibody region in both sequence and length, and thus also in structure.

1.2.1 RosettaAntibody

RosettaAntibody was developed in 2009⁷² to apply Rosetta principles to the prediction of antibody structure. RosettaAntibody uses a template-based homology modeling protocol followed by targeted structural refinement of difficult-to-predict regions of the antibody. Sequences for nine structural regions of the F_V (six CDRs, two domain frameworks, full F_V region) are queried against an antibody structural database, and the best matches are selected. The CDR templates are pasted onto the appropriate domain frameworks, and the resultant domain homology models are aligned to the full F_V template to provide a crude grafted model. The H3 is then completely remodeled, the light and heavy domains are redocked, and the CDR conformations are minimized in the context of the Rosetta score function. Thousands of candidate structures are stochastically generated, and the model with the lowest energy is selected.

1.2.2 Challenges in Antibody Homology Modeling

Because antibodies are highly homologous, it is easier to blindly model a generic antibody than it is to model a generic protein; however, there are some specific

CHAPTER 1. INTRODUCTION

challenges that arise from the extreme sequence diversity of the antibody’s antigen-binding site. In particular, modeling the CDR-H3 is incredibly difficult.⁸⁶ The H3 adopts no canonical conformations, has no conserved sequence, and ranges in length from 3 residues³⁹ to 61 residues.⁸³ The most effective H3 prediction techniques are based on complete remodeling of the H3 loop, but these are heavily limited by loop length becoming ineffective when the H3 is longer than about 14 residues.⁸⁶ Since the H3 is situated at the interface between the V_L and V_H domains, it is also difficult to predict the docked orientation of these two domains.⁶ The other CDRs are easier to predict, but these too can confound antibody-modeling protocols if they adopt a conformation not found in existing structural databases. These databases are almost exclusively comprised of human and murine antibodies, so antibodies derived from other species are more likely to have a rare non-H3 CDR conformation.⁸⁸ All of these difficult-to-predict regions occur at the antibody paratope, and errors in antibody homology modeling degrade the model of the antigen-binding site, making it more difficult to predict antibody-antigen complexes when an antibody homology model is required.

1.3 Goals and Outline of Thesis

In the remainder of this thesis, I describe the efforts I undertook and the progress I made toward improving the field of computational protein modeling, specifically to-

CHAPTER 1. INTRODUCTION

ward the goals of higher-fidelity antibody modeling and more general protein–protein docking. In particular, I describe new computational tools I built to address some of the limitations noted above.

In Chapter 2 (previously published⁵⁶), I describe my work on enhancements to the RosettaAntibody protocol. Motivated by community-wide failures in the prediction of antibody V_L – V_H orientation, I created the LHOC framework to unambiguously define the V_L – V_H orientation, and I built the multiple-template protocol to improve RosettaAntibody’s ability to predict these orientations. Combined, these two enhancements dramatically improve RosettaAntibody’s V_L – V_H orientation prediction fidelity, also surpassing a competing protocol, ABangle.³⁷

In Chapter 3 (previously published⁵⁵), I detail my efforts predicting protein–protein complexes in the CAPRI competition rounds 28-35. The challenges proved successful my novel anisotropic docking method, Ellipsoidal Dock. I also highlight the weaknesses in the field of protein–protein docking I uncovered during the CAPRI rounds, notably in the prediction of large and flexible complexes.

In Chapter 4, I describe my development of the Motif Dock Score (MDS) protocol. The failures in CAPRI made apparent the need for enhanced docking sampling methods, which themselves require a fast, accurate low-resolution sampling method to become feasible on current hardware. I show that MDS satisfies this niche, providing most of the accuracy of the Rosetta full-atom score function with only the overhead of one lookup per residue pair.

CHAPTER 1. INTRODUCTION

In Chapter 5, I summarize my contributions to the field of computational structure prediction, and I note the near-term future improvements that my work enables. I also take stock of some of the remaining challenges in protein structure prediction.

Chapter 2

Improved Prediction of Antibody

V_L-V_H Orientation

This is a pre-copyedited, author-produced version of an article accepted for publication in Protein Engineering Design & Selection following peer review. The version of record <Marze NA, Lyskov S, & Gray JJ, "Improved prediction of antibody VL-VH orientation," *Protein Eng. Des. Sel.* 29(10), 409-418.> is available online at: <https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/gzw013>

2.1 Overview

Antibodies are important immune molecules with high commercial value and therapeutic interest because of their ability to bind diverse antigens. Computational prediction of antibody structure can quickly reveal valuable information about the nature of these antigen-binding interactions, but only if the models are of sufficient quality. To achieve high model quality during complementarity-determining region (CDR)

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

structural prediction, one must account for the V_L - V_H orientation. I developed a novel four-metric V_L - V_H orientation coordinate frame. Additionally, I extended the CDR grafting protocol in RosettaAntibody with a new method that diversifies V_L - V_H orientation by using ten V_L - V_H orientation templates rather than a single one. I tested the multiple-template grafting protocol on two datasets of known antibody crystal structures. During the template-grafting phase, the new protocol improved the fraction of accurate V_L - V_H orientation predictions from only 26% (12/46) to 72% (33/46) of targets. After the full RosettaAntibody protocol, including CDR H3 remodeling and V_L - V_H re-orientation, the new protocol produced more candidate structures with accurate V_L - V_H orientation than the standard protocol in 43/46 targets (93%). The improved ability to predict V_L - V_H orientation will bolster predictions of other parts of the paratope, including the conformation of CDR H3, a grand challenge of antibody homology modeling.

2.2 Introduction

Antibodies are important immune molecules with high commercial value and therapeutic interest because of their ability to bind diverse antigens, from small molecules and short peptides to full-length proteins. Antibodies' binding diversity is a function of their hypervariable FV domains, each consisting of two immunoglobulin domains: V_L and V_H . The antigen-binding site (paratope) is located at six loops near the

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

V_L - V_H interface, known as complementarity-determining regions, or CDRs.

Many structural studies of the F_V have focused on the conformation of the CDRs, particularly CDR H3.^{3,61,83,85,91} Since the CDRs are attached to the framework of the V_L and V_H domains, any change in the relative orientation of the V_L and V_H domains will propagate to change the CDRs' relative orientation, and therefore, the shape of the paratope. Failing to account for the V_L - V_H orientation during CDR or paratope structure prediction dramatically hinders the quality of the output models, and recent evaluation found the V_L - V_H orientation to be a limiting factor in antibody structure prediction.⁸⁸

Abhinandan and Martin¹ were the first to codify a metric for measuring the V_L - V_H orientation. They defined the packing angle as a torsional angle between the primary axes of the V_L and V_H domains. Among the ~ 500 F_V crystal structures they examined, packing angle differed by as much as 30° . Chailyan *et al.*¹⁹ defined V_L - V_H orientation differently, via clustering. The resulting description was limited in scope: only two distinct orientational clusters and a distinct singleton were found; however, a number of key residues were found to correlate with the orientational clusters, indicating that V_L - V_H orientation may be predictable from sequence.

The Second Antibody Modeling Assessment (AMA-II) measured the ability of several computational antibody structural prediction methods to capture native V_L - V_H orientation in a blind prediction challenge. Two metrics were used to evaluate the antibody orientations generated in AMA-II: (1) an analogue to RMSDvariable as

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

described by Sela-Culang *et al.*,⁶⁸ and (2) the tilt angle as described in Almagro *et al.*⁶ While these measures encode more orientational information than the Abhinandan-Martin packing angle, both are pairwise difference metrics rather than absolute ones. A geometrically complete, absolute measure of V_L - V_H orientation, ABangle, was published by Dunbar *et al.*³⁷ ABangle is composed of one torsional angle, four plane angles, and one distance, representing the six degrees of freedom of the two-body V_L - V_H complex. The ABangle measure was applied in a study to predict V_L - V_H orientation. In tests on the AMA-II antibody set, the authors predicted ABangle metrics corresponding to an average RMSD of misorientation of 0.50 Å, performing better than the average competitor (0.63 Å), beating the average in 9 of 11 targets.¹⁷

RosettaAntibody is an application for blind prediction of antibody structure.^{72,88} RosettaAntibody operates in two phases: (1) template selection and grafting, wherein known antibody structure fragments are combined to create a coarse-grained model, and (2) structure refinement, which uses Monte Carlo perturbations with minimization to remodel the CDR H3 loop, refine all CDR loops, and redock the V_L and V_H domains.

Until recently, RosettaAntibody’s efficacy in predicting native V_L - V_H orientations had only been investigated implicitly by measuring RMSD values across all F_V residues. During the Second Antibody Modeling Assessment (AMA-II), RosettaAntibody’s orientation predictions were evaluated explicitly, comparing the packing angles of the Rosetta models to those of their corresponding crystal structures.⁸⁸ RosettaAn-

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

tibody compared favorably in most respects to the competing protocols, producing two sub-Ångstrom H3 models and achieving the best H3 model in four targets. However, V_L - V_H orientation was a weakness, as RosettaAntibody created a structure with sub-Ångstrom cross-domain RMSD for only 5 of 11 targets. V_L - V_H orientation prediction for targets with uncommon packing angles was particularly poor: all three targets with a packing angle more than one standard deviation removed from the database average were predicted incorrectly.

In this paper, I developed a novel four-metric V_L - V_H orientation coordinate frame, which I called Light-Heavy Orientational Coordinates (LHOC). Additionally, I extended the RosettaAntibody protocol with a new method to diversify V_L - V_H orientations by grafting multiple templates. I tested the new RosettaAntibody protocol on two datasets of known antibody crystal structures: a 46-member high-resolution antibody set, and the 11-member AMA-II dataset. I compared the performance of the new RosettaAntibody against the previous version, as well as against the ABangle method for predicting V_L - V_H orientation.

2.3 Materials and Methods

2.3.1 Orientational Coordinates Framework Calculation

The four coordinates used to describe V_L - V_H orientation (α , δ_{ID} , θ_L , and θ_H) are defined from a common framework of four non-atomic points at the V_L - V_H interface (Figure 2.1). Point 2 is located at the center of a conserved pair of beta-strands in the VL framework; it is defined as the centroid of the $C\alpha$ coordinates of residues L35-L38 and L85-L88 using Chothia numbering.³ Point 3 is the VH counterpart to point 2, defined as the centroid of the $C\alpha$ coordinates of residues H36-H39 and H89-H92, Chothia numbering. Point 1 is located nearer the CDRs than point 2, along the first principal component line of the coordinate set used to calculate point 2. Point 4 is the VH counterpart to point 1.

All coordinates were calculated with a Rosetta implementation of the above framework. α is defined in the same manner as Abhinandan and Martin;¹ specifically, it is defined as the dihedral angle between points 1, 2, 3, and 4. δ_{ID} is defined as the distance between points 2 and 3. θ_L is defined as the plane angle between points 1, 2, and 3. θ_H is defined as the plane angle between points 2, 3, and 4.

2.3.2 Orientational Coordinate Distance Measurement

Orientational Coordinate Distance (OCD) is calculated as:

$$OCD = \sum_{i=\{\alpha,\delta_{ID},\theta_L,\theta_H\}} \left(\frac{x_{i,A} - x_{i,B}}{\sigma_{i,DB}} \right)^2$$

where $x_{i,A}$ and $x_{i,B}$ represent the value of LHO metric i of structure A and structure B , respectively, and $\sigma_{i,DB}$ represents the standard deviation of the Gaussian distribution best fit to the database distribution of LHO metric i . The four values for i are α , δ_{ID} , θ_L , and θ_H . OCD is dimensionless.

2.3.3 RosettaAntibody Command Lines

The new MT protocol, part of the Rosetta software package, is available free of charge for academic and non-profit use at www.rosettacommons.org. The code used to generate data in this paper is available starting from release revision 57, deposited May 21, 2015. The MT protocol is currently available on the ROSIE public web server (rosie.graylab.jhu.edu⁵²).

To create the grafted structures, the following command line was used. The `homolog_exclusion` argument should be 99 when performing blind predictions, and 80 when evaluating algorithm performance on a known set.

```
antibody.py --both-chains <FASTA file> --relax
```

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

```
--homolog_exclusion=<99||80>

--multi-template-grafting --number-of-templates 10

--light_heavy-multi-graft

--filter-by-orientational-distance=1

--orientational-distance-cutoff 0.5
```

To create the candidate structures, the following command line was used for each grafted structure. `abH3.flags` is a text file containing the set of option flags for a standard RosettaAntibody run. The `cter_constraint` file is a two-line text file containing two atomic constraints; it is generated automatically by the previous command line. The grafted structure is one of 10 models generated by the previous command line. The `-nstruct` argument should be 1000 for the first grafted structure, and 200 for the other nine models.

```
antibody_H3.linuxgccrelease @abH3.flags

-s <grafted structure, 1 of 10> -nstruct <200||1000>

-constraints:cst_file <cter_constraint file>
```

2.3.4 Preparation of Antibody Database Set

The RosettaAntibody database consists of 1,040 antibody F_V crystal structures culled from the Protein Data Bank using the methods described by Sivasubramanian *et al.*⁷² One outlier antibody (1MCO) has an interdomain distance of 19.6Å, far-

ther removed from the second-largest interdomain distance than the second-largest is from the smallest. This antibody is highly irregular, with the F_{Ab} - F_C hinge region deleted,³⁵ explaining the unnaturally-large interdomain distance; this antibody was consequently removed from analyses of the RosettaAntibody database.

2.3.5 Preparation of Antibody Benchmark Sets

A high-resolution antibody set was compiled from the PyIgClassify database.² A series of restrictions was placed on the structures: a maximum resolution of 2.5 Å, a maximum R value of 0.2, a maximum B-factor of 80.0 Å² for each atom in the structure, an asymmetric unit containing only one copy of the F_V , a CDR H3 loop length between 9 and 20 residues, a human or mouse species tag, and no non-canonical or modified amino acid residues. Additionally, the set was filtered to remove antibodies with identical sequences in any of the heavy-chain CDR loops. Of the resultant 49 structures, three (1X9Q, 2W60, 3IFL) were eliminated because of challenges presented in sequence misalignment or numbering (e.g., 1X9Q is missing highly conserved heavy-chain residues C92 and W103).

The Second Antibody Modeling Assessment (AMA-II) antibody set consists of the 11 antibodies described in Almagro *et al.*⁶

2.4 Results

2.4.1 A New V_L - V_H Coordinate Frame

To describe the geometry of antibody V_L - V_H orientation, I developed a new coordinate frame (Figure 2.1) as an extension of the packing angle described by Abhinandan and Martin.¹ Three vectors compose the Abhinandan–Martin framework: two primary axis vectors, one each drawn through V_L and V_H , and a third vector linking the axis vectors tail-to-tail across the V_L - V_H interface. The Abhinandan–Martin packing angle (α) is defined as the apparent angle between the V_L and V_H vectors as seen when looking down the connecting line from V_H to V_L (Figure 2.1.B). The packing angle metric captures the set of V_L - V_H relative positions in which the V_L and V_H domains twist past each other, broadening or contracting the paratope. Figure 2.2 shows, however, that antibodies with identical α will not necessarily superimpose, and in practice, they often do not. This structural ambiguity is an inherent limitation of the α metric. Therefore, I sought a more complete description of the V_L - V_H orientation.

To capture more of the V_L - V_H orientation degrees of freedom, I repurposed the Abhinandan–Martin packing angle vector framework to define the other metrics: an interdomain distance (δID) and two plane angles, L-opening angle (θL) and H-opening angle (θH). δID as the length of the linking vector (Figure 2.1.C). θL and θH are defined as the plane angle between the linking vector and the V_L and V_H vectors,

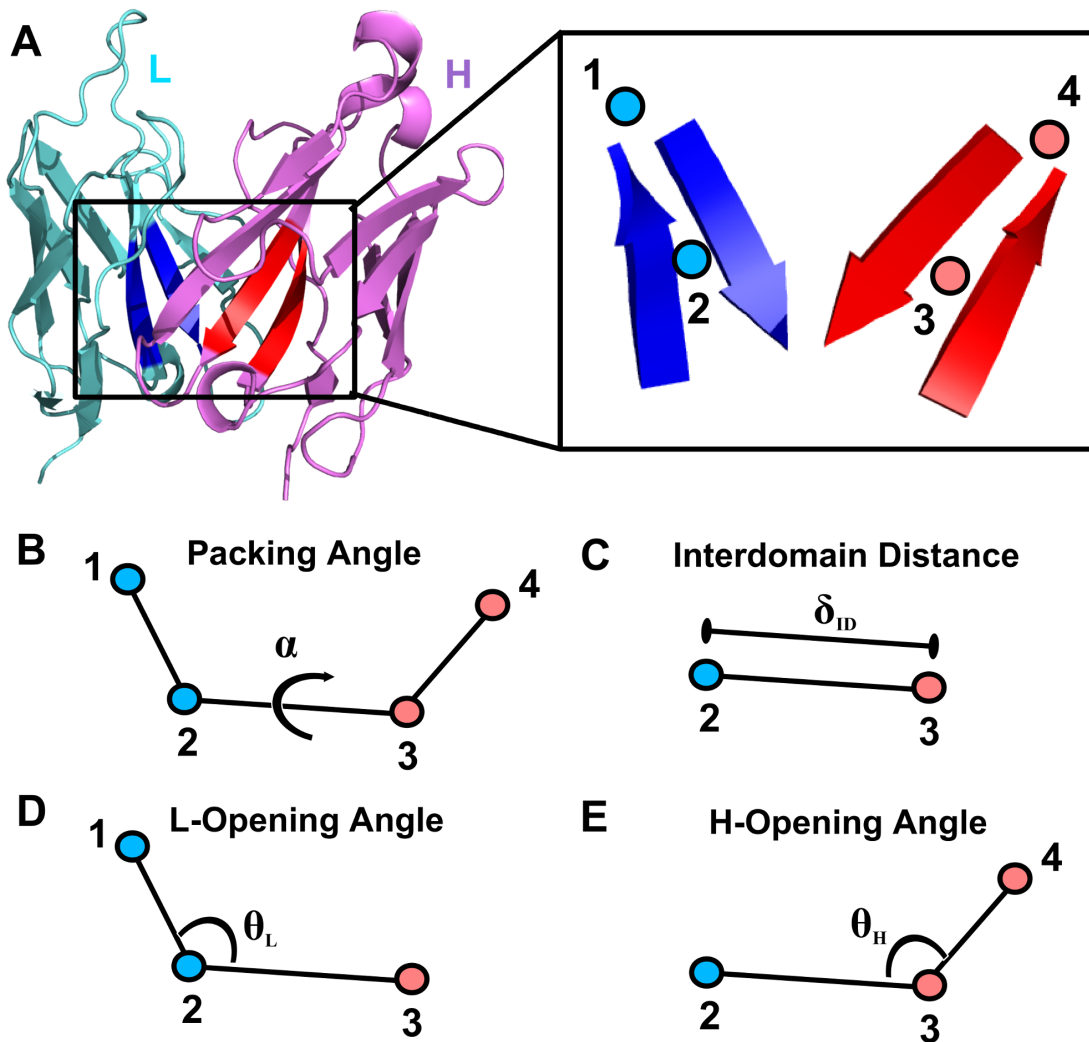


Figure 2.1: Orientational Coordinate (LHOC) definition. (A) F_V structure showing light chain (cyan), heavy chain (pink), and the key beta strands for defining the LHOC framework (Chothia numbering: L35-L38 and L85-L88 in blue and H36-H39 and H89-H92 in red, see 2.3 for details). The inset shows the placement of the four points which form the basis of the LHOC framework. (B) Packing angle, α , is the dihedral angle between points 1, 2, 3, and 4. (C) Interdomain distance, δ_{ID} , is the distance between points 2 and 3. (D) L-opening angle, θ_L , is the plane angle between points 1, 2, and 3. (E) H-opening angle, θ_H , is the plane angle between points 2, 3, and 4.

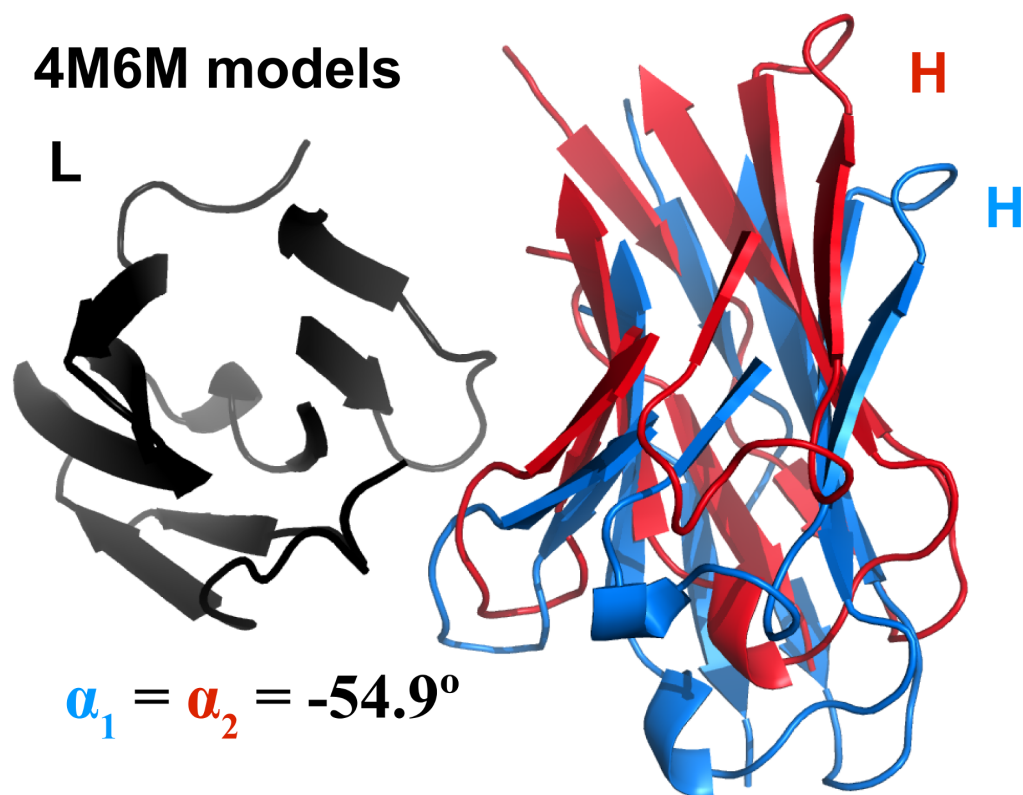


Figure 2.2: Two RosettaAntibody F_V models of AMA-II target 5 (PDB ID 4M6M) with equivalent values of packing angle. Structures have light chains (black) superimposed. Heavy chains are shown in red and blue. CDR residues (Chothia definition) are omitted for clarity.

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

respectively (Figure 2.1.D and 2.1.E). Together, I refer to the four coordinates (α , δ_{ID} , θ_L , and θ_H) as the Light-Heavy Orientational Coordinates (LHOC).

For LHOC to be a non-redundant coordinate frame and more descriptive than the Abhinandan-Martin packing angle, each coordinate must capture some component of V_L - V_H orientational diversity that is sufficiently independent from the components captured by other coordinates. To evaluate the effectiveness of the LHOC coordinate frame, I calculated the LHOC metrics for each antibody in a curated set of 1,040 antibody F_V crystal structures, representing a high- and medium-resolution (≤ 3.5 Å) subset of all antibodies in the Protein Data Bank.

Figure 2.3 shows distributions for each of the four LHOC metrics across all antibodies in the database. All three angle distributions are approximately Gaussian. Consistent with the prior use of packing angle to solely define V_L - V_H orientation,^{1,6} the α distribution is the largest component of diversity in V_L - V_H orientation, with a range of nearly 35° (mean (μ) = -52.3° , standard deviation (σ) = 3.9° , minimum = -70.9° , maximum = -36.7°). The two LHOC plane angle distributions each show a range approximately half as large as the α distribution. The θ_L distribution has a range of about 15° ($\mu = 97.2^\circ$, $\sigma = 1.9^\circ$, min = 89.3° , max = 104.4°), while the θ_H distribution has a range of about 20° ($\mu = 99.4^\circ$, $\sigma = 2.6^\circ$, min = 87.9° , max = 108.1°). The δ_{ID} distribution is also approximately Gaussian, but with a long right tail. While the bulk of the distribution, 1,030 of 1,040 structures, lies between 13.5 Å and 15.5 Å, nine of the ten remaining structures have a δ_{ID} between 15.5 Å and

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

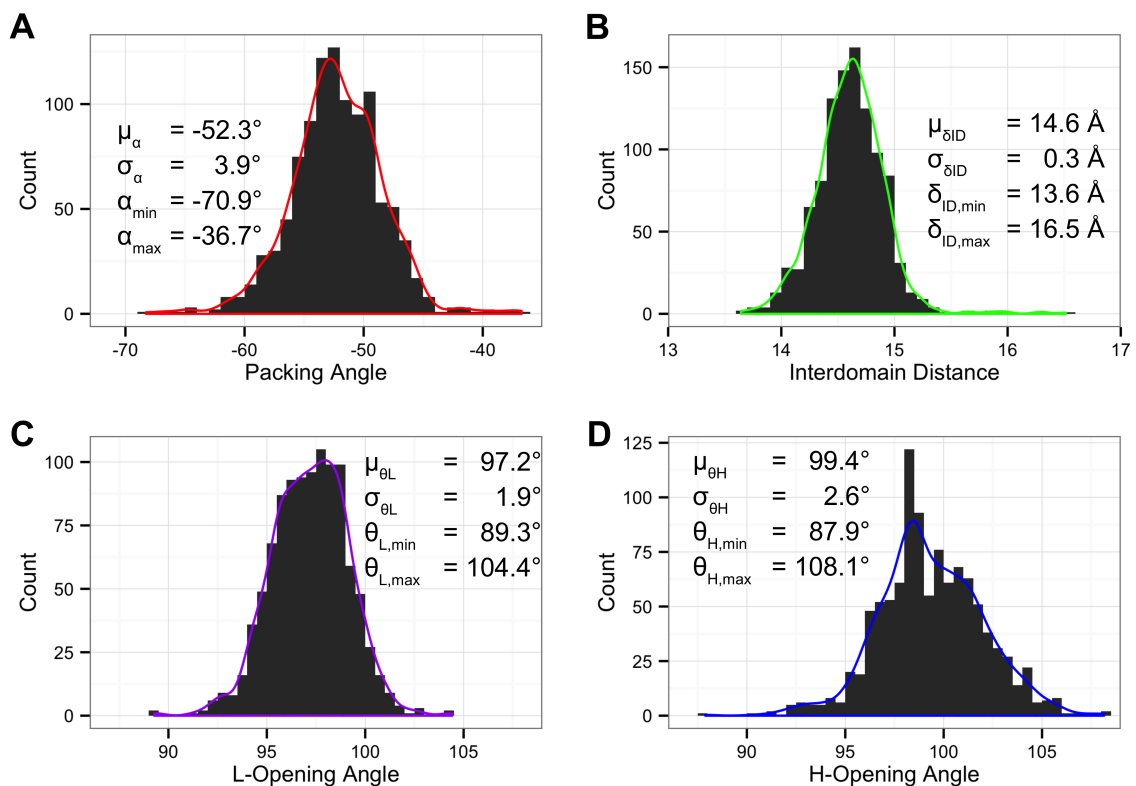


Figure 2.3: Histograms of each of the four LHOC metrics across the 1,040 structures in the Rosetta antibody database. Histogram bin widths are 1° for packing angle, α , (A), 0.1\AA for interdomain distance, δ_{ID} , (B), and 0.5° for plane angles, θ_L and θ_H , (C and D). Kernel density estimates of each distribution are shown as curves over the histograms.

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

16.5 Å.

To test the independence of the four LHOC metrics, I plotted all pairwise distributions of metrics for the database antibodies, shown in Figure 2.4. Five of the

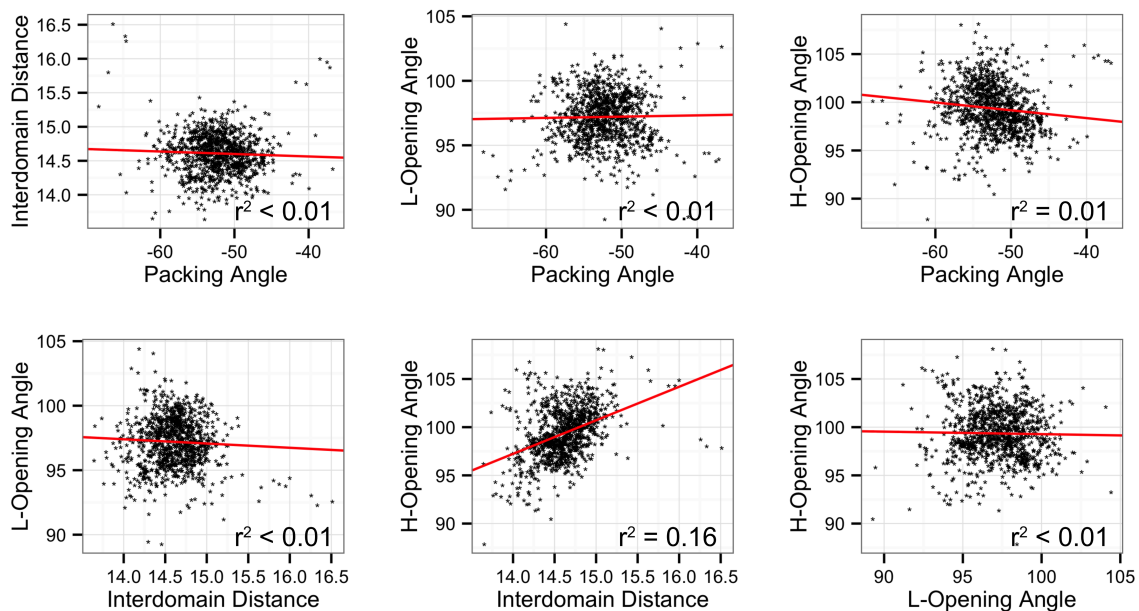


Figure 2.4: Pairwise distributions of each pair of LHOC metrics across the 1,040 structures in the Rosetta antibody database. A best-fit line (red) is drawn through each pairwise distribution, with its corresponding r^2 value labeled.

six pairs of metrics show no correlation ($r^2 \leq 0.01$), with approximately 2D-Gaussian distributions. The remaining pair, θ_H and δ_{ID} , show a small degree of correlation ($r^2 = 0.16$); antibodies with larger-than-average δ_{ID} tend to also have larger-than-average θ_H . Such a correlation could arise because the hinge of the θ_H definition differs from the physical hinge about which the V_L - V_H orientation actually varies between antibodies. If the physical hinge were upstream of the θ_H hinge, a naturally "open" antibody would have both a larger θ_H and a larger δ_{ID} . In this case, one

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

would also expect the antibody to also have a larger θ_L , as it is effectively a mirror image of θ_H ; however, there is no correlation seen between δ_{ID} and θ_L , suggesting that the mathematical and physical hinges are in a similar place. This implies that the correlation between θ_H and δ_{ID} is not due to misplacement of the LHOC framework, nor a redundant selection of coordinates to include in LHOC.

The four-coordinate nature of the LHOC framework allows it to describe more facets of V_L - V_H orientation than α alone, but it requires a combination metric to simplify the difference to one dimension. Therefore, I defined the Orientational Coordinate Distance (OCD) by summing the squared z-score deviations in each of the four LHOC base metrics (See Methods for details). Figure 2.5.A shows that a pair of antibodies with an OCD of 1.0 or less superimpose closely, and Figure 2.5.B shows that a pair of antibodies with an OCD of 2.0 or greater are clearly distinct.

Because changes in the different LHOC metrics exert different lever-arm effects on the antibody domains, and because the contributions to OCD can be dominated by a large variation in one or two LHOC metrics, two antibody pairs with the same OCD will not necessarily have the same RMSD between them. For example, two antibodies with a 3.0 OCD due only to a difference in packing angle will have a much larger RMSD than two antibodies with a 3.0 OCD due only to a difference in interdomain distance. Nonetheless, OCD and RMSD are loosely correlated: as shown in Figure 2.6, two structures with a high OCD tend to have a high RMSD as well. An OCD of 2.0 is roughly equivalent to an RMSD of 1 Å, although most 2.0 OCD

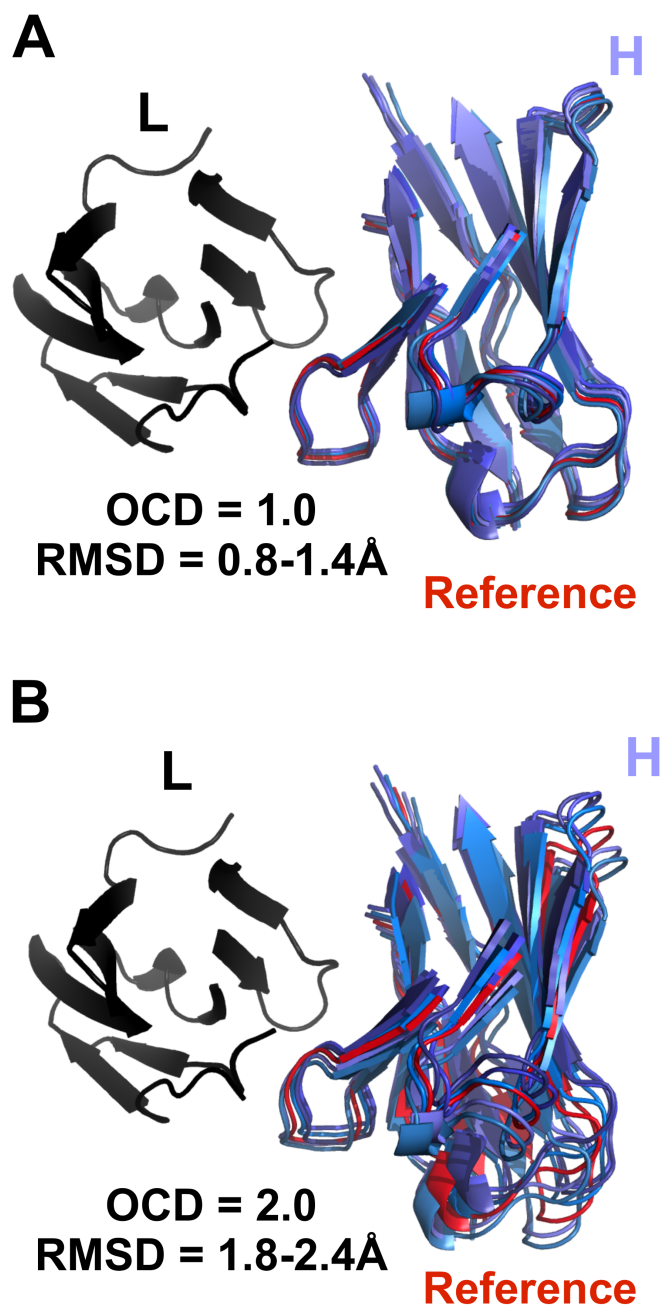


Figure 2.5: Comparison of five RosettaAntibody F_V models (blues) with 1.0 OCD (A) and 2.0 OCD (B) to a reference antibody F_V structure (red). Structures have light chains (black) superimposed. CDR residues (Chothia definition) are omitted for clarity.

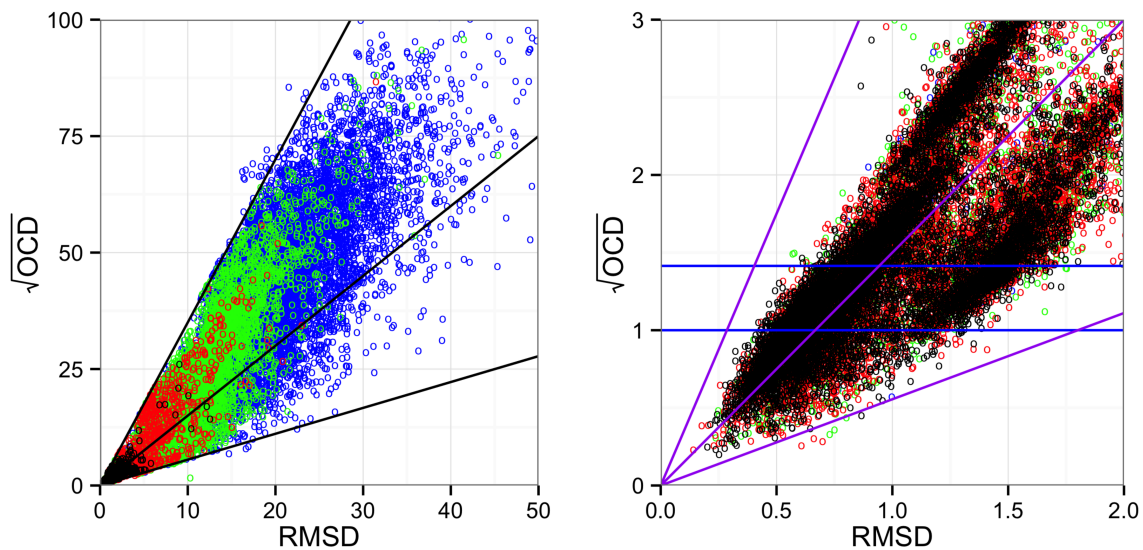


Figure 2.6: Correlation of OCD and RMSD in 40,000 candidate structures from rigid-body V_L - V_H docking of a 1DLF homology model. To linearize the correlation, the square root of OCD is plotted against RMSD. The different colors represent the scale of perturbations in the docking simulations (blue > green > red > black, corresponding to size of rotations and translations). Zero-intercept lines highlight the center and the bounds of the correlation, and their slopes are noted. Horizontal lines corresponding to OCD values of 1.0 and 2.0 are also shown. The data are plotted over the RMSD ranges 0 Å-50 Å (left) and 0 Å-2 Å (right).

structure pairs will have a larger RMSD due to intradomain variations.

2.4.2 V_L - V_H Orientation Prediction in Rosetta

With the OCD metric, I next sought to test the efficacy of RosettaAntibody at predicting correct V_L - V_H orientations. A preliminary examination of the RosettaAntibody candidate structures for one of the AMA-II targets with an incorrect V_L - V_H orientation prediction revealed that a wide range of V_L - V_H orientations were sampled by docking moves during the structure refinement phase – so wide, in fact, that nearly the entire database distribution is spanned in all coordinates. However, the lowest-scoring candidate structures, and thus, the ones selected as final models, had orientations quite similar to the starting point of the refinement trajectories, i.e., the grafted structure. To examine how the starting point biases the output orientations, I launched refinement trajectories from grafted structures with alternate V_L - V_H orientations. Figure 2.7 shows the orientation distributions of candidate structures generated by these runs. In each trajectory, there is a visible well in which low-scoring candidate structures tend to have orientations matching their individual grafted structures rather than converging to the native orientation. These data suggest that the refinement phase of RosettaAntibody has an effective limit on how far it can alter the V_L - V_H orientation. While more orientationally-distant structures can be sampled, these structures do not resemble natural antibodies, as evidenced by their high scores. This behavior is beneficial when the grafted structure has a native V_L - V_H

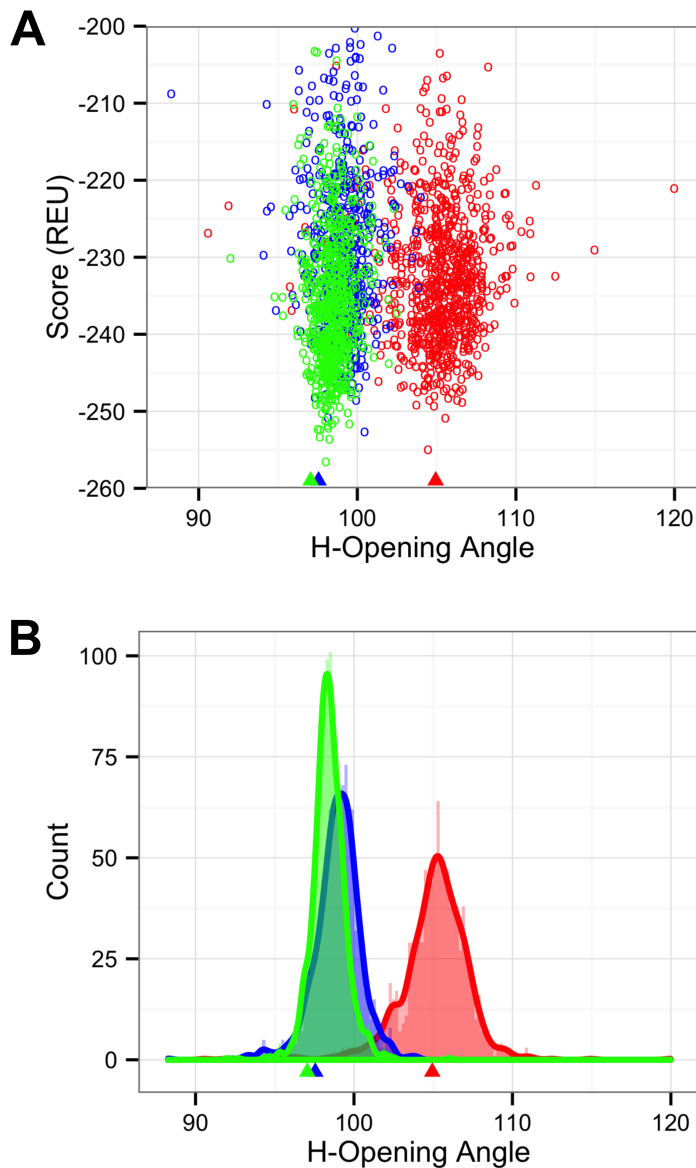


Figure 2.7: H-opening angle, θ_H , distributions among candidate structures generated by RosettaAntibody from three different starting grafted structures with different V_L - V_H orientations. (A) Plots of θ_H vs. score, showing scoring funnels for each of the three runs in a different color, with the grafted structure θ_H marked by a matching-color triangle below the x-axis. (B) Histograms and kernel density estimates for each of the three runs in a different color, with the θ_H of each grafted starting structure marked as in (A).

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

orientation, but in the general case, it indicates an inadequate search.

To attempt to produce low-scoring candidate structures near the native V_L - V_H orientation, I created a new RosettaAntibody grafting protocol that runs several trajectories rather than a single trajectory. A flowchart description of the protocol, called multiple-template grafting (MT), is shown in Figure 2.8 in the context of the

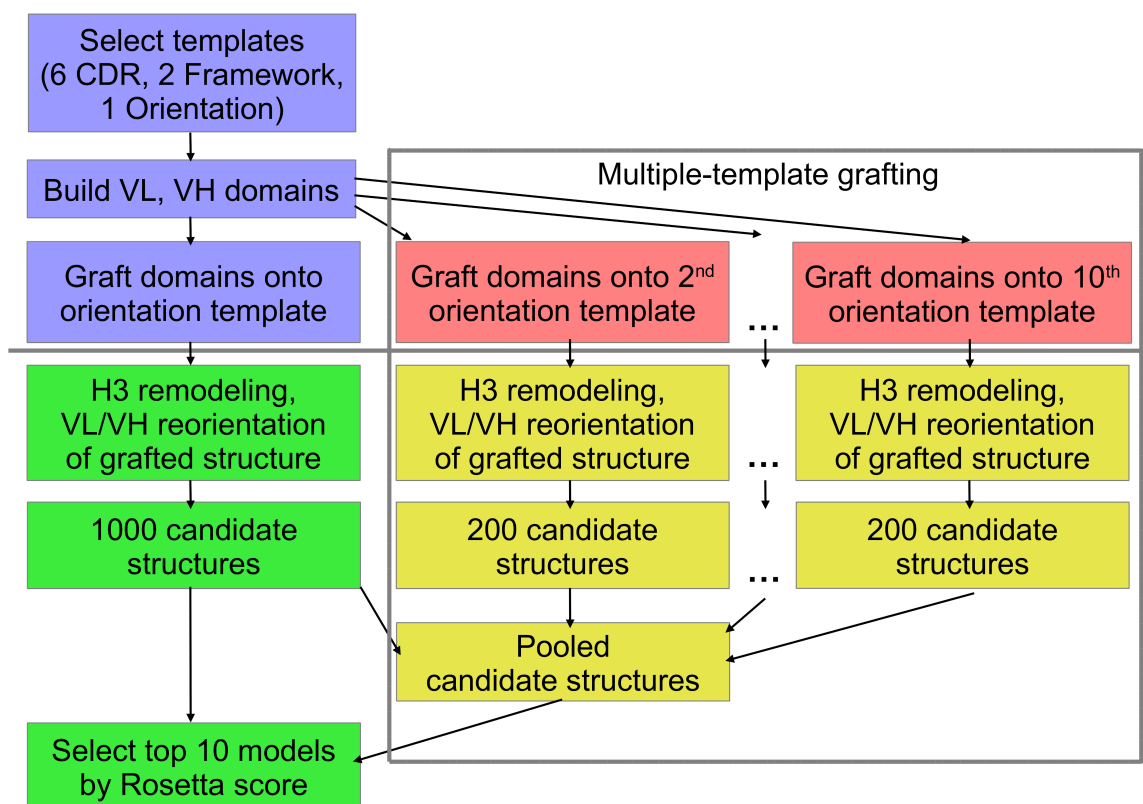


Figure 2.8: Flow chart for the RosettaAntibody protocol. The grafting phase is shown in blue and pink, above the solid gray line, while the refinement phase is shown in green and gold, below the solid gray line. Steps from the standard single-template grafting (ST) protocol are colored in blue and green. New steps added to create the multiple-template grafting (MT) protocol are colored in pink and gold and enclosed in the dashed gray box; the blue/green ST steps are also part of the MT protocol.

previous RosettaAntibody protocol, henceforth described as single-template graft-

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

ing (ST). Instead of creating only a single grafted structure during the first phase of RosettaAntibody, MT creates ten grafted structures from the ten best-matching (by BLAST alignment) V_L - V_H orientation templates. Additionally, to diversify the grafted structures, I enforce a minimum OCD cutoff value of 0.5 between all orientation template pairs, rejecting candidate templates with a lower OCD to any of the ten and replacing them with the next-best BLAST match. The number ten and the 0.5 OCD cutoff were selected to capture a near-native V_L - V_H orientation in all targets in my calibration set, the 11 AMA-II antibodies, while minimizing the number of redundant templates. Each grafted structure is refined in multiple independent RosettaAntibody refinement runs to create a pool of candidate structures: 1000 from the shared ST/MT grafted structure, and 200 each from the remaining nine MT grafted structures.

To evaluate the sampling efficacy of multiple-template grafting, I compared the performance of ST and MT RosettaAntibody on a benchmark set of 46 high-resolution, manually-curated antibody crystal structures from the Protein Data Bank (PDB).¹¹ Figure 2.9 shows the pairwise comparisons of OCD values between the ST and MT predictions for all targets. In the grafting phase of RosettaAntibody, the ST V_L - V_H orientation prediction was within 2.0 OCD of the native in only 26% (12/46) of targets. The MT predictions nearly tripled this, with the best match among the MT predictions within 2.0 OCD of native in 72% (33/46) of targets. Additionally, of the remaining 13 targets, 10 showed an improved OCD to native in their best MT

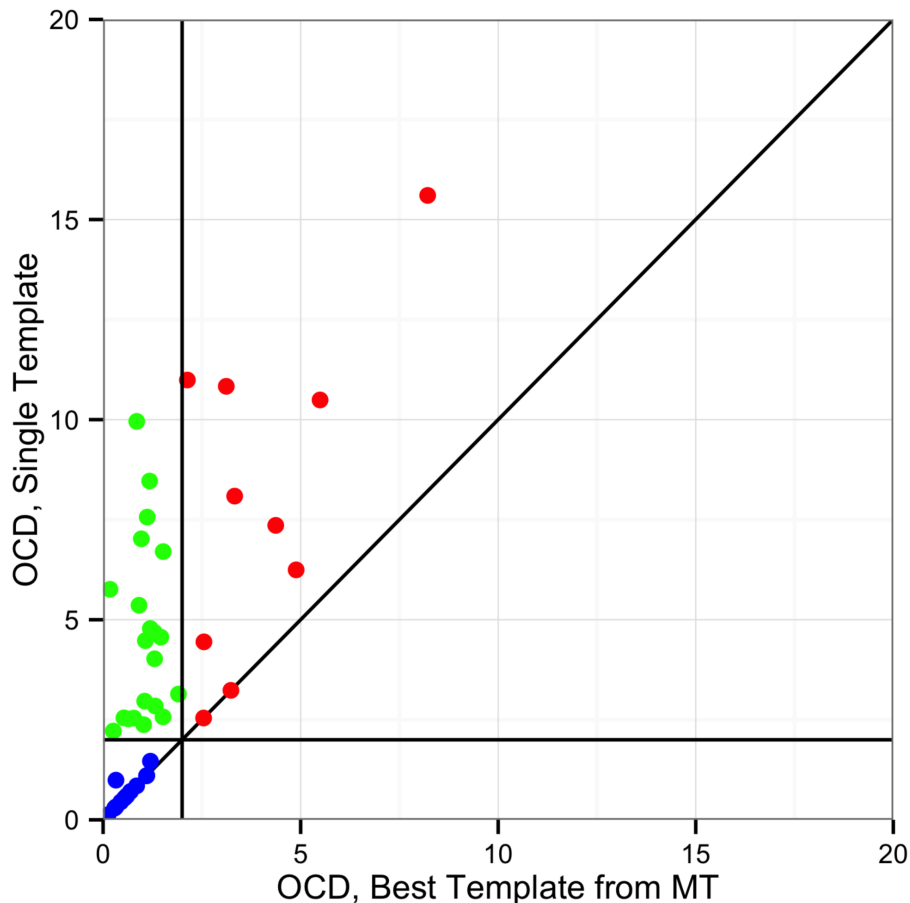


Figure 2.9: Comparison of V_L - V_H orientation prediction performance between MT RosettaAntibody and ST RosettaAntibody after the grafting stage for the 46 members of the benchmark set. The OCD between the native structure and the ST post-grafting stage structure is plotted against the lowest OCD between the native and any of the ten MT post-grafting stage structures. Targets where the best MT structure is the same as the ST structure appear on the $x = y$ line, also plotted. Targets where the best MT structure has a closer OCD to native than the ST structure are above the $x = y$ line. MT success cases ($OCD \leq 2.0$) are found to the left of the vertical $OCD = 2.0$ line, while MT failures ($OCD > 2.0$) are found to the right. Likewise, ST success cases are found below the horizontal $OCD = 2.0$ line, while failures are found above. The green points indicate the 21 targets that improved from a failure case to a success case when using the MT protocol, while blue points indicate the 12 targets that remained successes, and the red points indicate 10 of the 13 targets that remained failures (the other 3 have OCD values exceeding the bounds of the plot).

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

prediction versus the ST prediction.

After the RosettaAntibody refinement phase, including H3 remodeling and V_L - V_H re-orientation, the MT protocol produced more candidate structures within 2.0 OCD of native than the ST protocol in 43 of 46 targets (93%) (Figure 2.10.A). The

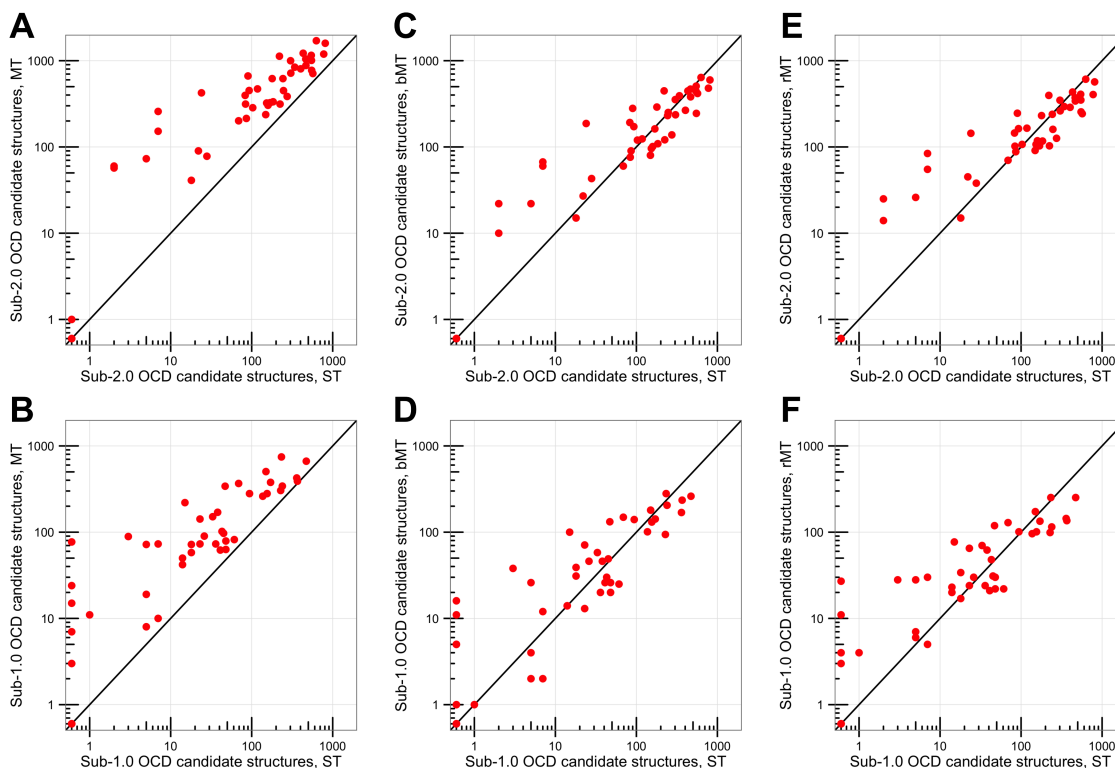


Figure 2.10: Performance of the full ST, MT, bMT, and rMT RosettaAntibody protocols on the 46 benchmark antibodies, showing the number of candidate structures with an OCD value below 2.0 (A, C, E) or below 1.0 (B, D, F) for the ST protocol vs. the MT (A and B), the bMT (C and D), and the rMT (E and F) protocols. The ST, the bMT, and the rMT protocols each include 1000 candidate structures total, while the MT protocol includes 2800 candidate structures.

remaining three targets all had poorly-predicted repertoires of grafted structures, in which none of the ten MT predictions (including the ST prediction) were closer than 15.0 OCD to native (Table 2.1). While the MT protocol generated more cases

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

under 2.0 OCD, it also required more total candidate structures for each target, 2,800 vs. 1,000, at the proportional cost of computing time (\sim 1440 CPU-hrs for the full MT protocol). To evaluate the candidate-structure-equivalent performance of the ST and MT protocols, I compared only the 1,000 lowest-scoring MT candidate structures against the 1,000 ST candidate structures; this is henceforth described as the biased MT (bMT) protocol. Additionally, to more fairly evaluate the time-equivalent performance of the ST and MT protocols, I also pared the output from the MT protocol to 1,000 randomly selected candidate structures per target, maintaining as best as possible the 5:1 ratio of input structures; this is henceforth described as the reduced MT (rMT) protocol.

Table 2.1: Summary of the performance of the full ST, MT, bMT, and rMT RosettaAntibody protocols for each of the 46 benchmark antibodies. Target totals represent the number of targets in the benchmark set for which a given protocol achieved a given threshold of success. The four success thresholds are: (1) a single sub-2.0 OCD grafted structure, (2) a single sub-1.0 OCD candidate structure, (3) at least 100 sub-2.0 OCD candidate structures, and (4) at least 10 sub-1.0 OCD candidate structures. The grafted models for the MT, bMT, and rMT protocols are identical, so the bMT and rMT columns were omitted from this portion of the table.

Target	OCD, grafted model		Lowest OCD candidate structure				Sub-2.0 OCD candidate structures				Sub-1.0 OCD candidate structures			
	ST	MT (best)	ST	MT	bMT	rMT	ST	MT	bMT	rMT	ST	MT	bMT	rMT
1dlf	15.60	8.21	1.09	0.42	0.46	0.42	5	73	22	26	0	15	5	4
1fns	3.14	1.91	0.22	0.08	0.08	0.08	302	999	354	348	47	341	132	119
1gig	2.97	1.05	0.53	0.10	0.10	0.10	103	285	120	107	18	72	39	34
1jfq	2.52	0.64	0.45	0.22	0.24	0.24	93	451	172	163	5	72	26	28
1jpt	5.76	0.17	0.08	0.06	0.06	0.06	809	1593	598	568	472	665	261	252
1mfa	6.25	4.89	0.59	0.52	0.51	0.52	28	78	43	38	5	8	4	6
1mlb	4.48	1.07	0.50	0.10	0.10	0.22	179	620	290	231	23	142	71	65
1mqk	2.84	1.32	0.07	0.04	0.13	0.04	340	841	392	294	94	280	140	101
1nlb	7.02	0.97	0.32	0.12	0.31	0.12	69	201	60	70	14	50	14	23
1oaq	4.78	1.20	0.31	0.10	0.10	0.18	84	314	76	102	23	73	13	24
1seq	0.60	0.60	0.30	0.06	0.06	0.24	304	717	236	262	33	151	58	70
2adf	2.38	1.03	0.29	0.15	0.22	0.15	153	323	96	107	36	73	20	24
2d7t	20.47	15.18	2.93	1.55	2.93	2.29	0	1	0	0	0	0	0	0
2e27	6.70	1.52	1.11	0.68	0.98	1.13	18	41	15	15	0	3	1	0
2fb4	7.56	1.12	1.04	0.11	0.19	0.11	7	152	67	55	0	24	11	11
2fbj	9.96	0.85	1.12	0.10	0.10	0.36	7	258	60	84	0	77	16	27
2r8s	0.99	0.33	0.07	0.07	0.10	0.15	550	773	245	255	228	304	94	99
2v17	10.83	3.12	0.22	0.16	0.16	0.34	83	396	192	145	26	90	46	30
2vxxv	0.45	0.45	0.12	0.12	0.12	0.12	775	1193	480	405	360	425	169	144

Table 2.1 ... continued

Target	OCD, grafted model		Lowest OCD				Sub-2.0 OCD				Sub-1.0 OCD			
	ST	MT (best)	ST	MT	bMT	rMT	ST	MT	bMT	rMT	ST	MT	bMT	rMT
2xwt	0.85	0.85	0.04	0.04	0.04	0.10	467	880	381	341	137	261	101	96
2ypv	8.09	3.33	0.41	0.41	0.82	0.41	22	90	27	45	7	10	2	5
3e8u	4.02	1.30	0.03	0.03	0.03	0.03	432	1220	446	433	150	504	180	173
3eo9	4.45	2.55	0.03	0.03	0.29	0.13	243	620	231	239	43	102	30	48
3g5y	2.57	1.52	0.38	0.10	0.10	0.13	403	805	266	288	38	170	46	62
3giz	5.36	0.91	0.96	0.36	0.36	0.36	2	60	10	25	1	11	1	4
3gnm	2.55	0.53	0.07	0.04	0.04	0.11	220	1127	447	396	69	367	149	129
3go1	2.22	0.26	0.10	0.06	0.06	0.06	169	326	162	103	45	97	49	31
3hc4	10.99	2.13	1.61	0.13	0.47	0.13	2	57	22	14	0	7	5	3
3hnt	4.56	1.46	0.43	0.14	0.14	0.20	90	665	280	246	15	220	100	77
3i9g	0.29	0.29	0.05	0.05	0.05	0.05	569	710	417	243	367	390	235	136
3liz	8.47	1.18	0.38	0.18	0.29	0.18	24	424	187	144	3	89	38	28
3lmj	2.54	0.78	0.63	0.17	0.46	0.37	159	304	101	118	14	42	14	20
3m8o	31.92	15.63	9.06	7.00	10.02	7.00	0	0	0	0	0	0	0	0
3mlr	130.15	130.15	49.42	49.42	50.09	52.15	0	0	0	0	0	0	0	0
3mxxw	1.47	1.20	0.06	0.06	0.07	0.07	547	1008	507	350	239	342	205	115
3nps	7.36	4.37	4.09	2.06	20.6	2.70	0	0	0	0	0	0	0	0
3oz9	2.54	2.54	0.65	0.39	0.86	0.60	86	214	90	88	5	19	2	7
3p0y	0.13	0.13	0.08	0.08	0.22	0.14	274	385	138	126	61	82	25	22
3t65	0.71	0.69	0.08	0.01	0.01	0.08	628	1704	639	610	233	745	280	252
3umt	1.10	1.10	0.25	0.10	0.10	0.29	184	334	109	117	48	79	26	30
3v0w	3.24	3.24	0.16	0.16	0.22	0.16	224	314	121	103	41	62	26	21
4f57	10.49	5.49	0.05	0.05	0.05	0.18	149	237	80	91	48	63	20	22
4h0h	4.69	1.30	0.62	0.20	0.21	0.20	118	471	124	165	7	73	12	30
4h20	0.32	0.32	0.16	0.10	0.13	0.10	465	1046	468	372	155	280	131	101

Table 2.1 ... continued

Target	OCD, grafted model		Lowest OCD candidate structure			Sub-2.0 OCD candidate structures			Sub-1.0 OCD candidate structures					
	ST	MT (best)	ST	MT	bMT	rMT	ST	MT	bMT	rMT	ST	MT	bMT	rMT
4hpy	0.54	0.54	0.32	0.17	0.17	0.30	247	450	251	160	18	58	31	17
4nzu	1.11	1.11	0.03	0.03	0.09	0.03	544	1154	444	408	171	378	142	134
Target Totals	Best Graft \leq 2.0 OCD		Best Structure \leq 1.0 OCD			\geq 100 Sub-2.0 OCD Models			\geq 10 Sub-1.0 OCD Models					
	12	33	37	42	42	41	27	36	29	31	30	39	35	35

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

The bMT protocol produced more sub-2.0 OCD candidate structures for 22 targets, with 20 targets generating fewer sub-2.0 OCD candidate structures than the ST protocol due to dilution effects (Figure 2.10.C). Likewise, the rMT protocol produced more sub-2.0 OCD candidate structures for 20 targets, and fewer sub-2.0 OCD candidate structures for 22 targets (Figure 2.10.E). The remaining four targets had no sub-2.0 OCD candidate structures created by either the ST, bMT, or rMT protocol (Table 2.1). When counting only sub-1.0 OCD structures, those with essentially identical V_L - V_H orientations to the native antibody, the rMT protocol fared better, with 25 targets improving on the ST counts, and only 16 worsening from dilution (Figure 2.10.F). The bMT protocol showed little improvement, bettering the ST counts in 21 targets, falling short of the ST counts in 18 targets, and matching the ST counts in the remaining 3 targets (Figure 2.10.D). Nearly all of the targets with fewer low OCD candidate structures in the rMT and bMT protocols still had at least 100 sub-2.0 OCD and 10 sub-1.0 OCD candidate structures, however, indicating that the dilution effects are largely benign.

I compared the grafting phase of RosettaAntibody, both the old ST protocol and the new MT protocol, against the recently published V_L - V_H orientation predictor, ABangle.¹⁷ The coordinate-by-coordinate ABangle prediction results for the AMA-II antibody set are published, allowing for a direct comparison of the two methods. Four of the ABangle coordinates, HL, dc, LC1, and HC1, are directly analogous to α , δ_{ID} , θ_L , and θ_H , respectively. All are calculated using a similar reference frame

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

centered on the same F_V residues, and the corresponding coordinate pairs populate native distributions of similar size and shape, albeit at different absolute values. By virtue of the similarity of these four ABangle coordinates to the four LHOC metrics, an OCD value can be calculated using the published model-to-native deviations in the four ABangle coordinates corresponding to LHOC.

Of the 11 AMA-II antibody targets, ABangle achieved a sub-2.0 OCD prediction for five. The original RosettaAntibody protocol (ST) performed similarly, shown in Figure 2.11.A, predicting a sub-2.0 OCD structure for four of the 11 targets, and predicting a structure with an OCD better than ABangle for five of the 11 targets. Interestingly, ABangle and ST RosettaAntibody have almost no overlap in their correct predictions, with only one target achieving a sub-2.0 OCD prediction from both methods. When the template with the best OCD of the ten models from the MT grafting prediction was used, however, RosettaAntibody substantially outperformed ABangle, as seen in Figure 2.11.B. RosettaAntibody predicted ten of 11 targets within 2.0 OCD of native, including six targets for which ABangle had made an incorrect prediction. The OCD values for each of the AMA-II antibody targets predicted by RosettaAntibody ST, RosettaAntibody MT (best prediction only), and ABangle, both as reported by Bujotzek *et al.*¹⁷ and as predicted by the ABangle server, are shown in Table 2.2. Counts of strong successes (OCD ≤ 1.0), total successes (OCD ≤ 2.0), and failures (OCD > 2.0) are included for each protocol.

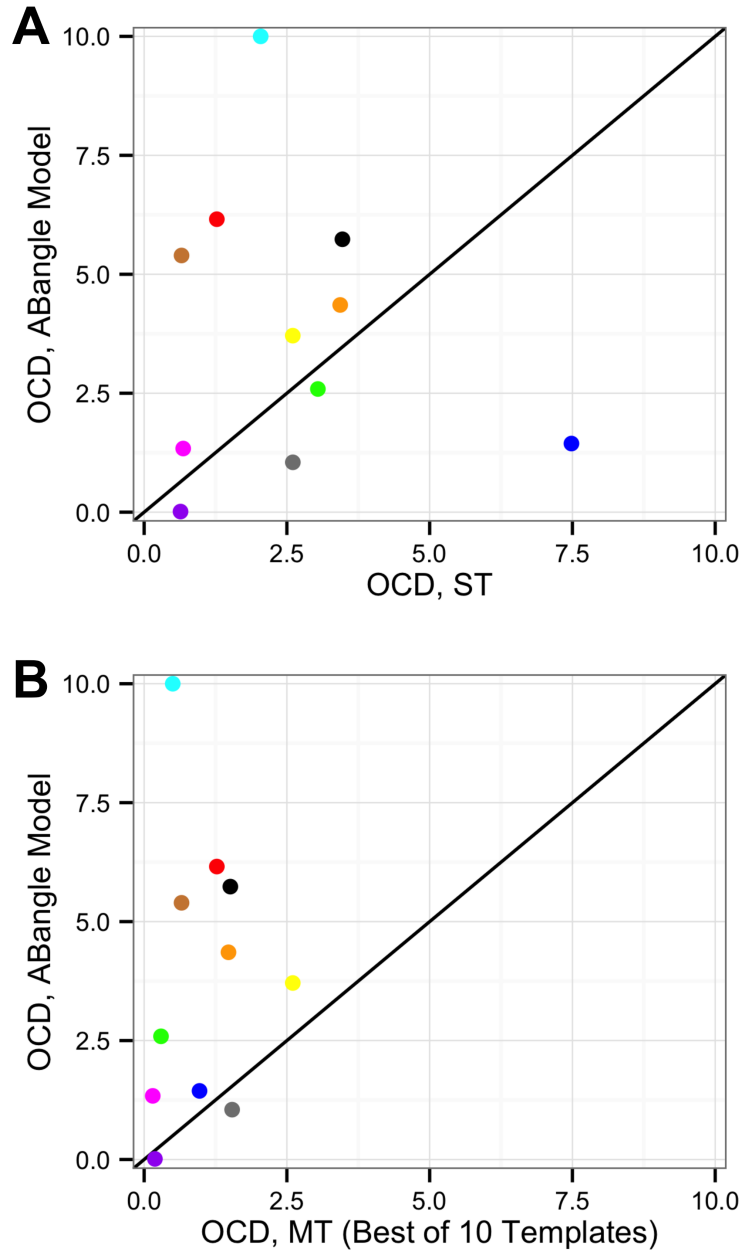


Figure 2.11: Results of ST (A) and MT (B) RosettaAntibody after the grafting stage for the 11 members of the AMA-II set compared to the predictions of ABangle.¹⁷ In (B), only the template with the lowest OCD is plotted; the other nine MT templates are omitted. Points above the line indicate targets in which the RosettaAntibody models are more accurate than the ABangle models, and vice versa.

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

Table 2.2: Performance of ST and MT RosettaAntibody and ABangle in capturing V_L - V_H orientation for the 11 members of the AMA-II antibody set. Both the ABangle results reported by Bujotzek *et al.*¹⁷ and results from the ABangle server are shown. † Best OCD of 10 MT grafted structures.

Target	ST	MT†	ABangle (paper)	ABangle (server)
1	2.60	1.54	0.38	1.05
2	3.47	1.51	2.61	5.74
3	7.48	0.97	2.60	1.44
4	3.04	0.30	1.07	2.59
5	1.27	1.27	8.15	6.16
6	2.60	2.60	1.53	3.71
7	3.43	1.48	4.96	4.36
8	0.64	0.19	3.36	0.01
9	0.68	0.15	1.41	1.34
10	2.04	0.50	0.78	15.11
11	0.66	0.66	4.36	5.40
Strong successes (≤ 1 OCD)	3/11 (27%)	6/11 (55%)	2/11 (18%)	1/11 (9%)
Successes (≤ 2 OCD)	4/11 (36%)	10/11 (91%)	5/11 (45%)	4/11 (36%)
Failures (> 2 OCD)	7/11 (64%)	1/11 (9%)	6/11 (55%)	7/11 (64%)

2.5 Discussion

Predicting V_L - V_H orientation in antibodies is not trivial, though it has been treated as such until recently, with no one quantifying it, let alone explicitly predicting it, until 2010.¹ The sequence signal determining V_L - V_H orientation is less strong, or at least less well-understood, than the conserved sequences of non-H3 CDR loops. Prediction is made more difficult by the wide-ranging yet fine-grained variation of V_L - V_H orientation: the V_L and V_H domains do not fall neatly into discrete canonical

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

conformations, and the qualities of a successful prediction are less clear than those of a CDR loop. Quantifying the orientation unambiguously is thus an important step toward "setting the goalposts" by defining the success case: where a predicted structure and a native structure have matching orientation definitions. The new framework, LHOC, with just a four-dimensional complexity, creates a functionally unambiguous orientation definition, where two structures with similar LHOC metrics will always superimpose within the tolerance of their intra-domain structural differences.

The addition of multiple-template grafting into RosettaAntibody advances V_L - V_H prediction. While the quick rMT protocol only makes slight gains on the ST protocol, sacrificing accuracy for speed, the full-length MT protocol makes nearly universal gains on the former standard, sampling orientationally-accurate candidate structures in 93% of the targets in my benchmark set. By including additional candidate V_L - V_H donor orientation models, MT RosettaAntibody also doubles the number of correctly predicted targets within the AMA-II benchmark set relative to the ABangle prediction method. Though ABangle's single prediction is more accurate, on average, than the ST prediction, the ten predictions from MT RosettaAntibody cover a larger conformational space, producing higher fidelity predictions overall. MT RosettaAntibody is not necessarily limited to using only RosettaAntibody predictions, however; it is easily extensible. Outside predictions, such as ABangle's, could replace one of the ten templates or be added as an eleventh, which would likely improve the predictive power further. A limitation of the new MT RosettaAntibody approach is that it

CHAPTER 2. ANTIBODY V_L - V_H ORIENTATION

requires significantly more computation time: more than 1000 CPU hours are needed per prediction.

The V_L - V_H orientation is only one part of the paratope orientation, but it is closely coupled to the other parts. Improving the ability to predict V_L - V_H orientation will improve the ability to predict the conformation of CDR H3, a grand challenge of antibody homology modeling. A correct V_L - V_H orientation places the H3 stem residues in the correct location, and it defines the available space through which the H3 loop can fold between the L and H chains. Conversely, better H3 prediction methods should also benefit orientation predictions by limiting the V_L - V_H geometries that can closely pack with the CDR H3. Ultimately, in antibody modeling, the whole is more than the sum of the parts.

Chapter 3

Modeling Oblong Proteins and Water-Mediated Interfaces with RosettaDock in CAPRI Rounds

28–35

Adapted from Marze NA[†], Jeliazkov JR[†], Roy Burman SS, Boyken SE, DiMaio F, & Gray JJ, "Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35," *Proteins* 85(3), 479-486. Copyright 2016 Wiley Periodicals, Inc. Reproduced with permission. [†]Equal-contribution authors

3.1 Overview

The 28th–35th rounds of the Critical Assessment of Protein Interactions (CAPRI) served as a practical benchmark for my RosettaDock protein–protein docking proto-

cols, highlighting strengths and weaknesses of the approach. I achieved acceptable or better quality models in three out of 11 targets. For the two α -repeat protein–green fluorescent protein (α rep–GFP) complexes, I used a novel ellipsoidal partial-global docking method (Ellipsoidal Dock) to generate models with 2.2 Å/1.5 Å interface RMSD, capturing 49%/42% of the native contacts, for the 7-/5-repeat α rep complexes. For the DNase–immunity protein complex, I used a new predictor of hydrogen-bonding networks, HBNet with Bridging Waters, to place individual water models at the complex interface; models were generated with 1.8 Å interface RMSD and 12% native water contacts recovered. The targets for which RosettaDock failed to create an acceptable model were typically difficult in general, as six had no acceptable models submitted by any CAPRI predictor. The UCH-L5–RPN13 and UCH-L5–INO80G de-ubiquitinating enzyme–inhibitor complexes comprised inhibitors undergoing significant structural changes upon binding, with the partners being highly interwoven in the docked complexes. My failure to predict the nucleosome–enzyme complex in Target 95 was largely due to tight constraints I placed on my model based on sparse biochemical data suggesting two specific cross-interface interactions, preventing the correct structure from being sampled. While RosettaDock’s three successes show that it is a state-of-the-art docking method, the difficulties with highly flexible and multi-domain complexes highlight the need for better flexible docking and domain-assembly methods.

3.2 Introduction

Proteins play important roles in cellular structure, metabolic activity, biochemical signaling, and multitudes of other biological functions. A protein’s function is determined by its three-dimensional structure, particularly how this structure interacts with other proteins or other biological molecules to form complexes. Consequently, if the structure of protein complexes can be predicted, the nature of their function can likewise be elucidated. Though experimental methods exist to determine protein structure (X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy among others), these are costly, time consuming, and low-throughput. Computational structure prediction is an alternative that can quickly and cheaply generate a structural model of a protein complex.

The Critical Assessment of Protein Interactions (CAPRI) is a long-running community-wide project that evaluates the performance of state-of-the-art computational protein–protein docking methods.⁷⁶ A set of experimentally determined protein complex structures are withheld before publication, and protein docking groups are invited to submit their computational predictions of these structures. These predictions are assessed for accuracy by comparison with the experimentally determined structures. Thus, CAPRI serves as an important benchmark to evaluate the state of the field of computational protein docking, and to reveal remaining challenges. The Gray Lab group has participated in CAPRI since its inception to evaluate the development of the RosettaDock docking method.³³ RosettaDock is, at its core, a Monte-Carlo

CHAPTER 3. CAPRI ROUNDS 28–35

based rigid-backbone docking method with side chain optimization. RosettaDock is extensible, and several ancillary protocols have proven effective in previous CAPRI rounds;⁷⁰ the conformer-selection protocol EnsembleDock²¹ and the flexible-loop induced fit protocol SnugDock⁷¹ are among the most broadly useful.

The prior set of CAPRI rounds (20–27) highlighted the need for docking tools that could do more than assemble unmodified globular proteins, namely model non-neutral pH, carbohydrates, and water-mediated interfaces.⁴⁰ This trend was continued by CAPRI rounds 28–35, with targets including a DNA/protein nucleosome, protein–peptide complexes, and more water-mediated interfaces. I did not attempt the peptide complexes, knowing that the Rosetta approaches would be predicted by the Furman team, who developed the Rosetta-based FlexPepDock. For the protein complexes, my team developed two new protocols to account for the eccentricities of certain targets in these rounds: (1) Ellipsoidal Dock, to account for oblong proteins such as those in Targets 96/97, and (2) HBNet with Bridging Waters, to predict hydrogen-bonding networks at water-mediated interfaces such as those in Targets 104/105.

CAPRI rounds 28–35 also continued to produce challenging globular-protein docking targets. Some of their challenges have already been well documented: proteins that exhibit large conformational changes on binding,^{13,47} partners that must be homology modeled before docking,⁵¹ and global docking targets lacking a homology complex or specific biochemical information about the binding site.⁶⁰ Two other challenges were more clearly defined by these CAPRI rounds: proteins that become

significantly entwined during docking (Targets 98–101), and partners requiring multi-domain assembly before docking (Targets 102/107). The clarification of remaining docking challenges aids the future development of RosettaDock, and the success of new methods such as Ellipsoidal Dock and HBNet provides a blueprint for future enhancements to the core protocol.

3.3 Methods and Results

We submitted predictions for 12 targets (13 if Target 107, the re-run of Target 102 is counted separately) across the seven standard CAPRI rounds from 28–35. (The hybrid CASP/CAPRI round 30 is not covered by the scope of this paper [see Lensink et al.⁵¹]; I did not submit predictions for any peptide docking targets, as my collaborators in the Furman lab are more adept at the peptide-specific Rosetta methods that I would have used.⁵⁴) I generated two medium- and one acceptable-quality predictions (Table 3.1). Additionally, I achieved one fair-quality water prediction among the two targets that required explicit water predictions. I present my successes first, along with descriptions of the corresponding novel methods used in these predictions. I follow with the targets in which my predictions failed, accompanied by descriptions of the targets’ difficulties, and lessons learned in *ex post facto* analysis.

Table 3.1: Summary of my results in CAPRI rounds 28-35 for targets I predicted. Target: the CAPRI target ID; Complex: the docking partners; Type: model for each docking partner: homology model (H) or unbound structure (U); f_{nat} : fraction of native contacts recovered; Lrmsd: ligand RMSD in Å; Irmsd: interface RMSD in Å; Quality: the CAPRI rating of the best predicted model: high (**), medium (**), acceptable (*), or incorrect (-); Method: the Rosetta docking protocol used to generate predictions; Search scale: breadth of the docking search. † Interface water prediction ratings.

Target	Complex	Type	Metrics of best model						Quality	Method	Search scale
			f_{nat}	Lrmsd	Irmsd	Quality	Method	Search scale			
59	Rps28b-Edc3	U-U	0.00	29.7	9.9	-	EnsembleDock	Local			
95	Nucleosome-Bmi1/ Ring1b-UbcH5c	U-U	0.00	23.2	11.4	-	Constrained	Local			
96	arep7-GFP	H-U	0.42	4.0	1.6	**	Ellipsoidal Dock	Partial global			
97	arep5-GFP	H-U	0.26	6.4	2.3	*	Ellipsoidal Dock	Partial global			
98	UCH-L5-RPN13	U-U	0.01	24.6	9.9	-	EnsembleDock	Global			
99	UCH-L5-Ub-RPN13	H-U	0.00	20.5	7.1	-	EnsembleDock	Global			
100	UCH-L5-Ub- INO80G	H-H	0.05	22.3	7.4	-	EnsembleDock	Global			
101	UCH-L5-INO80G	U-H	0.00	38.1	18.3	-	EnsembleDock	Global			
103	Ube27-FAT10	H-H	0.01	42.5	14.5	-	Constrained	Local			
104	DNase-ImAP41	H-H	0.07	13.4	5.7	-/bad†	EnsembleDock Standard/ HBNetBW†	Local			
105	DNase-ImS2	H-H	0.48	4.1	1.8	**/fair†	Standard/ HBNetBW†	Local			
102/107	HxuA-Hemopexin	U-U	0.00	27.4	17.0	-	Standard	Global			

3.3.1 Successes

3.3.1.1 Targets 96/97: α rep–GFP

In Targets 96 and 97, I was challenged to dock GFP to one of two α -repeat (α rep) proteins. The GFP sequence had ~ 5 point mutations relative to GFP of PDB ID 1JBZ;³⁶ I used RosettaDesign^{46,50} to make the appropriate point mutations on the crystal structure. The α rep proteins were highly homologous to the 6-repeat protein of 3LTJ,⁷⁵ but with one more (Target 96) and one fewer (Target 97) repeat subunit. To generate the α rep homology models, I spliced in/out a single non-terminal repeat from 3LTJ while maintaining the topological curvature of the template. I then optimized the α rep models, and to account for my uncertainty in the α rep structure I created 30-member docking ensembles using the Rosetta Relax^{15,58} protocol. Though no homologous complex was available in the PDB, the concave face of the α rep protein and the GFP β -barrel exhibited high shape complementarity; as such, I posited that the GFP must dock within the α rep concave face.

The geometric symmetry of GFP's β -barrel necessitated the use of global docking to properly place it in the α rep concave face. A new randomization method, Ellipsoidal Dock (Figure 3.1), was used for the global docking of GFP. The standard Rosetta global docking randomizes the Euler angles of the protein partners, implicitly treating them as spheres. When a partner is oblong, like GFP, such randomization inefficiently samples polar regions of ellipsoidal proteins and creates poor contacts in

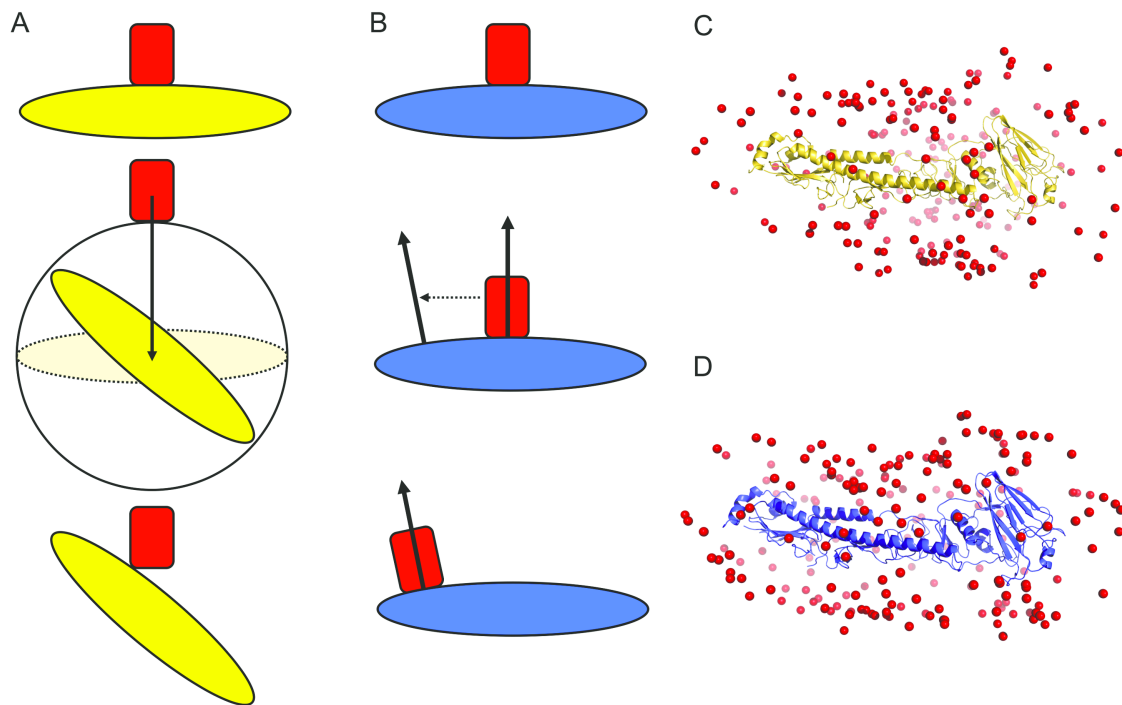


Figure 3.1: Global randomization in (A) standard Rosetta and (B) Ellipsoidal Dock. The standard global randomization pulls the docking partners apart, randomizes the Euler angles of the implicit sphere circumscribing one protein partner, and pulls the partners back into contact. Ellipsoidal Dock calculates the normal vectors from the surface of one partner at the starting point of contact, as well as at a random point on the surface, and superimposes the two normal vectors. The distributions of 200 oblong candidate complexes as generated using standard randomization (C) and Ellipsoidal Dock (D) are shown. The red spheres represent the center of mass of the antibody in each candidate complex.

the putative docked complexes (Figure 3.1.A/3.1.C). Ellipsoidal Dock corrects these issues by randomly selecting a point on the ellipsoidal surface approximation of the protein, and aligning the normal vector at this point to the normal vector from the other partner’s interface (Figure 3.1.B/3.1.D). The principle of aligning normal vectors to preserve shape complementarity is similar to that used in ICM-DISCO global docking;^{30,31} however, while ICM-DISCO is built for enumerative searches and uses a polyhedral surface approximation with one normal vector per face, Ellipsoidal Dock uses a smooth surface with a continuous distribution of normal vectors, appropriate for Rosetta’s stochastic Monte Carlo sampling methods.

The Ellipsoidal Dock protein surface approximation uses the standard equation for an ellipsoid:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

where the centroid of the protein’s $C\alpha$ atoms defines the origin, and the 1st, 2nd, and 3rd principal components of the $C\alpha$ atom set define the z , x , and y directions, respectively. The parameters c , a , and b are taken as twice the square root of the eigenvalue corresponding with the 1st, 2nd, and 3rd principal component eigenvectors, respectively. The surface area of an ellipsoid does not have a closed-form solution, so approximations are made to sample it evenly. I sample the z -coordinate first from a beta distribution with parameters $\alpha = 1.5$ and $\beta = 1.5$, scaling the distribution over the z -length of the ellipsoid. I also sample the x -coordinate from a beta distribution,

but with parameters between 0.5 (used when $\alpha' = \beta'$) and 1.0 (used when $\alpha' \gg \beta'$). Finally, I select the y-coordinate as either the positive or negative y-coordinate corresponding with the chosen x- and z-coordinates.

In addition to using Ellipsoidal Dock to fully sample the GFP orientation, I also used (1) a larger-than-standard initial translational/rotational perturbation (8 Å translational parameter vs. 3 Å standard) to broaden the search scope along the α rep crevice and (2) Rosetta EnsembleDock to sample different α rep conformers from my docking ensemble. I generated 20,000 candidate structures (decoys) for each target.

Among my 10 submitted models for each target, I achieved one medium-quality structure for Target 96 and two acceptable-quality structures for Target 97. My highest-quality structure for Target 96 (Figure 3.2.A) had a root-mean-squared displacement of interface atoms (Irmsd) of 1.577 Å and 0.420 fraction of native contacts (f_{nat}), both the second-best among all submissions. This structure was my highest-ranked model. My highest-quality structure for Target 97 (Figure 3.2.B) had an Irmsd of 2.251 Å and 0.256 f_{nat} .

3.3.1.2 Targets 104/105: DNase–Immunity protein

Targets 104 and 105 presented a dual challenge: first, to predict the complex structure of a DNase (PyoAP41 or PyoS2) with its cognate immunity protein (ImAP41 or ImS2), then to predict the mediating waters and side chains at the protein interface. While the DNase proteins had crystal structures available in the PDB, I had

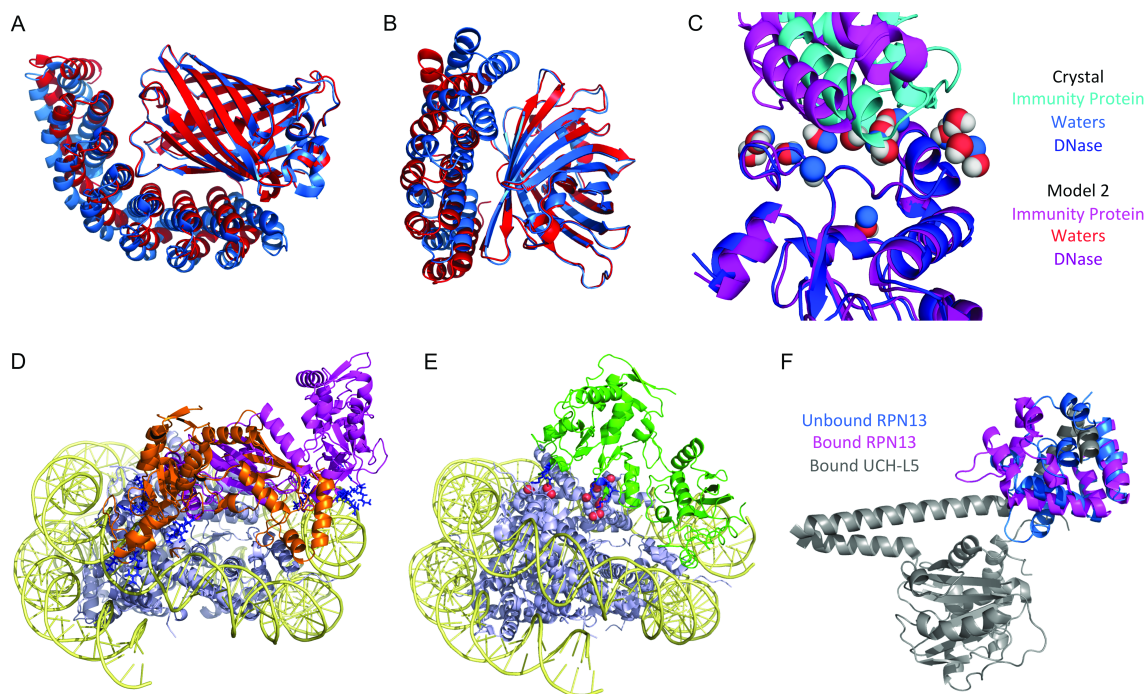


Figure 3.2: (A & B) My best-quality models for Targets 96 & 97 (red), superimposed with their complex crystal structures (blue). (A) shows my medium-quality model for Target 96; (B) shows the better of my two acceptable-quality models for Target 97. (C) My best medium-quality model for Target 105, superimposed with the crystal structure. My model is colored in red/purple shades, while the crystal structure is colored in blue shades. (D) Two of the three binding modes used in my docking simulations for Target 95. The mode implicated by Bentley et al.¹⁰ is shown in orange, while my first novel mode is shown in magenta. The six residues implicated as DNA-binding are colored in dark blue; in all my modes, these residues contact DNA directly. (E) The native binding mode for Target 95, shown in green. The six residues implicated in DNA-binding are again colored in dark blue; none of these residues contact DNA directly, instead making salt bridges with positively charged residue, colored in red. (F) The bound UCH-L5–RPN13 complex (grey and pink), with the unbound RPN13 superimposed on top (blue). Upon binding, the RPN13 helical bundle hinges open to accommodate the UCH-L5 C-terminal helix.

to generate generated homology models for ImAP41 and ImS2. The available homologous structure was colicin Im2 (chain A in PDB ID 3U43⁸⁹), a previous CAPRI target (T47) with 50% identity to ImAP41 and 59% identity to ImS2. The starting complexes were then generated by aligning the homology models (ImAP41 or ImS2) and structures (PyoAP41 or PyoS2) to their homolog’s position in PDB 3U43. For T104, I then used structural ensembles of PyoAP41 to account for a flexible loop at the interface and ran an local EnsembleDock to optimize the complex (50,000 decoys). For T105, I ran a local RosettaDock to optimize the complex (20,000 decoys). I used a new method for interface water predictions: HBNet with Bridging Waters (HBNetBW).

We expanded HBNet, a method for designing hydrogen bond networks,¹⁴ to include a statistical potential to capture water molecules that form bridging hydrogen bonds between side chains. The two-term potential utilizes the distance between the two protein atoms that hydrogen bond to the water molecule (acceptor or donor polar hydrogen) and the dihedral angle between those two atoms and their base atoms (e.g. the base atom for a carbonyl oxygen acceptor would be the carbon it is double bonded to, and the base atom of the polar hydrogen would be the heavy-atom donor that it is covalently bonded to). I calibrated the potential using interface waters from the Top 8000 dataset²³ and bicubic spline interpolation; the two-dimensional function that defines the bridging water score is:

$$score(a_1, a_2, a_3, a_4) = f(\text{distance}(a_2, a_3), \text{dihedral}(a_1, a_2, a_3, a_4))$$

where a_2 and a_3 are the protein atoms hydrogen bonded to the bridging water, a_1 is the base atom of a_2 , and a_4 is the base atom of a_3 . To identify water positions during HBNet search, if two rotamers have a bridging water score below a specified threshold, they are connected as part of a potential hydrogen bond network and an explicit water molecule is placed at ideal geometry relative to the hydrogen-bonding atoms. I ran HBNetBW on each docked backbone, sampling rotamers of the interface residues to identify the most satisfied networks. There is a substantial energetic penalty associated with burying polar atoms that do not participate in hydrogen bonds (either to solvent or other protein atoms); thus, I hypothesized that using this criterion would be advantageous for discriminating between docked complexes.

For Target 104, all of my models were incorrect. An *ex post facto* analysis revealed that my homology model had the correct complex orientation, and that my docking simulation moved the complex away from that conformation. For Target 105, all four of my submitted models were of medium quality, the best having an Irmsd of 1.757 Å and an 0.481 f_{nat} . Similar to Target 104, however, the unrefined homology model had a more native-like orientation than my docked model. One of my models from Target 105 had a fair-quality water prediction (Figure 3.2.C), with a waters-only f_{nat} of 0.118, indicating that HBNet can be useful even without a perfectly-aligned interface.

After the CAPRI blind challenge, I ran HBNetBW on the revealed crystal struc-

CHAPTER 3. CAPRI ROUNDS 28–35

tures for Targets 104 and 105 and the closest homology model to each. I removed water molecules from the structures. I then relaxed (cycles of minimization and side chain repacking) the structures using Rosetta. Next, I ran HBNetBW using identical parameters to those during analysis of submitted docked complexes. In regions of the interface where the backbone was close to that of the crystal structure, the native side chain hydrogen bond networks were largely recapitulated, and a couple of the bridging water molecules were placed in agreement with interface waters in the crystal structure; for example, running HBNetBW on the T105 homology model generated a network with a bridging water molecule between Tyr640, Tyr55, and His34 that is in very close agreement to the experimental crystal structure. However, many false-positive networks and water placements were also generated – multiple networks are identified for each fixed-backbone decoy, making it challenging to choose which networks and water placements to keep and which to discard. Ranking networks according to satisfaction and connectivity led to success in designed protein-only networks;¹⁴ however, as used here, these metrics are only as reliable as the bridging water identification and placement, and my results suggest that there is significant room for improvement to both.

3.3.2 Failures

3.3.2.1 Target 95: nucleosome–Bmi1/Ring1b–UbcH5c

The challenge in Target 95 was to dock the ubiquitinating enzyme complex Bmi1/Ring1b–UbcH5c to a nucleosome. The unbound forms of both partners were available, in 3RPG¹⁰ and 3LZ0⁷⁸ for the enzyme complex and the nucleosome, respectively. The scientists who solved the unbound enzyme crystal structure (3RPG) hypothesized a binding mode for the complex of interest predicated on two structural constraints: (1) a 2 Å distance constraint between the ubiquitin donor residue (UbcH5c, Cys85) and the ubiquitin acceptor residue(s) (nucleosome H2A, Lys119(/Lys118)), and (2) an ambiguous interaction constraint between four enzyme complex residues (Bmi1, Lys62/Lys64, Ring1b, Arg97/Arg98, all implicated through mutagenesis experiments as DNA-binding) and the nucleosome DNA.¹⁰ I used this binding mode as one starting structure for a local RosettaDock simulation (10,000 decoys). I also identified two other binding modes meeting the Bradley constraints (Figure 3.2.D) and launched local RosettaDock simulations from each of these starting structures (10,000 decoys each). Constraint (1) was enforced with a flat harmonic score function penalty during all docking runs, albeit at a looser minimum penalized distance of 15 Å while constraint (2) was used as a post-filter to remove any structures without at least one key residue contacting the nucleosome DNA.

All of my submitted models were incorrect (closest Irmsd: 11.4 Å). Examination

of the complex crystal structure (4R8P) showed that constraint (2) was not preserved. While the constraint assumed that the patch of four positively charged amino acids would contact the DNA directly, they in fact make salt bridges with the protein core of the nucleosome (Figure 3.2.E). As a result, my post-filter constraint prevented me from finding the correct binding mode.

3.3.2.2 Targets 98–101: UCH-L5(\pm Ub)–[RPN13 or INO80G]

Targets 98–101 provided a combinatorial docking challenge which asked me to dock deubiquitinating enzyme UCH-L5, with or without its conjugate ubiquitin (Ub), to either of two inhibitors, RPN13 or INO80G. Unbound structures of UCH-L5, RPN13, and Ub were available (3IHR, 2KQZ, and 1UBQ, respectively). I homology modeled INO80G by threading from PDB structure 2KQZ, loop-building, and refining in Rosetta. Additionally, I built a homology UCH-L5–Ub complex by aligning the two proteins to PDB structure 4IG7. Using the FloppyTail protocol⁴¹ I modeled the tails of RPN13, which are unresolved in 2KQZ, and the homologous regions of INO80G. I found no biochemical data or homology complexes that clearly identified a binding site, necessitating a global docking search. Due to the uncertainty in the monomer structures, I ran EnsembleDock with 30-member ensembles (generated by relaxing my top homology models). 20,000 decoys were generated for each target.

These targets were quite difficult: across all four targets, no CAPRI group submitted a model of acceptable-quality or better. Comparison of the complex binding

mode to the unbound structures revealed that RPN13 undergoes a significant conformational shift upon binding, in which a helical bundle hinges open to bind around a helical element from UCH-L5, which itself undergoes a substantial kinking upon binding (Figure 3.2.F). Though INO80G has no unbound structure to compare with its bound forms, the inhibitor is similarly entwined with the UCH-L5 helix. This binding mode is doubly difficult to predict. Firstly, predicting conformational change upon binding has been observed to be difficult in previous CAPRI challenges, particularly when the change is so large. Secondly, the degree of structural entwinement between the two partners requires a hybrid folding/docking algorithm to predict correctly: the bound forms of RPN13 and INO80G would have high energies in solution due to their open hydrophobic pocket, and even if these forms could be predicted, due to the high degree of entwinement they would be almost impossible to dock by rigid-body methods.

3.3.2.3 Target 102 (107): HxuA–Hemopexin

Target 102 challenged me to assemble and dock the multi-domain protein HxuA to the heme storage protein hemopexin. The challenge was repeated in Target 107 with the unbound structure of HxuA provided. The unbound structure of hemopexin was suggested in both challenges as PDB ID 1QHU.⁹ In my attempt to assemble HxuA, I used the Robetta server to predict the individual domains, then I used ClusPro^{25,26,42,43} to dock the domains together one-by-one, with Rosetta CCD^{18,81} used

to close the linking regions. Robetta predicted four domains, with domain 1 matching PDB ID (4I84⁸), domain 4 being a similar β -solenoid shape, and domains 2 and 3 being small helical and sheet linking domains, respectively. The unbound structure of HxuA provided in Target 107 revealed a few key errors in my assembled structure: (1) domain 2, predicted by Robetta as helical, is entirely a beta-sheet, (2) HxuA is not made up of distinct domains, but rather is a single extended β -solenoid, and (3) I mistakenly inserted domain 4 at the N-terminus, effectively inverting the entire domain. No binding site was identified through either homology or biochemical data, so I ran a Rosetta global dock, using Ellipsoidal Dock to account for the elongated HxuA β -solenoid. I produced 10,000 decoys for both Target 102 and Target 107.

In both targets, all of my submitted models were incorrect. The failures in Target 102 can largely be attributed to my incorrect model of HxuA. The failures in Target 107 are less easily attributed. A component of the failure is likely the size of the complex. HxuA is 884 residues; to fully sample the protein with a Rosetta global dock, I would generally produce between 100,000 and 1,000,000 decoys; however, the time constraints of the CAPRI competition limited me to 10,000 decoys. Perhaps more critically, though, the binding conformation is mediated by a 23-residue loop on HxuA that undergoes significant remodeling during binding, inserting into the heme-binding site of hemopexin. As my simulation did not account for any backbone flexibility, this loop was completely unavailable to the binding site in hemopexin.

3.3.2.4 Target 103: UBE2Z–FAT10

Target 103 presented the challenge of predicting the complex of UBE2Z, a ubiquitin-conjugating (E2) enzyme, and FAT10, a diubiquitin analogue. I used Modeller⁸⁴ to generate homology models for UBE2Z and FAT10 (from PDB structures 3CEG⁶⁹ and 4KSL,⁶⁶ respectively) for use in my docking ensembles. I ran a 50,000 decoy, high-perturbation local EnsembleDock from a putative binding conformation based on homology to other E2 ubiquitinating enzymes. During docking, I imposed a 16 Å distance constraint between the C-terminal residue of FAT10 and Cys-160 on UBE2Z, with the latter posited as the FAT10 carrier site. I achieved no models of acceptable or better quality, either in the evaluation of the full complex, or in the separate evaluations of the binding sites of either the C-terminal or N-terminal FAT10 domains.

3.3.2.5 Target 59: Rps28b–Edc3

In Target 59, I was challenged to dock ribosomal protein Rps28b to mRNA decapping enzyme Edc3. 20-member unbound NMR ensembles were provided for each partner: 1NE3⁹⁰ for Rps28b, and 4A53³² for Edc3. I ran 10,000 local EnsembleDock decoys from each of six putative binding sites manually identified by examining the solvent-accessible faces of Rps28b in the context of the ribosome and Edc3 in the context of homologous hexamers. I achieved no acceptable or better structures in my ten predicted structures, but I did achieve one acceptable structure in my 100 uploaded structures.

3.4 Discussion

My group’s CAPRI performance reveals strengths and limitations in my docking abilities. I achieved a successful prediction in 3 of 12 targets. Compared to the community as a whole, however, my performance is not atypical, as six targets did not elicit a single successful prediction from any team. I did not participate in the peptide docking targets, but my collaborators in the Furman lab did.⁵⁴ When the results of my submissions (3*/2**) are adjoined with those from the Furman lab (3*/3**/1***) on different targets, Rosetta docking approaches (6*/4**/1***) had acceptable or better predictions in six targets and medium or better predictions in four targets, which would position the combined ranking somewhere in the top ten of all predictors.⁵³ Other CAPRI predictors included Rosetta refinement in their approaches (Baker, Bradley, Guerois, etc.). In fact, Guerois, who incorporated Rosetta-based refinement as the final refinement and discrimination stage in their pipeline, predicted 9/18 targets correctly. Furthermore, for targets with at least one correct prediction, the Rosetta-based approaches yielded models closest to native. Thus, Rosetta remains a state-of-the-art computational tool that can successfully predict a diverse set of protein complexes.

The new Rosetta methods I tested during these CAPRI rounds worked especially well, leading to all three of my group’s success cases. Ellipsoidal Dock’s ability to appropriately adapt my search to the oblong shape of GFP led to my successes in predicting Targets 96 and 97, which were both difficult targets for the community.

CHAPTER 3. CAPRI ROUNDS 28–35

My group was the top predictor group for Target 96, and in the top-five for Target 97. HBNetBW, even its early stage of development, was able to achieve a fair-quality interface water prediction for Target 105 despite errors in the docked partners. However, like in previous CAPRI experiments, all of my successes were small protein complexes with little flexibility upon binding and with clues about the native binding sites, either by homology complex or obvious shape complementarity.

These CAPRI rounds reveal the shortcomings of my docking methods and the remaining challenges for the docking community as a whole. Large and multi-domain targets remain quite challenging, even when they are otherwise tractable docking challenges. Target 95, a 1639 residue complex that is nearly rigid upon binding and has an abundance of biochemical data restraining the complex, was only successfully predicted by three predictor groups, with only one medium-quality model between them. Target 102, a 1,098 residue complex that also exhibits a full loop remodeling at the active site upon binding, did not elicit a single acceptable-quality or better model from any predictor, even when the full unbound structure of the larger partner was provided. To allow prediction of large complexes, future global docking methods must be able to sample the resultant large conformational space more efficiently.

The latter large target is also indicative of the other key remaining challenge, large conformational changes during binding, which will often confound all existing docking methods. Targets 98-101 are all small complexes in which the structure of the binding residues of the larger partner are predictable by homology. Not one tar-

get, even T98 where unbound structures of both partners were provided, had a single correct model predicted. While this difficulty can be attributed to the conformational changes upon binding, it can also be attributed to the severe entwinement of the two partners in the bound state. This entwinement requires a more sophisticated set of docking methods where flexibility and docking orientation are sampled concurrently, as opposed to existing methods such as Rosetta's EnsembleDock, which largely separates the sampling of the flexibility and the docked conformation. The revelation of these CAPRI rounds that there are two distinct challenges in flexible docking will provide insight into future development of flexible docking methods.

Chapter 4

Motif Dock Score (MDS): A fast, accurate coarse-grained score function for flexible-backbone docking

4.1 Overview

Binding-induced conformational changes in protein complexes have confounded protein–protein docking algorithms by greatly increasing the degrees of freedom of the system. To properly sample the large conformational space in flexible docking protocols, a fast, accurate evaluation method is needed. I built Motif Dock Score (MDS) to address this need. MDS maps full-atom residue-pair energies, generated with the Rosetta REF15 score function, to a backbone-geometry-indexed hash ta-

CHAPTER 4. MOTIF DOCK SCORE

ble, allowing the RosettaDock protocol to score models using only a single lookup per residue pair. I optimized MDS for the enrichment of near-native models during a docking search, and I benchmarked its performance on a nine-target set of complexes. Without requiring additional computational time, MDS shows a ninefold improvement in near-native discrimination over centroid score ($N5 = 3.6$ vs. $N5 = 0.4$) and a threefold improvement in near-native enrichment ($N100 = 5.9$ vs. $N100 = 2.0$), ultimately showing a successful near-native enrichment in seven of the nine benchmark targets. MDS's performance compares favorably to other leading docking methods, particularly in regards to the prediction of difficult and flexible targets. Given its strong performance, MDS will be a critical component of next-generation protein–protein docking protocols.

4.2 Introduction

Conformational changes in proteins induced by binding have confounded protein–protein docking algorithms by greatly increasing the degrees of freedom to be sampled. While rotamer libraries have alleviated the sampling challenges for side-chains,⁴⁵ backbone flexibility remains a principal challenge in docking. Previous studies have found limited success by varying the backbone along a restricted set of coordinates^{57,59,79} or interface residues^{67,82} or by docking a small number of backbone conformations of the two partners.²¹

CHAPTER 4. MOTIF DOCK SCORE

RosettaDock has historically been among the top-performing methods for computational protein–protein docking.^{22,27,34,40,70} The last major version, RosettaDock 3.2, achieved a successful docking prediction on a majority of rigid body targets (58%) in the Docking Benchmark 3.0 set.²⁰ On the more flexible targets, however, RosettaDock (like other methods) performed poorly, only achieving a successful docking prediction on 29% of the medium-difficulty targets and 14% of the difficult targets. RosettaDock’s performance in CAPRI rounds since the last major version release mimicked the benchmark performance. On two rigid body targets, RosettaDock outperformed nearly all other docking methods, but on targets with significant conformational change upon binding, RosettaDock failed to produce any correct docked conformations.⁵⁵ The recent CAPRI rounds proved protein flexibility to still be a community-wide docking weakness, with several such targets eliciting no successful predictions from any method.⁵³

Flexible-backbone docking, as well as the other key remaining protein–protein docking challenges, global docking and docking of large multi-domain complexes, demands more algorithmic complexity to explore a larger conformational search space than rigid-body docking of small complexes.⁴⁷ These sampling approaches are more computationally intensive, and they must be balanced with scoring methods that are more efficient to prevent massive increases in computational resource utilization. RosettaDock’s current low-resolution scoring method, centroid score, is fast to calculate; however, it does not correctly identify native docked conformations,⁹² ultimately

CHAPTER 4. MOTIF DOCK SCORE

making it an inefficient score function. Thus, a new score function that is both fast and accurate is needed.

One promising score function, the residue-pair transform (RPX) score, was recently developed and used to design hydrophobic symmetric protein interfaces.²⁹ RPX score evaluates and mutates residue pairs using only the 6D transformation needed to superimpose the residues' N-C_α-C backbones onto each other. In a single lookup, RPX score queries this transformation against a pretabulated database of aliphatic amino acid pairs and their corresponding geometries and full-atom Rosetta scores. The pair score and sequence of the best amino acid pair from the database are then assigned to the queried residue pair.

In the project described in this chapter, I used the RPX framework to build a new low-resolution protein interface score function, Motif Dock Score (MDS). I optimized MDS in the context of the RosettaDock protocol, selecting for enrichment of near-native decoys. I then examined the performance of MDS, both singly and in concert with enhanced sampling method Adaptive EnsembleDock, on a benchmark set of nine targets randomly selected from Docking Benchmark 5.0.

4.3 Methods

4.3.1 PDB Curation

I culled the PDB for all crystal structures containing two or more interacting protein chains and a resolution of 3.0 Å or better. I also removed any structures present in the DockingBenchmark 5.0 to be used as a test set. In the remaining set, PDB structures with more than two chains in their asymmetric unit were further divided such that one structure represented every pair of interacting protein chains in their asymmetric unit. The PDB structures were then stripped of all HETATM lines and non-canonical amino acids. My curated set contains 154,955 protein–protein complex structures from 103,017 PDB entries.

4.3.2 Motif Querying

Each structure in the protein interface set was loaded into Rosetta and scored with a full-atom score function; the resultant energies were decomposed onto the set of interacting residue pairs. The system was queried for cross-chain pairs of residues within 10 Å of each other with a pair score below a constant energy cutoff (typically 0 kcal/mol; i.e. residue pairs that are net-attractive). For each residue pair in the filtered residue set, I calculated the six-dimensional transform needed to superimpose one amino acid backbone onto the other (three-dimensional Cartesian translation

and three-dimensional Euler Angle rotation). Each pair score was stored with its corresponding 6D-transform as a one-line motif.

4.3.3 Score Grid Generation

A score grid is initialized with a translational and rotational grid size. One by one, motifs are analyzed. The motif 6D-transform is binned, and the corresponding bin in the score grid is queried. If the bin is empty, the motif score is saved as the bin score. If the bin is populated, either an aggregation or minimization method can be used to calculate the new bin score. The aggregation method is calculated by converting the motif score to a pseudo-count weighted by its favorability, then summing the count with the current bin score. The minimization method simply calculates the lower of the old bin score and the motif score, and saves it as the new bin score. If smoothing is being used, the neighboring bins are also queried, with the favorability of the score determining the radius within which the bin scores are updated. Once all motifs have been analyzed, the populated bins are assigned a hash value and, to minimize its memory footprint, only the hashed bins are stored.²⁹

4.3.4 Motif Dock Score

Motif Dock Score uses the same algorithmic framework as RosettaDock, described in Gray *et al.*,³³ with modernizations described in Chaudhury and Gray,²¹

CHAPTER 4. MOTIF DOCK SCORE

Chaudhury *et al.*,²⁰ and Marze *et al.*⁵⁵ The standard low-resolution score function (`interchain_cen`) is replaced with a motif-based score function, called `motif_dock_score`. The score function consists of a new scoring term, `motif_dock`, and a clash penalty (`interchain_vdw`). The `motif_dock` term is a residue pair energy that acts only on cross-chain residue pairs within 10 Å of each other. The residue pairs are scored by calculating their 6D-transform, converting this to the hash value of the corresponding 6D bin, querying the hash table, and reporting the bin score. If the bin is empty (i.e. there are no matches for the hash), the pair score will either be zero if no penalty is used, or 0.5 kcal/mol, if a penalty is used.

4.3.5 Benchmark Set Generation

I built two benchmark sets using subsets of the Docking Benchmark 5.0 set.⁸⁰ The first, a set of eleven targets for rescoring, was randomly selected from the rigid-body subset of Docking Benchmark 5.0 to provide ample near-bound structures to optimize motif scoring’s near-native discrimination ability. To generate the rescoring sets for each target, I ran the standard RosettaDock protocol²⁰ on the unbound complex structures, including translation and rotation perturbations (mean = 3 Å translation, 8° rotation) to the ligand (the smaller protein partner) to disrupt existing interfaces. The second set, a small representative docking benchmark, was generated by selecting four rigid-body targets (1EFN, 1GLA, 2A1A, 2FJU), three medium-difficulty targets (1LFD, 2CFH, 3AAA), and two difficult targets of different categories (2OT3, 3F1P)

from the Docking Benchmark 5.0.

4.3.6 Generation of Backbone Ensembles

To generate diversity in backbone conformations for the Adaptive EnsembleDock runs, I used three conformer generation methods: perturbation of the backbones along the normal modes by 1 Å,⁷ refinement using the Relax protocol in Rosetta,⁷⁴ and backbone flexing using the Rosetta Backrub protocol.⁷³ Each was shown to generate different modes of backbone motions, overlapping by 30-50% with the actual directions of motion between the unbound and bound states.⁴⁷ Since the normal mode analysis generated the largest deviations, I used 40 normal mode conformers, 30 Relax conformers and 30 Backrub conformers to generate the ensemble of 100 conformers.

4.3.7 Benchmark Evaluation Metrics

I evaluated the results of the docking benchmark runs using three bootstrapped metrics: N5, as described by Chaudhury *et al.*,²⁰ and N100 and N1000, described analogously to N5 as the counts of near-native decoys among the lowest-scoring 100 and 1000 candidate structures, respectively. As in Chaudhury *et al.*,²⁰ a decoy must have an interface RMSD ≤ 4.0 Å, and a run is categorized as "successful" when N5 ≥ 3 . I use more lenient success criteria for my N100 and N1000 metrics (N100 ≥ 30 , N1000 ≥ 150), as these are metrics of near-native enrichment rather than near-native

discrimination. Furthermore, I also evaluate the N5, N100, and N1000 metrics after only the low-resolution phase of docking as measures of the near-native enrichment and discrimination of the low-resolution docking protocol alone. Here, I use a more lenient measure of near-native candidate structures, 6.0 Å centroid RMSD, to account for the limitations in measuring RMSD in the centroid phase (incompletely resolved side chains, lever-arm effects away from the interface, etc.). As in Chaudhury *et al.*,²⁰ the metrics are bootstrapped using 1000 resamples of the docking decoy sets.

4.4 Results

My goal was to create a fast, accurate score function for low-resolution protein–protein docking, so ultimately, I wanted to examine the docking performance of my Motif Dock Score protocol. Since, however, I developed MDS on top of the RPX framework, which was built for interface design, I first needed to optimize the score function for protein–protein docking.

4.4.1 Optimization of Motif Dock Score

MDS depends on a discrete space tabulation of all-atom energies; therefore, there are three scoring grid structural factors that might affect the performance of MDS: bin size, smoothing factor, and population method. For my first round of optimization, I examined these factors. I identified Intuitively, bin size has the most direct effect on

CHAPTER 4. MOTIF DOCK SCORE

scoring; a larger bin is more likely to properly identify a low-energy interaction but it is also more likely to assign this low-energy interaction incorrectly to a geometrically dissimilar residue pair; conversely, a smaller bin, due to its highly constrained geometry and the limitations of the motif database, is less likely to assign a low score to an interacting residue pair. I therefore chose five bin sizes to examine, each with a fixed width for the three translation bins and a fixed angle for the three rotation bins. The largest bin has an 8 Å width parameter and a 36° angle parameter, while the smallest bin has an 0.5 Å width parameter and a 12° angle parameter.

The smoothing factor's effect on scoring is more indirect. While a large smoothing factor will populate more bins with interaction energies than a small smoothing factor, the actual conformational space populated also depends on the bin size. I chose three smoothing factors to examine: 0.3, 1.0 (the RPX score default), and 3.0.

There are two simple methods to populate the score bins. The first, minimization, stores only the best-scoring interaction in each bin as the bin score. The second, aggregation, increments a pseudo-count to the bin score for each low-scoring interaction in the bin, weighted by the score of the interaction. The effects on scoring are not obvious, as the former rewards very strong interactions, even if they are only observed once in the PDB, while the latter rewards moderately strong interactions that are seen frequently in the PDB.

From my culled PDB set, I attempted to build experimental scoring grids using all permutations of three factor sets, for a total of 30 scoring grids. Four of these grids,

CHAPTER 4. MOTIF DOCK SCORE

all using the $0.5 \text{ \AA}/12^\circ$ bin size, could not be built because the fine grid size required over 40 GB of memory to hold the scoring grid in memory during its population, exceeding my hardware limit. With each of the remaining 26 scoring grids, I rescored a docking decoy set to evaluate grid performance. Two of the 26 score grids did not complete the rescoring trials, and are omitted from my analysis.

In the benchmark, smoothing factor had little to no effect on decoy enrichment. Figure 4.1.A shows the nearly identical enrichments of near-native decoys observed in the top 10% of scores for three score grid trials using the maximization population method and only differing in smoothing factor. When aggregation is used, the effect of the smoothing factor becomes slightly more erratic, sometimes improving on the enrichment shown by the equivalent minimization grid, sometimes worsening the enrichment, and sometimes showing no difference; the magnitude of this effect is marginal, however (see Figure 4.1.B). This volatility is at least partially due to a quirk of the aggregation method which populates some commonly observed bins with exceptionally low scores (at least an order of magnitude larger than a typical low-scoring bin). As shown in Figure 4.2, these bins act as quanta, separating the decoys into distinct tiers based on the number of interactions in these low-scoring bins. Such stratification is not desirable, as a single interaction in a false binding site could overwhelm all of the contributions from interactions in the proper binding site.

In contrast to the other two factors, bin size had a larger effect on performance, as shown in Figure 4.3. Both the largest and smallest bins performed poorly. The

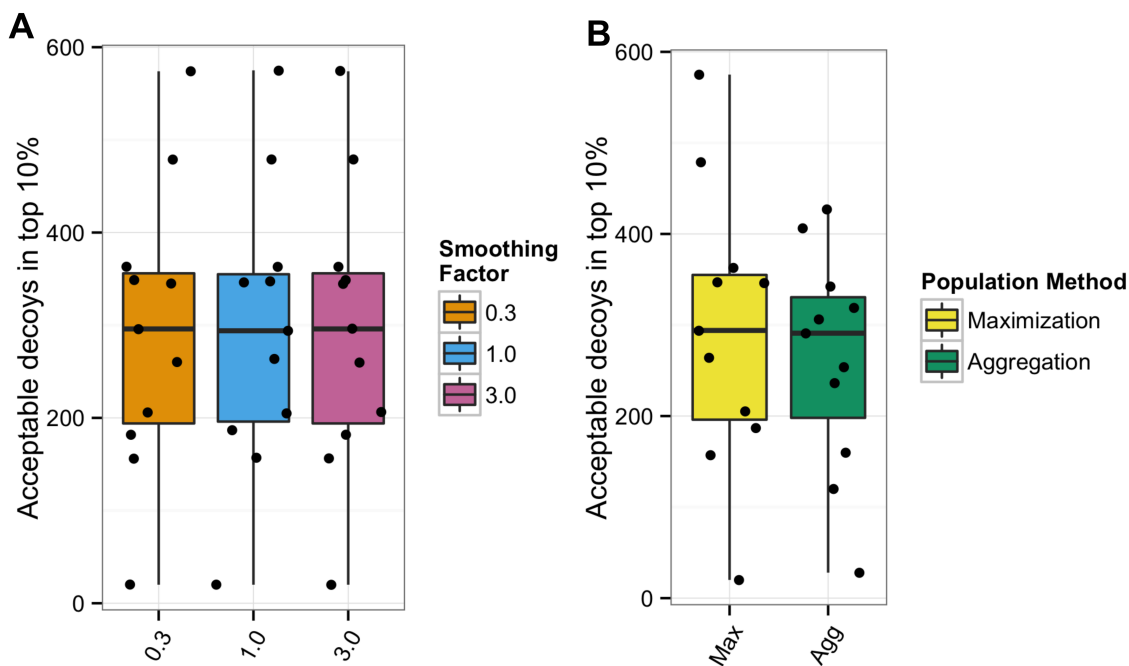


Figure 4.1: Near-native enrichment for one representative MDS score table formulation for (A) each of the three smoothing factors, and for (B) each of the two population methods. Each formulation trial is represented by a boxplot showing the enrichment of CAPRI-rating acceptable models⁵³ within the lowest-scoring 1,000 models out of a set of 10,000 (top 10%) for each of 11 protein–protein complexes. (A) All trials shown have a bin size of $2 \text{ \AA}/22.5^\circ$ and use the maximization population method. All three smoothing factors produce nearly identical near-native enrichments. (B) Both trials have a bin size of $2 \text{ \AA}/22.5^\circ$ and a smoothing factor of 1.0. The aggregation method (green fill) has erratic, but ultimately minor results on near-native enrichment.

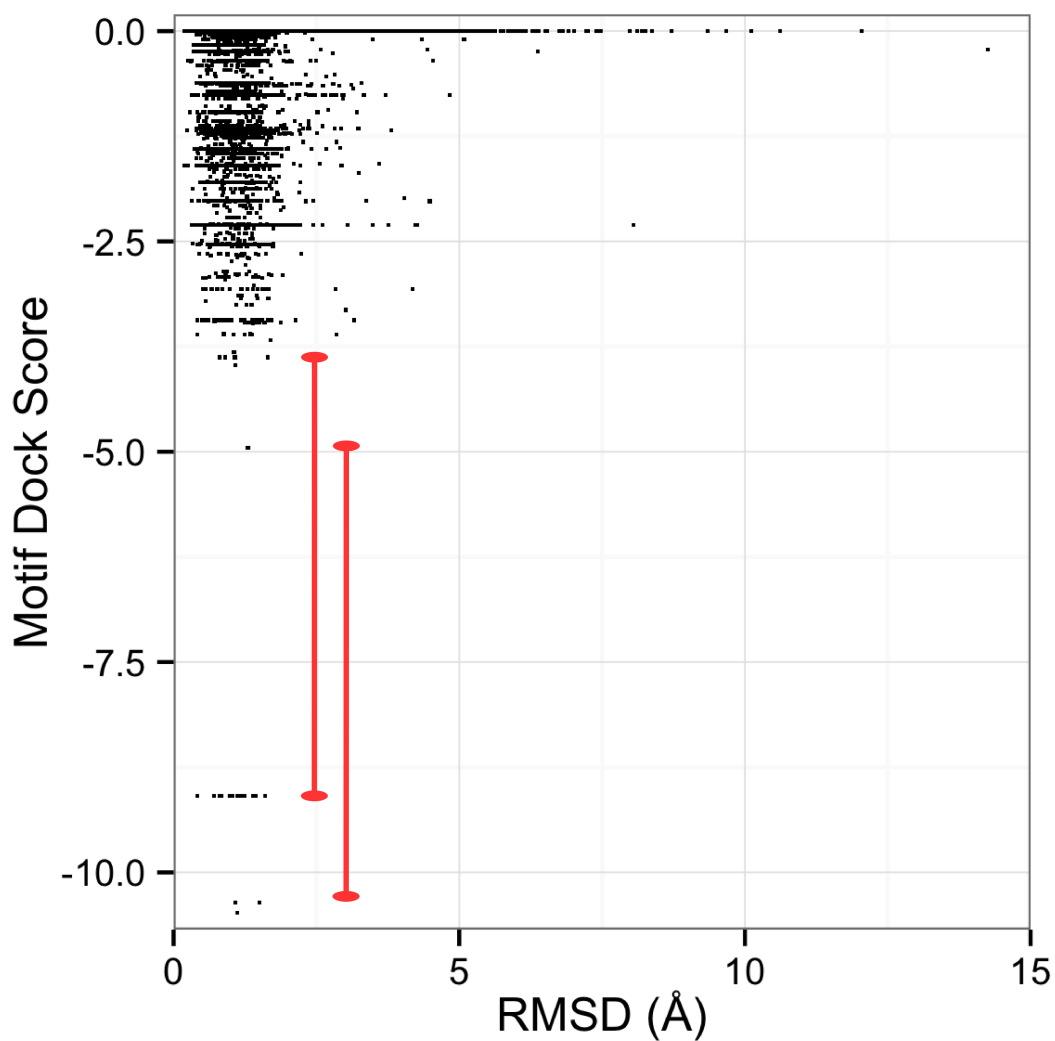


Figure 4.2: Motif Dock Score (MDS) vs. RMSD to native for all models, as scored by a score table trial with bin size = $0.5\text{\AA}/12^\circ$, smoothing factor = 0.3, and population by aggregation. Stratification of scores caused by a low number of contributing pair interactions is evident in upper left and lower left, with many models scoring identically. Large scoring gaps, caused by a single very favorable pair interaction, are highlighted in red.

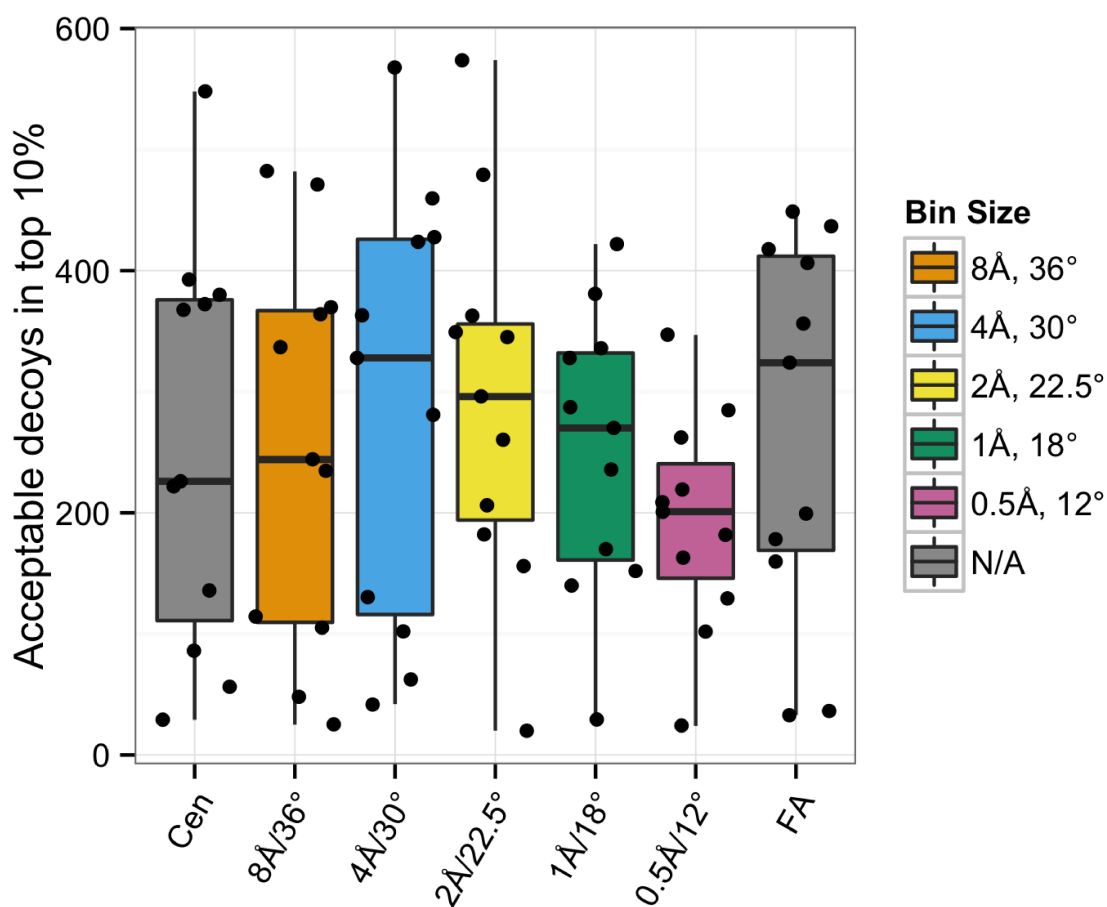


Figure 4.3: Near-native enrichment for one representative MDS score table formulation for each of the five examined bin sizes. Each formulation trial is represented by a boxplot showing the enrichment of CAPRI-rating acceptable models within the lowest-scoring 1,000 models out of a set of 10,000 (top 10%) for each of 11 protein–protein complexes. All trials shown have a smoothing factor of 1.0 and populate bins by maximization. The enrichment performance for RosettaDock’s original centroid scoring method (Cen) and the full-atom scoring method (FA) are shown in gray for comparison. The intermediate bin sizes show the best near-native enrichment.

CHAPTER 4. MOTIF DOCK SCORE

largest bin size likely convolutes too many distinct geometries into a single bin, giving it the inability to discriminate near-native decoys from non-native decoys. Similar shortcomings plague the centroid score, which represents the protein side chain with only a single pseudo-atom, and which performs almost identically to the largest bin size. Conversely, I found the smallest bin sizes create a grid that is too sparsely populated to score most pair interactions; decoys scored by the smallest bin size typically have fewer than 20 non-zero pairwise interactions. Though the highest average performance was seen in the second-largest bin size (4 Å/30°), the middle bin size (2 Å/22.5°) has an only slightly lower average, but a much higher lower quartile, with a profile closely matching the full-atom performance. Thus, I selected the 2 Å/22.5° bin size, 1.0 smoothing factor, and minimization population method as the optimal set of factors.

For my second round of optimization, I examined the underlying score function used to generate the residue pair motifs. Initially, I built motif tables from my PDB set using four score functions: `talaris2014`, the former Rosetta standard,⁶² and three iterations of the current Rosetta standard, `beta_july15`,⁶⁴ `REF15` (formerly known as `beta_nov15`),⁵ and `beta_nov16`.⁶⁴ I also examined the score cutoff for storing a motif; with the `talaris2014` score function, I tested the standard cutoff, 0 REU, as well as cutoffs of 1 REU and 2 REU to include more residue pair motifs in the motif table. From these motif tables, I built score grids using the optimal factors from the previous optimization trials. I then rescored my docking decoy set with these

CHAPTER 4. MOTIF DOCK SCORE

six score grids and compared the decoy enrichments to those of the optimal trial in the previous set. The previous set of score grids were built from a single motif set generated using the RPX minimal form of the Rosetta score function²⁹ that included only Van der Waals interactions, an implicit solvation potential, and a side chain hydrogen bonding potential.

All of the complete Rosetta score functions significantly underperformed the minimal Rosetta score function (results for `talaris2014` shown in Figure 4.4). The largest difference between the minimal score function and the `talaris2014` and `beta` score functions is the lack of single-body energy terms. These terms describe things like the likelihood of seeing certain rotamers, how well the residue fits in Ramachandran space, and corrections for zeroing the energy needed to swap amino acids in a protein sequence. I suspected that these energy terms were describing contributions that were largely redundant within the given residue-residue transform geometry, but that were nevertheless overwhelming the effects of pairwise interactions that created favorable motifs in the minimal score function implementation. To overcome these effects, I built motif tables from the same four new score functions with all single-body terms removed. I also attempted to build motif tables with higher storage cutoffs (3 REU and 6 REU) to capture a larger fraction of the relevant interactions, but the protocol quickly ran into the memory ceiling as the number of motifs per PDB structure approached the order of 10^3 , and these trials had to be aborted.

The motif sets including only two-body terms were much more successful in dis-

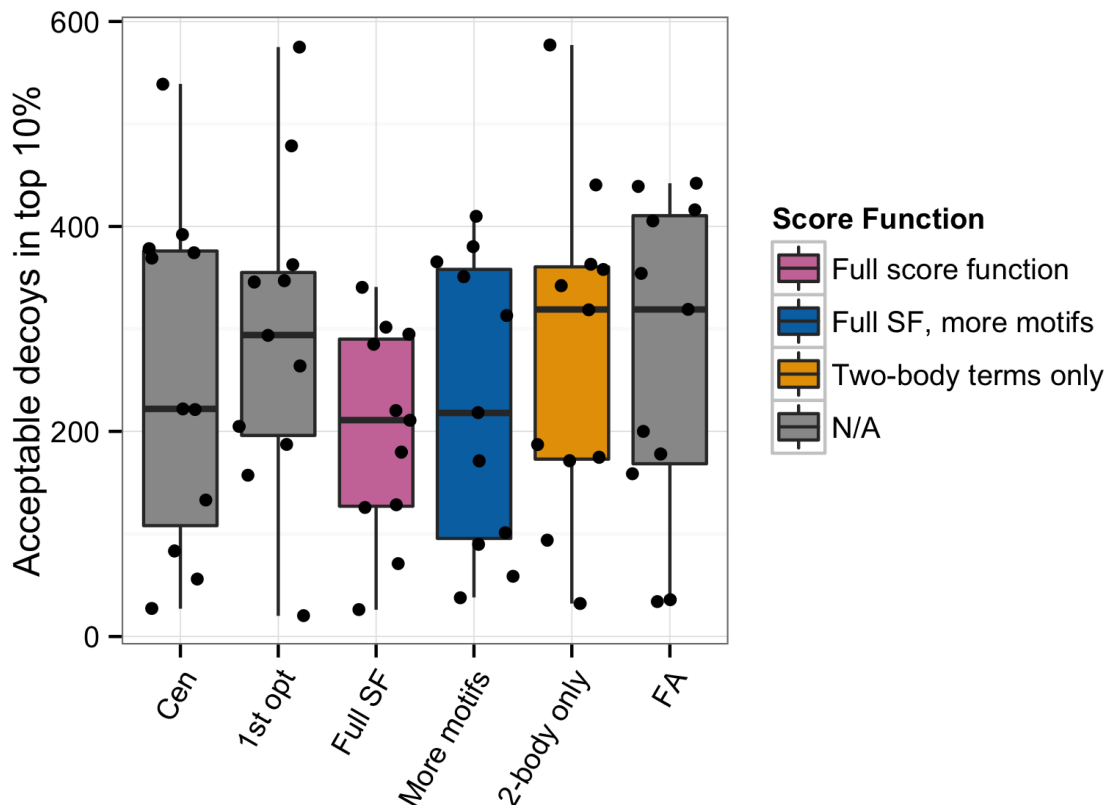


Figure 4.4: Near-native enrichment for three different `talaris2014`-based motif-generating score functions. Each formulation trial is represented by a boxplot showing the enrichment of CAPRI-rating acceptable models within the lowest-scoring 1,000 models out of a set of 10,000 (10%) for each of 11 protein–protein complexes. All trials use the optimized score table structural formulation (bin size = $2\text{\AA}/22.5^\circ$, smoothing factor = 1.0, population method = maximization). The score functions are examined in their complete state, both with a standard motif-storage cutoff (pink fill) and a more inclusive motif-storage cutoff (blue fill), as well as absent single-body score terms (orange fill). The enrichment performance for the centroid scoring method (Cen), the best trial from the first round of optimization, using the default minimal score function (1st opt), and the full-atom scoring method (FA) are shown in gray for comparison. The complete score function underperforms the minimal score function, with the enrichment improving slightly when more motifs are included. When only two-body terms are included, the `talaris2014` score function outperforms the minimal score function.

CHAPTER 4. MOTIF DOCK SCORE

criminating near-native decoys, as shown in the comparison of all `talaris2014` trials in Figure 4.4. Among the four score functions without single-body terms, The `REF15` score function, which I chose as the optimum for motif generation, had the highest average decoy enrichment of all score grids tested. These results are shown in Figure 4.5. The `REF15` is also the most recent stable, released version of the `beta` score function.⁵

To incorporate Motif Dock Score into the RosettaDock protocol, few modifications were needed. RosettaDock’s existing low-resolution phase, centroid mode, represents the protein with a full-atom backbone and a single pseudo-atom in place of the side chain. Since MDS only requires backbone atomic coordinates and side-chain identity to calculate the transform needed to identify the proper scoring bin, the centroid representation was sufficient for calculating MDS. I had previously created the Rosetta scoring term `motif_dock`, so I only needed to wrap the `motif_dock` score term into a single-term score function that could be called during low-resolution docking instead of the existing centroid scoring function.

During error testing, a shortcoming of my MDS single-term score function became evident. Because only low-scoring pair motifs are stored in the scoring grid, the single-term MDS has no knowledge of high-energy pair interactions. This absence of penalty is not important when the interaction is mildly poor. For example, if a charged residue is placed in a hydrophobic pocket, it would likely have a MDS of zero, as opposed to a full-atom score of a few positive REU; however, the energy difference between

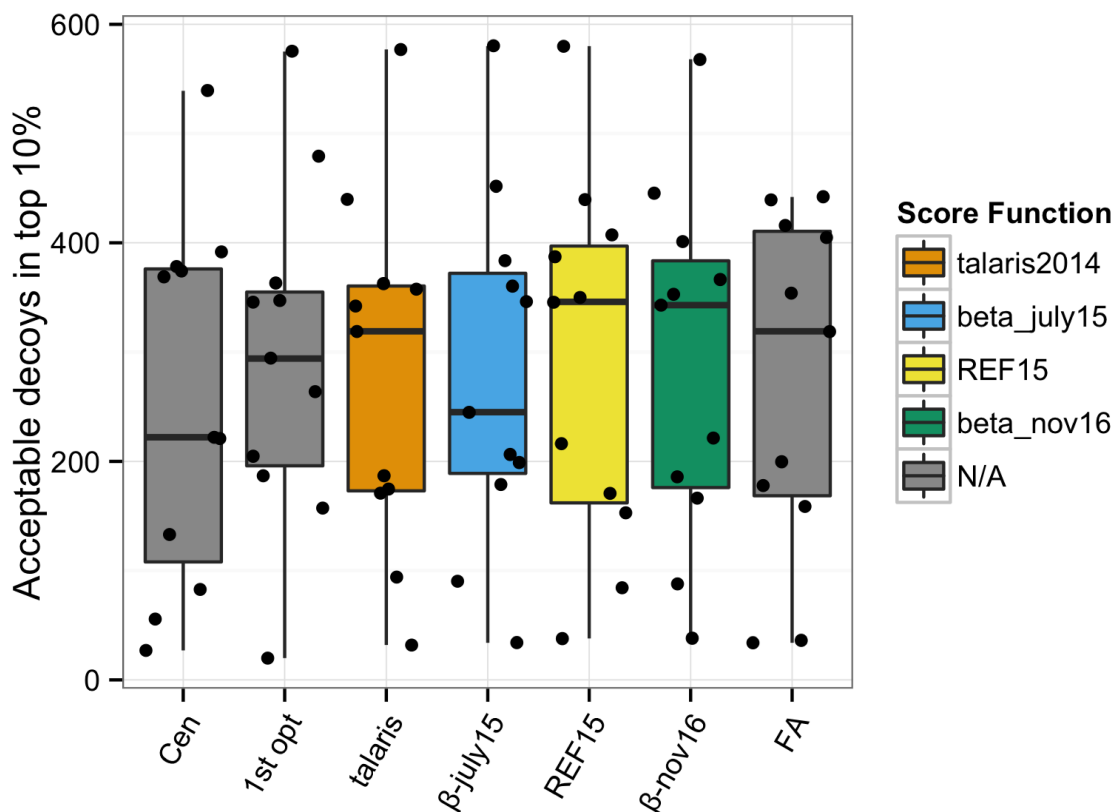


Figure 4.5: Near-native enrichment for four different motif-generating score functions. Each formulation trial is represented by a boxplot showing the enrichment of CAPRI-rating acceptable models within the lowest-scoring 1000 models out of a set of 10,000 (10%) for each of 11 protein–protein complexes. All trials use the optimized score table structural formulation (bin size = $2\text{\AA}/22.5^\circ$, smoothing factor = 1.0, population method = maximization). Four score functions are examined: `talaris2014` (orange), `beta_july15` (blue), `REF15` (yellow), and `beta_nov16` (green). The score functions are examined absent their single-body score terms. The enrichment performance for the centroid scoring method (Cen), the best trial from the first round of optimization, using the default minimal score function (1st opt), and the full-atom scoring method (FA) are shown in gray for comparison. The `talaris2014`, `REF15`, and `beta_nov16` score functions all outperform the minimal score function, with the latter two slightly outperforming even the full-atom score function in regards to near-native enrichment.

CHAPTER 4. MOTIF DOCK SCORE

such a conformation and a more favorable conformation, such as exposing the charged residue to solvent, is still the same order of magnitude it would be during full-atom scoring. The absence of penalty becomes severely detrimental in the case of very poor conformations, such as when two residues have a substantial steric clash. In such a case, the full-atom score would give an energy penalty of hundreds or thousands of REU, while the single-term MDS gives no penalty. This leads to the docking behavior seen in Figure 4.6, where some trajectories will bring the docking partners closer and

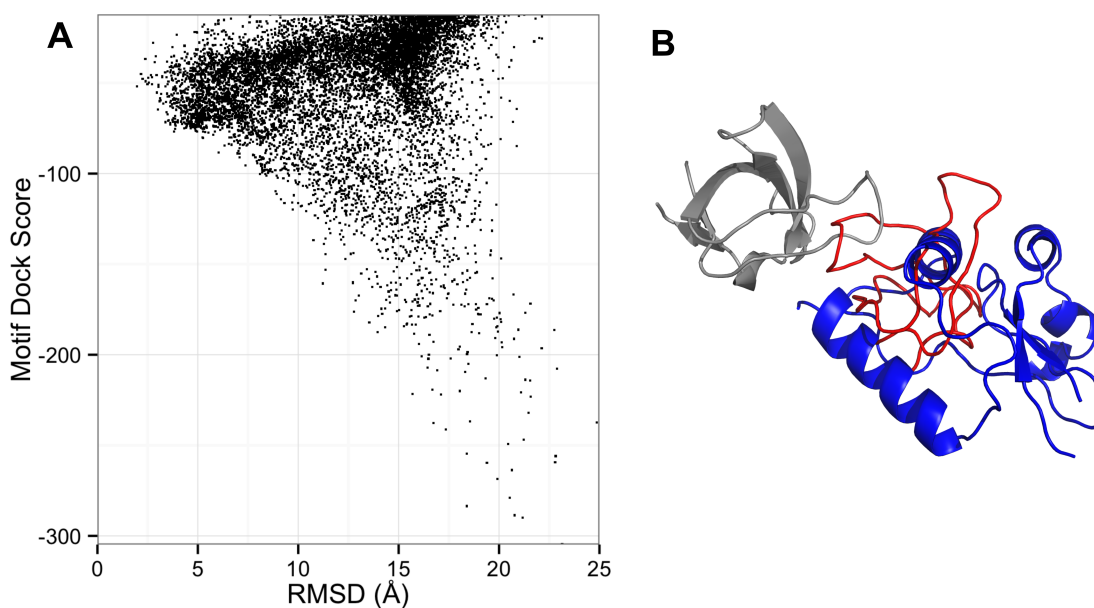


Figure 4.6: (A) Motif Dock Score (MDS) vs. RMSD-to-native for models generated by MDS protocol. Lower-scoring models are more distant from native structure, contrary to the optimal behavior. (B) Structure of a low-scoring model (red/blue) compared to the native complex (gray/blue). Absent any penalty for overlapping residues, the docking trajectory embeds the red partner inside the blue partner to maximize residue–residue contacts.

closer together to the point of embedding one in the other. These conformations score much better than realistic conformations, since placing a residue in the middle of its

CHAPTER 4. MOTIF DOCK SCORE

docking partner gives it more opportunities to find a residue partner in a favorable scoring bin without accounting for the massive penalty from all the overlapping atoms.

To compensate for the lack of a clash penalty, I tried two approaches. In the first, I simply added the centroid clash score (`interchain_vdw`) as an extra term in the MDS score function. In the second, I added a penalty (0.5 REU) for any observed motif whose geometric bin is unpopulated in the score grid. Both were successful in preventing the docking partners from embedding in each other. Testing both methods, as well as a hybrid method, on a small benchmark set revealed that the centroid clash score alone provides the best near-native decoy enrichment (Table 4.1).

Table 4.1: Enrichment performance (N100) of MDS protocol on nine targets after addition of centroid clash score, zero-count penalty, and both. N100 averages and number of N100 successes are also shown for each score addition.

	Number of models within 6 Å		
	Centroid Clash Score	MDS Penalty	Hybrid
1EFN	43	29	25
1GLA	5	2	0
1LFD	95	96	97
2A1A	58	19	12
2CFH	98	98	96
2FJU	55	2	0
2OT3	0	0	0
3AAA	2	0	0
3F1P	41	28	27
Average	44	28	27
Successes	6	2	2

4.4.2 Benchmarking Motif Dock Score’s Performance

To evaluate the docking accuracy of MDS, I compared its performance against a baseline method, the current centroid low-resolution docking mode, on a nine-target representative benchmark set. I also evaluated the performance of MDS in concert with Adaptive EnsembleDock, a more-efficient conformational sampling method that produces larger conformer ensembles for each of the docking partners than the standard ensemble generation method. Previously, EnsembleDock was able to sample 10 backbones per partner at a computational cost of 1-2 minutes per model. Adaptive EnsembleDock provides a significant speed-up, allowing me to sample 100 backbone conformations per partner in similar time. As a second control, I also evaluated the performance of Adaptive EnsembleDock alone. For each of the four conditions, I ran a full local EnsembleDock, generating 10,000 decoys per target. The docking runs are evaluated for near-native discrimination by the N5 metric, and for near-native enrichment by the N100 and N1000 metrics, both after the low-resolution phase and after the full protocol (see Section 4.3.7 for details). These metrics are detailed for all docking runs in Table 4.2.

Table 4.2: Bootstrapped N5, N100, and N1000 metrics for four docking protocols: (1) Centroid protocol, (2) Adaptive EnsembleDock protocol, (3) MDS Protocol, and (4) Hybrid MDS/Adaptive EnsembleDock protocol. N5/N100/N1000 metrics are calculated for both the low-resolution and high-resolution phases of the protocols. Average metrics across nine benchmark targets are recorded in table, as are the summed probabilities of success.

		Low-Res			High-Res		
		N5	N100	N1000	N5	N100	N1000
Average	Centroid	0.3	21	220	2.3	44	305

CHAPTER 4. MOTIF DOCK SCORE

Table 4.2 ... continued

		Low-Res			High-Res		
		N5	N100	N1000	N5	N100	N1000
	Adaptive EnsembleDock	0.3	14	155	2.0	41	300
	MDS	2.1	45	349	2.1	41	334
	MDS/AED	2.2	42	339	2.3	46	334
Expected Successes (Out of 9)	Centroid	0.4	2.0	4.0	3.9	5.0	5.0
	Adaptive EnsembleDock	0.4	2.0	3.0	3.7	4.6	5.0
	MDS	3.6	5.9	6.0	3.7	5.3	6.9
	MDS/AED	3.8	5.5	6.0	4.3	5.2	7.0

An example of a successful docking simulation, the MDS run for target 1LFD, is shown in comparison to an unsuccessful docking simulation, the centroid run for target 1LFD, in Figure 4.7. All candidate structures generated by the low-resolution

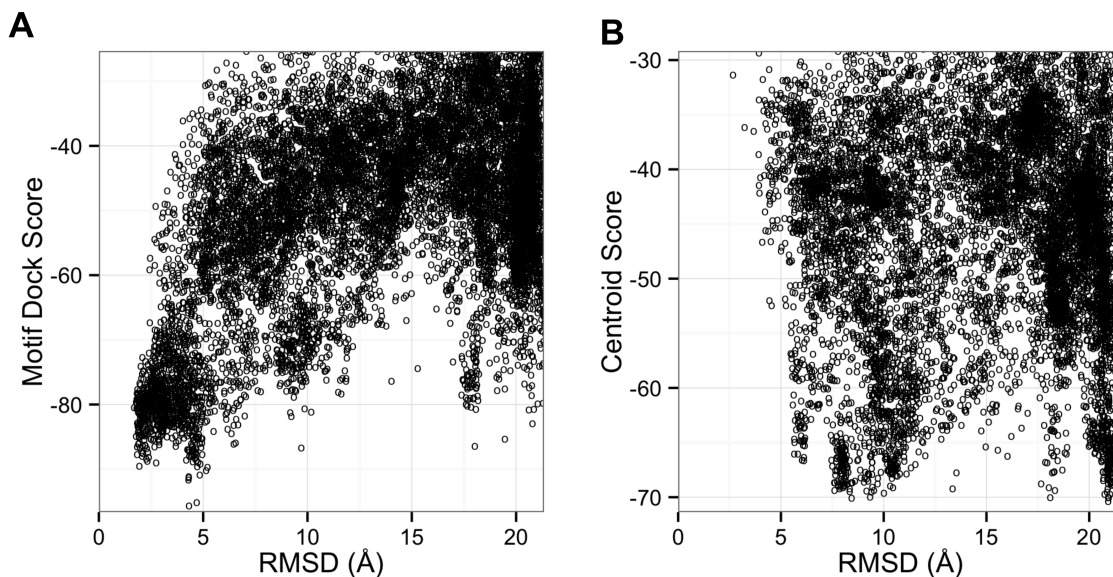


Figure 4.7: Low-resolution score vs. RMSD to native for target 1LFD. (A) 10,000 models generated by MDS protocol. MDS generates a large number of near-native candidate structures, and discriminates them from incorrect models. All metrics indicate success: N5 = 5, N100 = 95, N1000 = 750. (B) 10,000 models generated by centroid protocol. Centroid score does not generate many near-native candidate structures, and it cannot distinguish them from incorrect models. All metrics indicate failure: N5 = 0, N100 = 0, N1000 = 23.

CHAPTER 4. MOTIF DOCK SCORE

phase of docking are plotted, comparing their low-resolution score to their RMSD values. In the MDS run (Figure 4.7.A), a clear "funnel" can be seen in the plot, with the lowest-scoring models being near-native. This plot corresponds to the metrics $N5 = 5$, $N100 = 95$, and $N1000 = 750$, all strong successes. This behavior is optimal in a blind docking prediction, where only the score is available for discrimination of candidate structures. Conversely, in the centroid run (Figure 4.7.B), no funnel is observed; the lowest-scoring models are nearly all incorrect, and few near-native models are sampled at all. This plot corresponds to the metrics of $N5 = 0$, $N100 = 0$, and $N1000 = 23$, all failures.

The behavior seen in Figure 4.7 is repeated throughout the benchmark, with the largest performance gains observed in the low-resolution metrics when MDS is added to the baseline protocol. The expected number of $N5$ successes increases from 0.4 to 3.6 (out of 9), showing a ninefold improvement in near-native discrimination. Perhaps more notably, though, the expected number of $N100$ successes rises from 2.0 to 5.9 (out of 9), and the MDS $N1000$ metric shows nearly 7 successes (out of 9), missing only 2OT3 and 3AAA. 2OT3 is a difficult target whose internal motions upon binding prevent MDS from sampling decoys closer than 6 Å RMSD, however a low-scoring structural cluster appears around 8 Å that contains several models (~ 20) that maintain 10% of the native contacts, the threshold for an acceptable model in the CAPRI experiments.⁵³ 3AAA is a medium-difficulty target for which near-native decoys are present in the top 1000 (and the top 100), but not at a high

CHAPTER 4. MOTIF DOCK SCORE

enough enrichment to count as a success. These results suggest that MDS can be a useful docking filter, with both its top 10% and top 1% of decoys showing successful enrichment of near-native decoys in a majority of targets.

Conversely, Adaptive EnsembleDock shows no broad improvements over the baseline RosettaDock, with the two protocols exhibiting similar N5, N100, and N1000 metrics. Likewise, the combined MDS/Adaptive EnsembleDock protocol performs similarly to the MDS protocol alone. Anecdotally, however, there are indications that Adaptive EnsembleDock is occasionally sampling near-native conformations more effectively, most visibly in the high-resolution N5 metric for target 3F1P, a difficult target. In the MDS protocol, the high-resolution N5 has an average value of 0.6, with a 0.0 probability of success; when Adaptive Ensemble Dock is added to make the hybrid protocol, the N5 value jumps to 3.1, corresponding to an 0.7 probability of success. Plots of the 3F1P results of the two protocols (Figure 4.8) reveal that the hybrid method has a score funnel centered at approximately 3.5 Å RMSD, just below the N5 cutoff, as opposed to the MDS score funnel centered at approximately 5.0 Å RMSD, just above the N5 cutoff.

Plots of all target results from all four protocols, divided by target difficulty, can be found in Figures 4.9 (rigid-body), 4.10 (medium), and 4.11 (difficult).

Because future docking protocols will require a low-resolution phase that is both discriminating and fast, I also evaluated the computational overhead for all four protocol conditions. The baseline protocol, MDS protocol, and hybrid protocol all

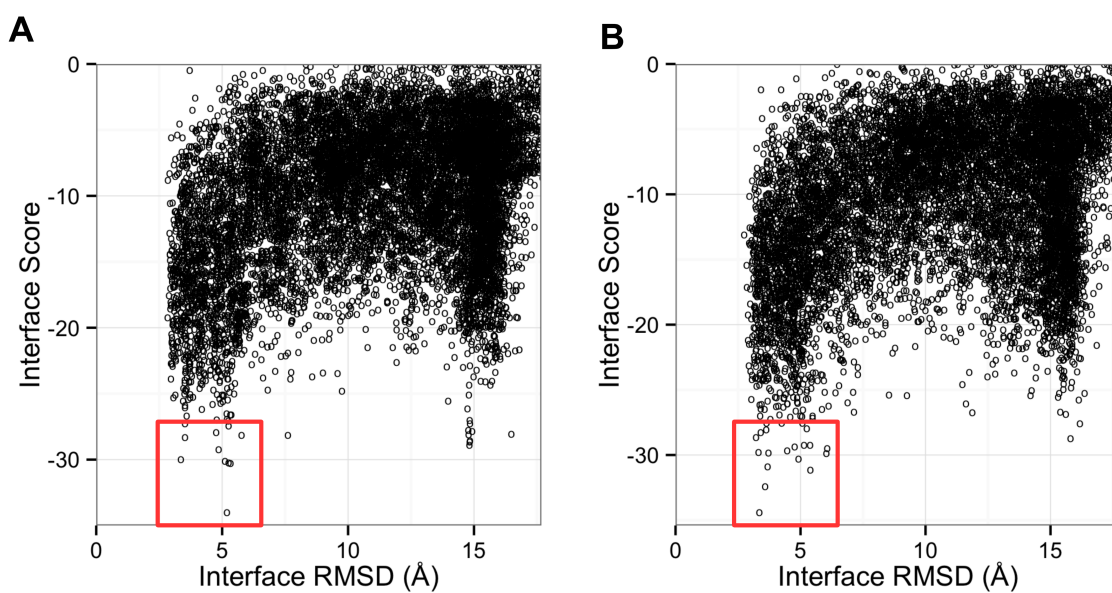


Figure 4.8: Interface Score vs. Interface RMSD to native for target 3F1P. (A) 10,000 models generated by MDS protocol. $N_5 = 0.6$ (failure). (B) 10,000 models generated by hybrid MDS/Adaptive EnsembleDock protocol. $N_5 = 3.1$ (marginal success). The low-scoring region of interest is highlighted, showing the marginally better discrimination by the hybrid protocol.

CHAPTER 4. MOTIF DOCK SCORE

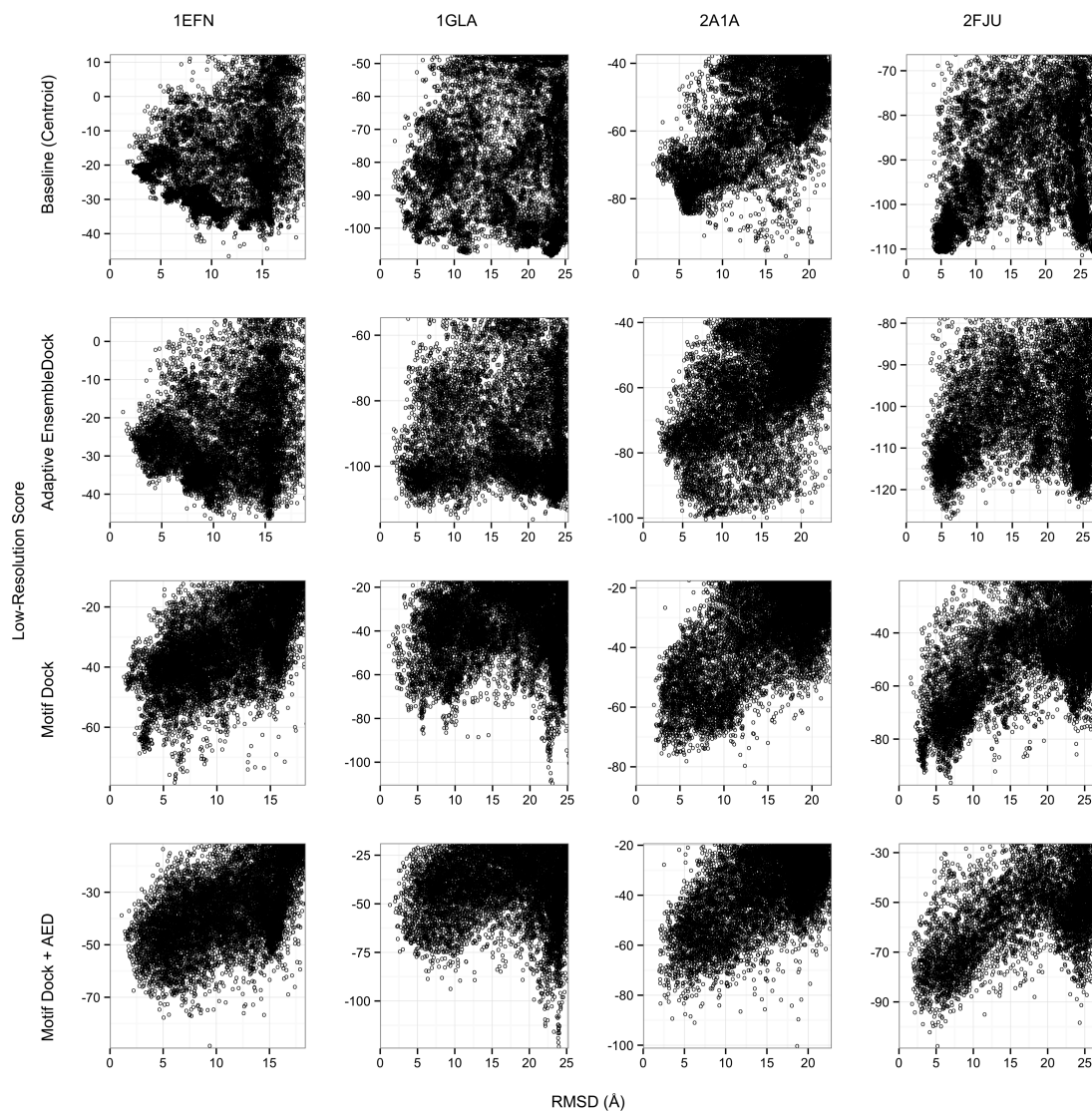


Figure 4.9: Score vs. RMSD for 10,000 models for each of four rigid-body (easy) benchmark targets. Results are shown for (1) Centroid protocol, (2) Adaptive Ensemble Dock protocol, (3) MDS Protocol, and (4) Hybrid MDS/Adaptive Ensemble Dock protocol.

CHAPTER 4. MOTIF DOCK SCORE

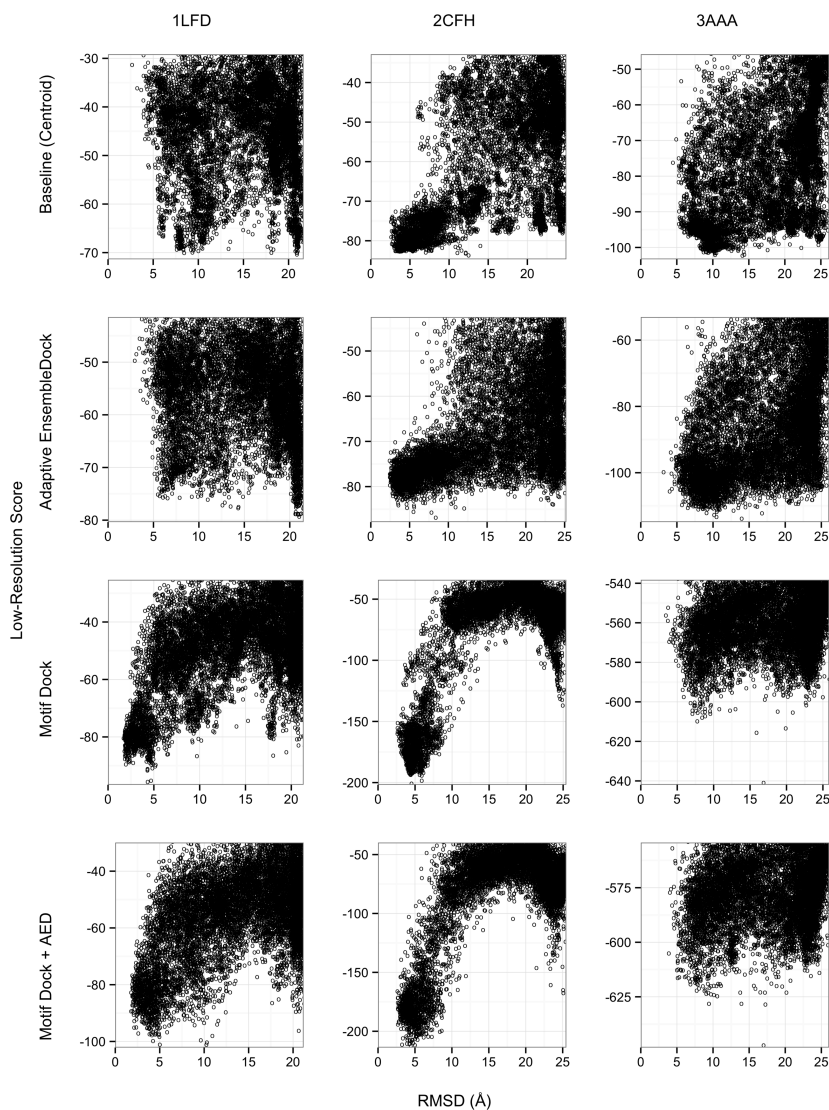


Figure 4.10: Score vs. RMSD for 10,000 models for each of three medium benchmark targets. Results are shown for (1) Centroid protocol, (2) Adaptive Ensemble Dock protocol, (3) MDS Protocol, and (4) Hybrid MDS/Adaptive Ensemble Dock protocol.

CHAPTER 4. MOTIF DOCK SCORE

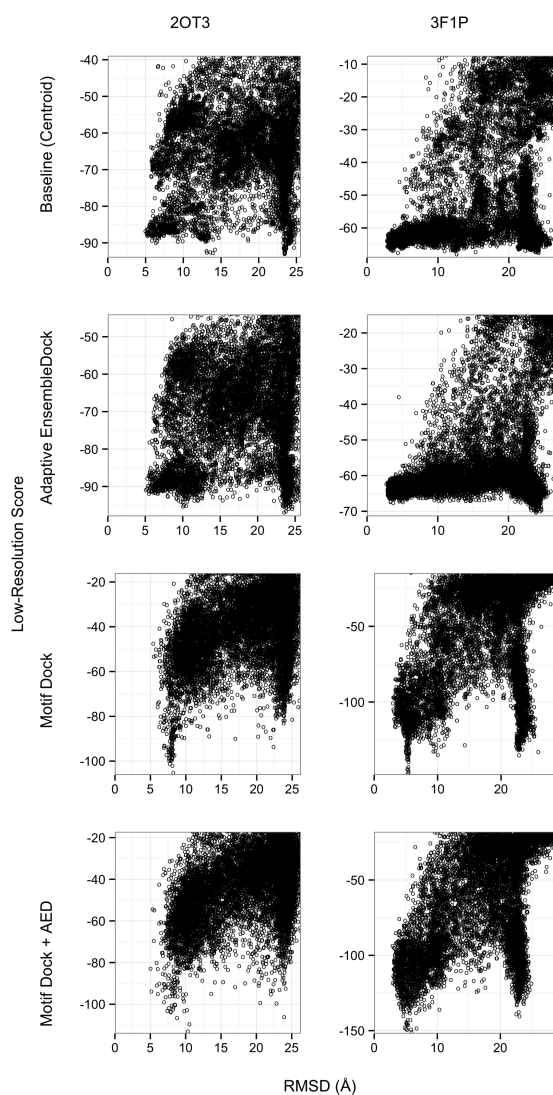


Figure 4.11: Score vs. RMSD for 10,000 models for both difficult benchmark targets. Results are shown for (1) Centroid protocol, (2) Adaptive Ensemble Dock protocol, (3) MDS Protocol, and (4) Hybrid MDS/Adaptive Ensemble Dock protocol.

CHAPTER 4. MOTIF DOCK SCORE

have approximately equal runtimes. The Adaptive EnsembleDock protocol requires about twice the runtime as the other three protocols, likely due to a higher rate of conformer swap rejection. MDS and Adaptive EnsembleDock both have additional memory overhead. The MDS protocol requires about 2 GB of memory to store the score table, while the Adaptive EnsembleDock protocol requires anywhere from 200 MB to 1 GB of extra memory to store the larger protein ensembles, dependent on the size of the protein. The extra memory requirements are roughly additive, with the hybrid protocol requiring approximately 3 GB for a small target and 4 GB for a medium-sized target, compared to only 0.6 GB and 1.1 GB respectively for the baseline protocol.

Encouraged by the promise MDS showed in local docking simulations, I also tested MDS in global docking simulations. Bootstrapping of the local docking results suggested that MDS may be able to perform similarly to a full local docking simulation with as few as 500 decoys, with only one target’s funnel degrading at that decoy count. As such, I ran MDS’s global simulations with a low decoy count of 10,000, roughly 1%-0.1% of the decoys of a typical global docking run. The global docking results, as displayed in Figure 4.12, clearly show that 10,000 decoys were not sufficient to produce thorough global sampling, with no more than a handful of sub-10 Å decoys generated for any target. The results also show, however, that global docking with MDS recapitulates the same energy landscape observed in local docking, and that four targets do not find a well populated lower-scoring false energy funnel. This sug-

CHAPTER 4. MOTIF DOCK SCORE

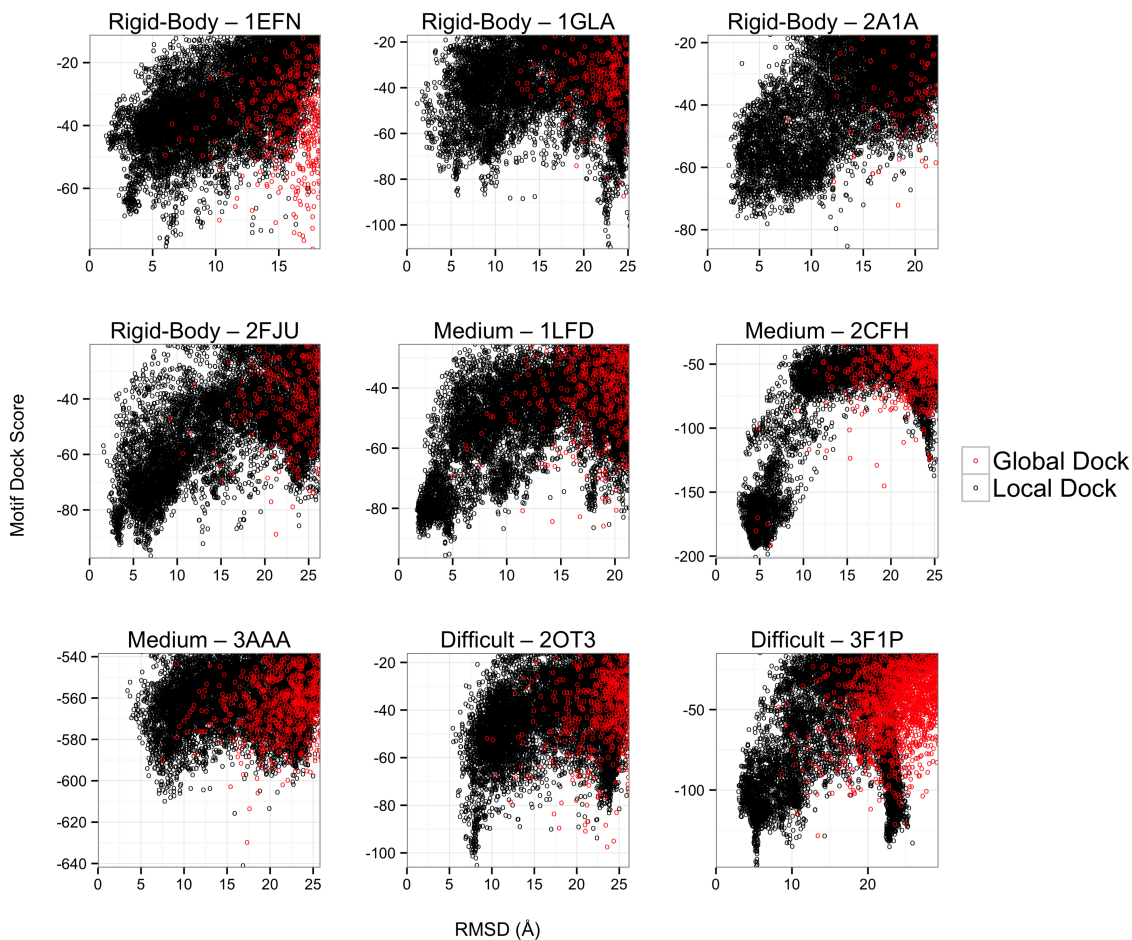


Figure 4.12: MDS vs. RMSD to native for each of nine benchmark targets, arranged by target difficulty. 10,000 models generated by local MDS protocol are shown in black; 10,000 models generated by global MDS protocol are shown in red. The global sampling finds the same energy landscape as the local sampling, but rarely finds near-native structures with the same total number of models

gests that a standard MDS global docking run might produce and enrich near-native decoys in many targets.

4.5 Discussion

Computational protein-protein docking can be confounded by a diverse set of problems, including backbone flexibility upon binding, global docking searches, and lack of structural knowledge of the docking partners. Within the Rosetta framework, however, all of these problems can potentially be addressed by intelligently increasing the conformational sampling space of the docking protocol. Sampling increases must be offset with efficiency gains, however, to prevent the computational costs from exploding. Motif Dock Score (MDS), by virtue of its fast and accurate scoring, will provide the necessary efficiency offsets by filtering the larger model sets and passing only a small fraction to the expensive high-resolution phase. As such, MDS will be a critical component of future enhancements to the RosettaDock protocol.

The MDS protocol that I created shows a marked improvement in accuracy over centroid scoring. MDS triples the number of targets in which the top 1% of models are significantly enriched with near-native structures, and it is seven to nine times as effective for discriminating top models, as evidenced by the bootstrapped N5 metrics. More generally, MDS captures nearly all of the discriminatory power of the full-atom score function upon which it is based, exhibiting similar low-resolution and

CHAPTER 4. MOTIF DOCK SCORE

high-resolution N5, N100, and N1000 metrics. Most importantly for a low-resolution score function, MDS achieves these gains in accuracy without sacrificing computational efficiency, running in roughly equivalent time to the centroid scoring method. The Adaptive EnsembleDock method was less successful, showing minimal gains in simulation accuracy in only a small number of docking targets.

MDS also compares favorably to other docking protocols. Table 4.3 describes recent published results from five leading docking methods: HADDOCK,⁷⁷ iATTRACT,⁶⁷ ClusPro,⁴⁴ ZDOCK,⁶⁵ and RosettaDock (v3.2).²⁰ While the methods have different scopes and report their results in different forms, I was able to assign an N# success metric (analogous to N5, N100, etc.) to each method. In general, the methods are good at docking easy, rigid-body targets ($\sim 50\%$ accuracy or better), but they are all poor when the targets become more flexible ($< \sim 30\%$ accuracy on difficult targets). Albeit on a small benchmark set, MDS maintains this level of accuracy for easy targets (100%) while showing dramatically improved accuracy for difficult, flexible targets (60%). Given these drastic gains, MDS is potentially the first step toward a paradigm shift in protein-protein docking where complexes with moderate backbone flexibility become tractable docking targets.

Table 4.3: Comparison of five leading docking methods and MDS. *Nearest-native structures from rigid-body docking selected for refinement. †Half successes awarded for targets with multiple binding sites evaluated, where at least one but not all binding sites are captured. ‡2.5Å cutoff for near-native structures. §For Docking Benchmark sets, includes both medium and difficult targets. For CAPRI sets, includes targets without at least one high-quality prediction by any predictor.

Method	Description	Flexibility?	Benchmark Set	Docking Search	Success Metric	Performance		
						All Targets	Easy Targets	Difficult Targets§
HADDOCK ⁷⁷	Restraint-based docking and minimization	Yes	CASP-CAPRI	Mixed global/local	N10 = 1	16/25 (64%)	12/12 (100%)	4/13 (31%)
iATTRACT ⁶⁷	Rigid-body docking, interface refinement	Yes	Docking Benchmark 4.0	Global*	N200 = 30	64/166 (39%)	55/119 (46%)	9/47 (19%)
ClusPro ⁴⁴	FFT docking, cluster evaluation	No	CAPRI Rds. 13–35	Mixed global/local	N10 = 1	19/42 (45%)	12.5†/16 (78%)	6.5†/26 (25%)
ZDOCK ⁶⁵	FFT docking, model evaluation	No	Docking Benchmark 4.0	Global	N100 = 1‡	65/176 (37%)	58/121 (48%)	7/55 (13%)
RosettaDock ²⁰	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 4.0	Local	N5 = 3	56/115 (49%)	49/84 (58%)	7/31 (23%)
MDS	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 5.0	Local	N1000 = 150	7/9 (78%)	4/4 (100%)	3/5 (60%)

Chapter 5

Conclusion

5.1 My Contributions

Protein–protein interactions underlie countless physiological processes, and they contribute to diseases as diverse as diabetes, cancer, Alzheimer’s, and viral infections. As humanity learns more about the structure of protein–protein interactions, critical biological pathways like the immune system will become better understood, new therapies and cures will become evident, and human health will improve. While experimental techniques are valuable to the determination of protein complex structures, computational modeling is an important alternative when experimental techniques are too slow, too expensive, or infeasible for certain targets. For computational techniques to have real-world utility, they must be as accurate and as high-throughput as possible, and they should, *in toto*, be able to address a wide range of problems. To

CHAPTER 5. CONCLUSION

this end, my graduate research has focused on developing and evaluating new computational methods that improve the fidelity, speed, and scope of protein structure prediction.

My thesis work divided broadly into three areas of focus, each highlighted by a thesis chapter. Chapter 2 outlined my work toward improving antibody homology modeling. Motivated by the weaknesses in predicting or even defining antibody V_L - V_H orientation uncovered by the Second Antibody Modeling Assessment,⁸⁸ I designed a four-coordinate geometric framework, LHOc, that can be used to define an empirically unambiguous V_L - V_H orientation for any antibody structure; I also developed a metric, OCD, that measures the distance in LHOc geometric space between any two antibody structures. Incorporating LHOc and OCD, I built the multiple-template grafting method for RosettaAntibody, tripling the protocol's accuracy in prediction of V_L - V_H orientation and surpassing the prediction accuracy of the competing method, ABangle. My contributions are included in the current standard version of the RosettaAntibody protocol.⁸⁷

In Chapter 3, I described my work designing and testing sampling methods for protein-protein docking. After observing that RosettaDock's global search poorly sampled oblong proteins, I designed Ellipsoidal Dock, a new global sampling method in which the docking partners are approximated as ellipsoids rather than spheres. My work on the CAPRI blind prediction challenges proved the utility of Ellipsoidal Dock, with the protocol producing acceptable or better models for two complexes of oblong

CHAPTER 5. CONCLUSION

proteins, the structures of which eluded most groups. During my work with CAPRI, I also uncovered a number of existing weaknesses in the RosettaDock protocol and community-wide blind spots in protein–protein docking, highlighting the need for advanced docking protocols that can handle large complexes, global searches, and backbone flexibility.

Chapter 4 detailed my work on a low-resolution docking score function, Motif Dock Score (MDS). I built MDS using the motif score framework, a utility for querying features of crystal structures and hashing them into a geometrically indexed table, and optimized it for enrichment of near-native structures during low-resolution docking. MDS fully replaces the old low-resolution centroid score function³³ in RosettaDock, requiring less time to run and selecting near-native decoys at a rate comparable to the high-resolution full-atom score function, with the only drawbacks being an increased memory footprint and a one-time calculation of a motif database. The increased calculation speed of MDS makes feasible docking protocols requiring decoy counts one or two orders of magnitude higher than current protocols, setting the table for future protein–protein docking efforts to address the limitations observed in Chapter 3.

5.2 Future Directions

Despite recent advances, there is still much work to do in the field of protein structure prediction. There are a number of approaches that I believe have the potential to improve the fidelity of antibody structure prediction and the scope of protein-protein docking prediction. I will also highlight a few community-wide issues that, if addressed, will increase the speed of advances in the field.

5.2.1 Antibody Homology Modeling

Homology-based antibody modeling is limited by three stages: prediction of the V_L - V_H orientation, prediction of the CDR loops, and prediction of the antibody-antigen binding mode.

5.2.1.1 V_L - V_H Orientation

In regards to the V_L - V_H orientation, while my LHOC coordinate frame and multiple-template grafting mode have improved the sample space of the cross-domain docking, they have only partially solved the identification of the correct V_L - V_H orientation. Though multi-template grafting ensures the domain orientations are quite diverse, they typically score in similar ranges, making it difficult to distinguish the correct orientation from the incorrect ones.

Some of this ambiguity is due to an insufficient score function. RosettaAntibody

CHAPTER 5. CONCLUSION

was last benchmarked before MDS was developed and before the REF15 score function⁵ became the Rosetta standard; a new benchmark with these advanced score functions would likely show at least an incremental increase in prediction fidelity. A larger portion of the ambiguity, however, is likely the fault of our template selection method. As shown in Chapter 2, the V_L - V_H orientation of decoys is heavily biased by the V_L - V_H orientation of the original grafted template. Since we do not often select orientation templates that match the native antibody structure, our decoy pool is often dominated by structures with incorrect V_L - V_H orientations. Thus, better prediction of V_L - V_H orientation is required.

I believe that a comprehensive machine-learned classifier is the most logical approach to V_L - V_H orientation prediction. While the null approach would simply train on full F_V sequences, there are a number of other, more intelligent feature options. In particular, the V -genes of the light and heavy chains are likely of critical importance to the V_L - V_H orientation, as they have well-conserved structures that form the majority of the V_L - V_H interface. The CDR lengths, especially the H3, are also intriguing features, since the breadth of the paratope is influenced by the sterics of the CDR loops. Other features, such as species, degree of humanization, and isotype should also be considered. With a large number of features, given the sparseness of the antibody structural data set, there exists a possibility that the machine-learned classifier will be overfitted. Counteracting this tendency, however, is easily achieved with the forced diversification, such as the one encoded in the multiple-template grafting pro-

CHAPTER 5. CONCLUSION

tol. The classifier could be forced to make ten distinct template predictions using different subsets of the feature set, mitigating any overfitting effects.

Though better V_L-V_H orientation prediction will improve our modeling fidelity, the hypervariable nature of antibodies puts a limit on how effective a sequence-based prediction can be: a template-based method can only predict orientations for which a template exists. The resulting gaps in structural space must be filled with other methods. The RosettaAntibody protocol fills these gaps using random-magnitude and -direction Monte Carlo moves. While such moves do fill the missing structural space, they do not do so efficiently, with a significant fraction of the accepted moves placing the domains into non-antibody-like orientations. A more efficient method would sample only the appropriate orientational space, easily defined by the set of all known antibody structures. Instead of sampling in random directions and magnitudes, the Monte Carlo moves would be directed along the principal components of V_L-V_H orientational space. A larger number of reasonable conformations would be sampled in the same computational time, enhancing the efficiency of the protocol. This method could be implemented using the LHOC framework, but it would be better implemented on a geometrically complete V_L-V_H orientation framework such as ABangle³⁷ where the principal components could be easily translated to xyz space.

5.2.1.2 CDR Prediction

Ultimately, the V_L - V_H orientation must also be predicted in congress with the second impediment to antibody modeling, prediction of CDR conformations. Broadly, CDR prediction breaks down into the easy problem of canonical CDR prediction, which has been largely solved, and the hard problem of H3 prediction, which is still largely unsolved. Canonical CDRs are predicted correctly by RosettaAntibody about 80% of the time.⁸⁸ While incremental improvements in sequence-based prediction will eventually be made, I believe the simpler path forward is to assign accurate confidences to CDR predictions. If the incorrect predictions are not distinguished from correct predictions in confidence, there is no way to improve upon the incorrect predictions without affecting the correct predictions. Conversely, if the correct predictions have a high confidence and the incorrect predictions have a low confidence of matching their predicted conformations, the low confidence CDRs become prime targets for additional structural prediction, such as scoring multiple canonical conformations in situ or modeling the CDR de novo. These confidence values could be assigned simply using BLAST scores, or more complexly using an HMM with knowledge of key amino acid-position pairs for each canonical CDR structure. Assignment of accurate confidence values combined with structural remodeling of the low-confidence CDRs should lead to improved fidelity of canonical CDR prediction.

Unlike the canonical CDRs, the H3 has no clear path forward for improved predictions that I can see. H3 modeling is the quintessential protein modeling problem:

CHAPTER 5. CONCLUSION

native-like conformations are rarely sampled, and when they are, they are not easily discriminated from non-native conformations. A great leap forward is needed to truly solve H3 prediction. In the meanwhile, incremental gains can likely be made by increasing the speed at which H3 conformations can be sampled and scored, allowing for a larger search space to be explored. A modified version of the MDS score function is a good first candidate for such a speed-up. The modified score function would need to calculate not just cross-chain residue pairs, but also intrachain loop-framework and cross-loop residue pairs; it would also need to include single-body score terms for the H3 residues. Such a score function should be developed and benchmarked in the context of RosettaAntibody.

5.2.1.3 Antibody–Antigen Docking

The final hindrance to antibody modeling, antibody–antigen docking, provides an interesting test case for the future of protein–protein docking. To date, protein–protein docking has largely focused on docking one partner, made up of one or more peptide chains, to another partner, also made up of one or more peptide chains. The protein complexes modeled successfully generally exhibit little rearrangement upon binding, with no topology changes or large loop motions. Furthermore, the proteins are generally part of some native pathway, meaning they have co-evolved and typically have a good degree of shape complementarity, giving insight into the binding site and restricting the docking search space. Antibody–antigen docking breaks all

CHAPTER 5. CONCLUSION

of these rules: the antigen binds to the interface of two antibody domains, making antibody–antigen docking a three-body problem instead of the normal two-body problem; antigens don’t co-evolve with antibodies, and the complex shape complementarity is entirely determined by transient loops on the antibody, leaving little indication of the binding site; and antibodies can exhibit large H3 loop motions upon binding to the antigen. Each of these three problems, multi-body docking, global docking, and flexible docking, represents a standing challenge in protein–protein docking; to achieve consistent antibody–antigen docking, all three of these challenges must be addressed. Fortunately, through work presented in this thesis and advances elsewhere, Rosetta is primed to tackle all of these challenges in a next-generation docking protocol.

5.2.2 Protein–Protein Docking

5.2.2.1 Multi-Body Docking

Rosetta does not currently support multi-body docking, but just a few adaptations are needed to make it possible. The biggest of these is a required update to the FoldTree framework. The basic FoldTree implementation defines each peptide chain’s primary structure as an **edge**; the edges are connected in series by **jumps**, which are virtual rigid ties that preserve the peptide chains’ relative orientation.⁸¹ During docking, one of these rigid jumps is made flexible, allowing one half of the complex to move relative to the other half. This implementation is insufficient in

CHAPTER 5. CONCLUSION

a multi-body context; while the firstmost and lastmost protein partners would be able to move independently, any middle partners would not be able to move relative to the rest of the complex, as their motions would propagate to the chains either upstream or downstream. Current Rosetta protocols for pseudo-multi-body docking (e.g. SnugDock) avoid this problem by creating a new FoldTree each time a different protein partner moves relative to the complex; however, this behavior is not multi-body docking, but iterative two-body docking.

True multi-body docking can be achieved with a new type of FoldTree. This "Universal" FoldTree, rather than connecting chains in series, would function more like a hierarchical tree. An antibody-specific example of the Universal FoldTree is compared to the base Rosetta FoldTree in Figure 5.1. The Universal FoldTree will have a root node representing either some point in space, or some relative point in the complex, like the center of mass. The root node will be connected by jumps to daughter nodes representing the independent components of the complex; the daughter nodes can in turn be connected by jumps to their own daughter nodes, representing subunits of the independent components. For example, an antibody-antigen complex may have a Universal FoldTree where the root has two daughters, the antigen and the antibody, and the antibody has two daughters, the light and the heavy chain. This hierarchy is infinitely extensible to as many docking partners and/or as many tiers of independence as needed. Using this Universal FoldTree, each partner would be able to move independently relative to the rest of the complex. Additionally,

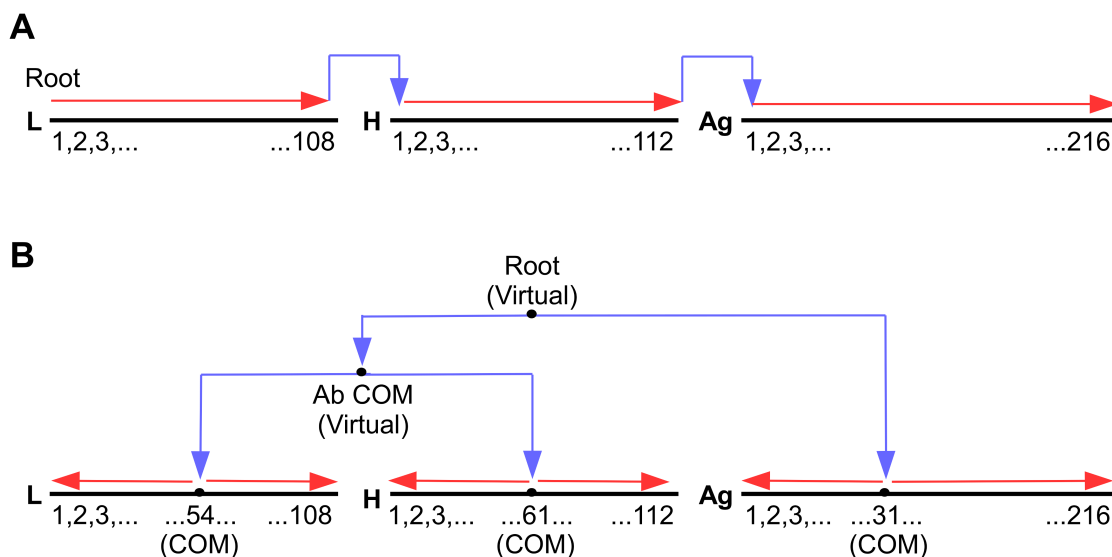


Figure 5.1: Representation of (A) the base Rosetta FoldTree and (B) the Universal FoldTree for the antibody–antigen docking case. The path of traversal and propagation of motion is shown by the direction of the arrows, with protein edges shown in red and jumps shown in blue. (A) In the base FoldTree, the root is residue 1 of the first protein chain. In antibody–antigen docking, motions propagate downstream from the first residue of the light chain to the last, then to the heavy chain, then to the antigen. The central chain (antibody heavy chain) cannot be moved independently, and any motions internal to the antibody cause motions in the antigen. (B) In the Universal FoldTree, the root is an arbitrary point in space known as a virtual residue. The motions propagate to the center of mass (COM) of the antibody and antigen, and further from the antibody COM to the light and heavy chain COMs. From the COMs, the motions propagate both upstream to the beginning of the protein chains and downstream to the end of the protein chains. Each chain can move independently to the others, the antibody can move independently to the antigen, and motions internal to a chain only propagate within the chain.

CHAPTER 5. CONCLUSION

two or more docking partners would also be able to move simultaneously using this framework, allowing ternary or higher order docking moves. An antibody–antigen-specific precursor to the Universal FoldTree is currently being developed by Jeliazko Jeliazkov.³⁸ Once completed and generalized, it should allow for comprehensive multi-body docking in Rosetta.

5.2.2.2 Flexible Docking

The improvements to Rosetta’s docking protocol shown in Chapter 4 are the first step in an ongoing effort to enhance flexible docking. The lower-memory ensembles, faster conformer switching, and more accurate low-resolution scoring provide the basis for several planned expansions of the Rosetta flexible-docking protocol. Larger, more diverse backbone conformer sets will be generated using aggressive methods such as normal mode analysis and kinematic closure to better mimic the conformer selection mechanism. Conformers and complexes will be perturbed and minimized in Cartesian space along principal components of motion observed by the diversification protocols to better mimic the induced fit and allostery mechanisms. A targeted induced fit method in which only a small number of loops or regions are made flexible during docking, enabled by the Universal FoldTree, will be used to address targets such as antibodies or T102 from Chapter 3. Altogether, I believe these planned enhancements have the potential to significantly improve Rosetta’s flexible docking performance.

5.2.2.3 Global Docking

Global docking has long been a weak point of RosettaDock, generally being outpaced by faster FFT methods and outperformed by hybrid methods that incorporate biochemical and evolutionary data.⁵³ While RosettaDock is unlikely to surpass the performance of hybrid methods without also incorporating bioinformatics, it can still be an effective tool for truly blind prediction targets, especially if the speed gap between it and the FFT methods can be narrowed. Both Ellipsoidal Dock and MDS should prove, in a full benchmark, to provide substantial efficiency boosts to global docking in Rosetta. Further efficiency gains can likely be made by incorporating the new Rosetta job distribution protocol, JD3,⁴⁹ into RosettaDock. JD3 allows for Monte Carlo trajectories to be modified or aborted at protocol checkpoints. Combined with the MDS score function and other quality filters, JD3 can be used to remove poor-scoring and clearly non-native decoys at the earliest stage possible, greatly reducing the computational time wasted on dead-end trajectories; additionally, the results of the early trajectories can be used to bias later trajectories, allowing both targeted sampling of low-scoring regions and increased sampling of poorly sampled geometric space, further increasing global docking efficiency within Rosetta.

5.2.3 External Limitations to Protein Structure Prediction

Beyond the technical limitations, there are also external factors that hinder protein structure prediction. RosettaAntibody and RosettaDock, as well as any other protocols (Rosetta or otherwise) that use knowledge-based potentials, are inherently limited by the databases they use for training. As an example, the PDB is extraordinarily valuable to the Rosetta score function, with its hundreds of thousands of structures providing the basis for Rosetta's hydrogen bonding energies, rotameric potential, Ramachandran potential, and many other score terms. Despite its value, the PDB is deeply imperfect. It is severely lacking in some protein classes (e.g. membrane proteins, non-human-/non-murine-derived antibodies), making it difficult to model these proteins; in other places, it is highly redundant, biasing potentials derived from an unculled set. It is poorly curated, with little vetting for errors in structure assignment or file formatting, necessitating a high degree of standardization before use.¹¹ The former shortcoming should abate naturally, albeit slowly, as more structures are deposited, with the sparse regions filling in and the dense regions regressing to the mean; this process could be quickened if owners of proprietary databases (for example, pharmaceutical companies) could be convinced to release at least the non-profit-generating portions of their databases. The latter shortcoming, however, will require a concerted effort to remedy, as depositors currently have lit-

CHAPTER 5. CONCLUSION

the incentive to correct their errors. Because knowledge-based potentials are only as good as their input data, improving the quality and size of databases like the PDB will bestow far-reaching benefits on a large number of protein structure prediction protocols.

In a similar vein, computational protocols can only run as fast as their computational resources allow. As processor technology becomes faster and cheaper, and as public computational resources become more available, the speed of computational structure prediction will increase, and more complex protocols will become computationally viable. Protocols should also begin to fully integrate with parallel processing to make full use of the newest hardware.

Perhaps the single largest barrier to computational structure prediction is the difficulty of use of most scientific software, especially packages that are publicly available. Scientific software is typically single use and not generalizable, not broadly stable or robust across platforms or inputs, not well documented or supported, and not efficient in either time or memory. These deficiencies are especially true for the Rosetta package, which is developed in a massively collaborative manner by a relatively flat hierarchy of academic developers. Most Rosetta developers are amateur programmers with little formal training in optimization or memory management. Since most developers are academics, there is little incentive to clean up or robustify the code once the scientific problem is solved. Without a true project manager to whom all developers are accountable, the writing of documentation and software tests is not

CHAPTER 5. CONCLUSION

strongly enforced. Finally, developer turnover is high, with many leaving the project once they complete their degree or post-doctoral fellowship, often as the only person with full knowledge of their code's function and use and the only one responsible for ensuring their code is not rendered non-functional by one of the hundreds of other developers; as such, many Rosetta protocols quickly become defunct, or worse, silently produce incorrect output. Taken together, these issues substantially delay protocol development and make it difficult to onboard new users and developers, artificially limiting the utility and reach of Rosetta; solving them will go a long way toward advancing the field of protein structure prediction.

Bibliography

- [1] K. R. Abhinandan and A. C. R. Martin. Analysis and prediction of VH/VL packing in antibodies. *Protein Engineering, Design and Selection*, 23(9):689–697, 2010.
- [2] J. Adolf-Bryfogle, Q. Xu, B. North, A. Lehmann, and R. L. Dunbrack. PyIgClassify: a database of antibody CDR structural classifications. *Nucl. Acids Res.*, 43(D1):D432–D438, 2015.
- [3] B. Al-Lazikani, A. M. Lesk, and C. Chothia. Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, 273(4):927–948, 1997.
- [4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular biology of the cell. *ISBN: 0-8153-3218-1*, 4th Edition:"The Generation of Antibody Diversity", 2002.
- [5] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, and K. Kappel et al.

BIBLIOGRAPHY

- The rosetta all-atom energy function for macromolecular modeling and design. *JCTC*, Early Access, 2017.
- [6] J. C. Almagro, A. Teplyakov, J. Luo, R. W. Sweet, S. Kodangattil, F. Hernandez-Guzman, and G. L. Gilliland. Second antibody modeling assessment (AMA-II). *Proteins*, 82(8):1553–1562, 2014.
- [7] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515, 2001.
- [8] S. Baelen, F. Dewitte, B. Clantin, and V. Villeret. Structure of the secretion domain of HxuA from *Haemophilus influenzae*. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, 69(Pt.12):1322–1327, 2013.
- [9] E. N. Baker, M. Paoli, B. F. Anderson, H. M. Baker, W. T. Morgan, and A. Smith. Crystal structure of hemopexin reveals a novel high-affinity heme siteformed between two β -propeller domains. *Nat. Struct. Biol.*, 6(10):926–931, 1999.
- [10] M. L. Bentley, J. E. Corn, K. C. Dong, Q. Phung, T. K. Cheung, A. G. Cochran, P. Adams, P. Afonine, G. Bunkoczi, and V. Chen. Recognition of UbcH5c and the nucleosome by the Bmi1/Ring1b ubiquitin ligase complex. *EMBO J.*, 30(16):3285–3297, 2011.

BIBLIOGRAPHY

- [11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28(1):235–242, 2000.
- [12] G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. S. R. K. Purlavarti, Q. Kaas, and A. Eletsky et al. Accurate de novo design of hyperstable constrained peptides. *Nature*, 538(1):329–335, 2016.
- [13] A. M. Bonvin. Flexible protein-protein docking. *Current Opinion in Structural Biology*, 16(2):194–200, 2006.
- [14] S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. Dimaio, and J. H. Pereira et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science*, 352(6286):680–687, 2016.
- [15] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5472):1868–1871, 2005.
- [16] T. J. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, and D. Baker. Exploring the repeat protein universe through computational protein design. *Nature*, 528(1):580–584, 2015.
- [17] A. Bujotzek, J. Dunbar, F. Lipsmeier, W. Schafer, I. Antes, C. M. Deane, and

BIBLIOGRAPHY

- G. Georges. Prediction of VH–VL domain orientation for antibody variable domain modeling. *Proteins*, 83(4):681–695, 2015.
- [18] A. A. Canutescu and R. L. Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.*, 12(5):963–972, 2003.
- [19] A. Chailyan, P. Marcatili, and A. Tramontano. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS Journal*, 278(16):2858–2866, 2011.
- [20] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman, and J. J. Gray. Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLOS ONE*, 6(8):e22477, 2011.
- [21] S. Chaudhury and J. J. Gray. Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *J. Mol. Biol.*, 381(4):1068–1087, 2008.
- [22] S. Chaudhury, A. Sircar, A. Sivasubramanian, M. Berrondo, and J. J. Gray. Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6–12. *Proteins*, 69(4):793–800, 2007.
- [23] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. MolProbity: All-

BIBLIOGRAPHY

- atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 66(Pt.1):12–21, 2010.
- [24] S. A. Combs, S. L. DeLuca, S. H. DeLuca, G. H. Lemmon, D. P. Nannemann, E. D. Nguyen, J. R. Willis, J. H. Sheehan, and J. Meiler. Small-molecule ligand docking into comparative models with rosetta. *Nature Protocols*, 8(1):1277–1298, 2013.
- [25] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho. ClusPro: a fully automated algorithm for protein-protein docking. *Nucl. Acids Res.*, 32(1):W96–W99, 2004.
- [26] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1):45–50, 2004.
- [27] M. D. Daily, D. Masica, A. Sivasubramanian, S. Somarouthu, and J. J. Gray. CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock. *Proteins*, 60(2):181–186, 2005.
- [28] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, and H. L. Axelrod et al. Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature*, 473(1):540–543, 2010.

BIBLIOGRAPHY

- [29] J. A. Fallas, G. Ueda, W. Sheffler, V. Nguyen, D. E. McNamara, B. Sankaran, J. H. Pereira, F. Parmeggiani, T. J. Brunette, and D. Cascio et al. Computational design of self-assembling cyclic protein homo-oligomers. *Nature Chem.*, 9(4):353–360, 2017.
- [30] J. Fernandez-Recio, M. Totrov, and R. Abagyan. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1):113–117, 2003.
- [31] J. Fernandez-Recio, M. Totrov, and R. Abagyan. Identification of protein–protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, 335(3):843–865, 2004.
- [32] S. A. Fromm, V. Truffault, J. Kamenz, J. E. Braun, N. A. Hoffmann, E. Izauralde, R. Sprangers, M. Albrecht, T. Lengauer, and G. Badis et al. The structural basis of Edc3- and Scd6-mediated activation of the Dcp1:Dcp2 mRNA decapping complex. *EMBO J.*, 31(2):279–290, 2012.
- [33] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, 331(1):281–299, 2003.
- [34] J. J. Gray, S. E. Moughon, T. Kortemme, O. Schueler-Furman, K. M. S. Misura, A. V. Morozov, and D. Baker. Protein–protein docking predictions for the CAPRI experiment. *Proteins*, 52(1):118–122, 2003.

BIBLIOGRAPHY

- [35] L. W. Guddat, J. N. Herron, and A. B. Edmundson. Three-dimensional structure of a human immunoglobulin with a hinge deletion. *PNAS*, 90(9):4271–4275, 1993.
- [36] G. T. Hanson, T. B. McAnaney, E. S. Park, M. E. P. Rendell, D. K. Yarbrough, S. Chu, L. Xi, S. G. Boxer, M. H. Montrose, and S. J. Remington. Green fluorescent protein variants as ratiometric dual emission pH sensors. *Biochemistry*, 41(52):15489–15494, 2002.
- [37] J. Dunbar, A. Fuchs, J. Shi, and C. M. Deane. Abangle: characterising the VH–VL orientation in antibodies. *Protein Eng. Des. Sel.*, 26(10):611–620, 2013.
- [38] J. R. Jeliazkov. *Personal Communication*.
- [39] T. Keitel, A. Kramer, H. Wessner, C. Scholz, J. Schneider-Mergener, and W. Hohne. Crystallographic analysis of anti-p24 (HIV-1) monoclonal antibody cross-reactivity and polyspecificity. *Cell*, 91(6):811–820, 1997.
- [40] K. P. Kilambi, M. S. Pacella, J. Xu, J. W. Labonte, J. R. Porter, P. Muthu, K. Drew, D. Kuroda, O. Schueler-Furman, R. Bonneau, and J. J. Gray. Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. *journalname*, 81(12):2201–2209, 2013.
- [41] G. Kleiger, A. Saha, S. Lewis, B. Kuhlman, and R. J. Deshaies. Rapid E2-E3 assembly and disassembly enable processive ubiquitylation of cullin-RING ubiquitin ligase substrates. *Cell*, 139(5):957–968, 2009.

BIBLIOGRAPHY

- [42] D. Kozakov, D. Beglov, T. Bohnuud, S. E. Mottarella, B. Xia, D. R. Hall, and S. Vajda. How good is automated protein docking? *Proteins*, 81(12):2159–2166, 2013.
- [43] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins*, 65(2):392–406, 2006.
- [44] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov, and S. Vajda. The ClusPro web server for protein–protein docking. *Nature Protocols*, 12(2):255–278, 2017.
- [45] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4):778–795, 2009.
- [46] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *PNAS*, 97(19):10383–10388, 2000.
- [47] D. Kuroda and J. J. Gray. Pushing the backbone in protein-protein docking. *Structure*, 24(10):1821–1829, 2016.
- [48] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35(2):133–152, 1999.
- [49] A. Leaver-Fay. *Personal Communication*.
- [50] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, and W. Sheffler et al. Rosetta3: An

BIBLIOGRAPHY

- object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, 487(1):545–574, 2011.
- [51] M. F. Lensink, S. Velankar, A. Kryshchak, S.-Y. Huang, D. Schneidman-Duhovny, A. Sali, J. Segura, N. Fernandez-Fuentes, S. Viswanath, and R. Elber et al. Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins*, 84(S1):323–48, 2016.
- [52] S. Lyskov, F.-C. Chou, S. O. Conchuir, B. S. Der, K. Drew, D. Kuroda, J. Xu, B. D. Weitzner, P. D. Renfrew, P. Sripakdeevong, and B. Borgo et al. Serverification of molecular modeling applications: The Rosetta Online Server that Includes Everyone (ROSIE). *PLOS ONE*, 8(5):e63906, 2013.
- [53] S. J. Wodak M. F. Lensink, S. Velankar. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins*, 84(S1):323–348, 2016.
- [54] O. Marcu, E.-J. Dodson, N. Alam, M. Sperber, D. Kozakov, M. F. Lensink, and O. Schueler-Furman. FlexPepDock lessons from CAPRI peptide-protein rounds and suggested new criteria for assessment of model quality and utility. *Proteins*, 85(3):445–462, 2017.
- [55] N. A. Marze, J. R. Jeliazkov, S. S. Roy Burman, S. E. Boyken, F. DiMaio, and J. J. Gray. Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35. *Proteins*, 85(3):479–486, 2017.

BIBLIOGRAPHY

- [56] N. A. Marze, S. Lyskov, and J. J. Gray. Improved prediction of antibody VL-VH orientation. *Protein Eng. Des. Sel.*, 29(10):409–418, 2016.
- [57] E. Mashlach, R. Nussinov, and H. J. Wolfson. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucl. Acids Res.*, 38(W):W457–W461, 2010.
- [58] K. M. S. Misura and D. Baker. Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, 59(1):15–29, 2005.
- [59] I. H. Moal and P. A. Bates. SwarmDock and the use of normal modes in protein-protein docking. *Int. J. Mol. Sci.*, 11(10):3623–3648, 2010.
- [60] I. S. Moreira, P. A. Fernandes, and M. J. Ramos. Protein-protein docking dealing with the unknown. *J. Comput. Chem.*, 31(2):317–342, 2010.
- [61] B. North, A. Lehmann, and R. L. Dunbrack. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.*, 406(2):228–256, 2011.
- [62] M. J. O’Meara, A. Leaver-Fay, M. D. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, and B. Kuhlman. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.*, 11(2):609–622, 2015.
- [63] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E.

BIBLIOGRAPHY

- Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, 2017.
- [64] H. Park and F. DiMaio. *Unpublished work, accessible via RosettaCommons git archive.*
- [65] B. G. Pierce, Y. Hourai, and Z. Weng. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLOS ONE*, 6(9):e24657, 2011.
- [66] E. Rivkin, S. M. Almeida, D. F. Ceccarelli, Y.-C. Juang, T. A. MacLean, T. Srikumar, H. Huang, W. H. Dunham, R. Fukumura, and G. Xie et al. The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis. *Nature*, 498(7454):318–324, 2013.
- [67] C. E. M. Schindler, S. J. de Vries, and M. Zacharias. iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement. *Proteins*, 83(2):248–258, 2015.
- [68] I. Sela-Culang, S. Alon, and Y. Ofran. A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J Immunol*, 189(10):4890–4899, 2012.
- [69] Y. Sheng, J. H. Hong, R. Doherty, T. Srikumar, J. Shloush, G. V. Avvakumov, J. R. Walker, S. Xue, D. Neculai, and J. W. Wan et al. A human ubiquitin

BIBLIOGRAPHY

- conjugating enzyme (E2)-HECT E3 ligase structure-function screen. *Mol. Cell. Proteomics*, 11(8):329–341, 2012.
- [70] A. Sircar, S. Chaudhury, K. P. Kilambi, M. Berrondo, and J. J. Gray. A generalized approach to sampling backbone conformations with rosettdock for CAPRI rounds 13-19. *Proteins*, 78(15):3115–3123, 2010.
- [71] A. Sircar and J. J. Gray. SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.*, 6(1):e10000644, 2010.
- [72] A. Sivasubramanian, A. Sircar, S. Chaudhury, and J. J. Gray. Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*, 74(2):497–514, 2009.
- [73] C. A. Smith and T. Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, 380(4):742–756, 2008.
- [74] M. D. Tyka, D. A. Keedy, I. Andre, F. DiMaio, Y. Song, D. C. Richardson, J. S. Richardson, and D. Baker. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.*, 405(2):607–618, 2011.
- [75] A. Urvoas, A. Guellouz, M. Valerio-Lepiniec, M. Graille, D. Durand, D. C. Desravines, H. van Tilbeurgh, M. Desmadril, and P. Minard. Design, produc-

BIBLIOGRAPHY

- tion and molecular structure of a new family of artificial alpha-helical repeat proteins (α Rep) based on thermostable HEAT-like repeats. *J. Mol. Biol.*, 404(2):307–327, 2010.
- [76] S. Vajda, I. A. Vakser, M. J. E. Sternberg, and J. Janin. Modeling of protein interactions in genomes. *Proteins*, 47(4):444–446, 2002.
- [77] A. Vangone, J. P. G. L. M. Rodrigues, L. C. Xue, G. C. P. van Zundert, C. Geng, Z. Kurkcuoglu, M. Nellen, S. Narasimhan, E. Karaca, and M. van Dijk et al. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins*, 85(3):417–423, 2017.
- [78] D. Vasudevan, E. Y. D. Chua, and C. A. Davey. Crystal structures of nucleosome core particles containing the '601' strong positioning sequence. *J. Mol. Biol.*, 403(1):1–10, 2010.
- [79] V. Venkatraman and D. W. Ritchie. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins*, 80(9):2262–2274, 2012.
- [80] T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jimenez-Garcia, P. A. Bates, and J. Fernandez-Recio et al. Updates to the integrated protein-protein interaction benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.*, 427(19):3031–3041, 2015.

BIBLIOGRAPHY

- [81] C. Wang, P. Bradley, and D. Baker. Protein–protein docking with backbone flexibility. *J. Mol. Biol.*, 373(2):503–519, 2007.
- [82] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein–protein docking. *Protein Sci.*, 14(5):1328–1339, 2005.
- [83] F. Wang, D. C. Ekiert, J. Ahmad, W. Yu, Y. Zhang, O. Bazirgan, A. Torkamani, T. Raudsepp, W. Mwangi, and M. F. Criscitiello et al. Reshaping antibody diversity. *Cell*, 153(6):1379–1393, 2013.
- [84] B. Webb and A. Sali. Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, 47(5.6):1–32, 2014.
- [85] B. D. Weitzner, R. L. Dunbrack, and J. J. Gray. The origin of CDR H3 structural diversity. *Structure*, 23(2):302–311, 2015.
- [86] B. D. Weitzner and J. J. Gray. Accurate structure prediction of CDR H3 loops enabled by a novel structure-based C-terminal constraint. *J. Immunol.*, 198(1):505–515, 2017.
- [87] B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R. L. Dunbrack, and J. J. Gray. Modeling and docking of antibody structures with Rosetta. *Nature Protocols*, 12(1):401–416, 2017.

BIBLIOGRAPHY

- [88] B. D. Weitzner, D. Kuroda, N. Marze, J. Xu, and J. J. Gray. Blind prediction performance of RosettaAntibody 3.0: Grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins*, 82(8):1611–1623, 2014.
- [89] J. A. Wojdyla, S. J. Fleishman, D. Baker, and C. Kleanthous. Structure of the ultra-high-affinity colicin E2 DNase–Im2 complex. *J. Mol. Biol.*, 417(1–2):79–94, 2012.
- [90] B. Wu, A. Yee, A. Pineda-Lucena, A. Semesi, T. A. Ramelot, J. R. Cort, J.-W. Jung, A. Edwards, W. Lee, and M. Kennedy et al. Solution structure of ribosomal protein S28E from methanobacterium thermoautotrophicum. *Protein Sci.*, 12(12):2831–2837, 2003.
- [91] H. Xu, A. G. Schmidt, T. O’Donnell, M. D. Therkelsen, T. B. Kepler, M. A. Moody, B. F. Haynes, H.-X. Liao, S. C. Harrison, and D. E. Shaw. Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins*, 83(4):771–780, 2015.
- [92] Z. Zhang and O. F. Lange. Replica exchange improves sampling in low-resolution docking stage of RosettaDock. *PLOS ONE*, 8(8):e72096, 2013.

Vita

VITA

CURRICULUM VITAE

NICHOLAS A. MARZE

BORN DECEMBER 21, 1988
IN ALLENTOWN, PENNSYLVANIA

EDUCATION

PH.D., CHEMICAL AND BIOMOLECULAR ENGINEERING, Aug. 2011–Jun. 2017 (expected)

Johns Hopkins University, Baltimore, MD

Thesis Title: “Building new tools for protein–protein docking and antibody modeling”

Advisor: Jeffrey J. Gray

4.0 GPA (through Fall 2016)

BACHELOR OF CHEMICAL ENGINEERING, Aug. 2007–May 2011

University of Delaware, Newark, DE

Honors Degree with Distinction, double minor in Biochemistry & Biochemical Engineering

Thesis Title: “Quantifying the relative contributions of two pathways to
odd-chain-length fatty acid synthesis in 3T3-L1 adipocytes”

Advisor: Maciek R. Antoniewicz

3.97 GPA, *magna cum laude*

RESEARCH EXPERIENCE

GRADUATE RESEARCH ASSISTANT, Jeffrey J. Gray Lab

Johns Hopkins University, Baltimore, MD

Nov. 2011–present

- Created a new framework for defining and predicting V_L – V_H orientation
- Developed RosettaAntibody homology modeling protocol, including feature addition, porting from python to C++, and standardization/documentation
- Designed and wrote computational method for efficient docking of oblong proteins (Ellipsoidal Dock)
- Added and optimized a fast low-energy screening method in RosettaDock protocol
- Headed lab team in blind prediction of protein complexes from Critical Assessment of PRediction of Interactions (CAPRI) rounds 28–35

UNDERGRADUATE RESEARCH ASSISTANT, Maciek R. Antoniewicz Lab

University of Delaware, Newark, DE

Jun. 2009–May 2011

- Cultured lipogenic mammalian cells, standardized and optimized culture protocol
- Designed and ran *in vitro* ^{13}C tracer experiments, collected concentration data using gas chromatography–mass spectroscopy, built models of lipid production pathways from data using computational metabolic flux analysis

SUMMER RESEARCH ASSISTANT, M. Kathryn Iovine Lab

Lehigh University, Bethlehem, PA

Jun. 2008–Aug. 2008

- Sorted zebrafish embryos for use in regenerative medicine research

VITA

PUBLICATIONS

1. **N.A. Marze**, J.R. Jeliaskov, S.S. Roy Burman, S.E. Boyken, F. DiMaio, J.J. Gray “Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35”. *Proteins*, 85(3), 479-486 (2017).
2. B.D. Weitzner, J.R. Jeliaskov, S. Lyskov, **N.A. Marze**, D. Kuroda, R. Frick, N. Biswas, J.J. Gray. “Modeling and docking antibody structures with Rosetta”. *Nat. Protoc.*, 12(2), 401-416 (2017).
3. **N.A. Marze**, S. Lyskov, & J.J. Gray. “Improved Prediction of Antibody V_L-V_H Orientation.” *PEDS* 29(10), 409-418 (2016).
4. M.F. Lensink et al. (102 authors, including **N. Marze**) “Prediction of homo- and hetero-protein complexes by *ab-initio* and template-based docking: a CASP-CAPRI experiment”. *Proteins* 84(S1), 323-348 (2016).
5. S.B. Crown, **N. Marze**, M.R. Antoniewicz. “Catabolism of Branched Chain Amino Acids Contributes Significantly to Synthesis of Odd-Chain and Even-Chain Fatty Acids in 3T3-L1 Adipocytes.” *PLoS ONE* 10(12): e0145850 (2015)
6. B.D. Weitzner, D. Kuroda, **N. Marze**, J. Xu, & J.J. Gray. “Blind Prediction Performance of RosettaAntibody 3.0: Grafting, Relaxation, Kinematic Loop Modeling, and CDR Optimization.” *Proteins* 82(8), 1611-1623 (2014)

CONFERENCE PRESENTATIONS

*Poster presentation, speaker underlined

1. **N.A. Marze**, B.D. Weitzner, J.R. Jeliaskov, S. Lyskov, D. Kuroda, R. Frick, N. Biswas, J.J. Gray. “The New RosettaAntibody”. RosettaCON 2016, Leavenworth, WA. August 9, 2016.
2. **N.A. Marze**, J.R. Jeliaskov, S.S. Roy Burman, D. Kuroda, S. Boyken, J. Koehler Leman, J.J. Gray. “Accurate global docking in RosettaDock is masked by difficult multi-domain targets in CAPRI rounds 31–35”. 6th CAPRI Evaluation Meeting, Tel Aviv, Israel. April 17, 2016.
3. **N.A. Marze** & J.J. Gray. “Using Multiple Templates to Improve Antibody Modeling”. WinterRosettaCON March 2016, Burlington, VT. March 10, 2016.
4. ***N. Marze** & J.J. Gray. “Two New Rosetta Utilities”. RosettaCON 2014, Leavenworth, WA, July 31, 2014.
5. ***N. Marze** & J.J. Gray. “Docking Antibody Repertoires in Rosetta”. RosettaCON 2012, Leavenworth, WA, July 29, 2012.
6. ***N.A. Marze**, S.B. Crown, M.R. Antoniewicz. “Elucidating the Effect of Passage Number, Culture Flask Type, and Serum Presence on 3T3-L1 Adipocyte Metabolism”. 2010 AIChE Annual Meeting, Salt Lake City, UT, November 8, 2010.

LEADERSHIP EXPERIENCE

RESEARCH MENTOR

Johns Hopkins University, Baltimore, MD

- Naireeta Biswas, Master's student, Summer 2016–present
- Gabriela Sinclair, undergraduate student, Fall 2015–Summer 2016
- Nikhil Shah, undergraduate student, Fall 2014–Summer 2016
- Michael Porter, undergraduate student, Fall 2013–Summer 2014

VITA

RECRUITMENT CHAIR, Graduate Student Liaison Committee

Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD

Lead Recruitment Chair, 2013–Mar. 2016

Assistant Recruitment Chair, Jun. 2012–May 2013

- Organized and oversaw annual 3-day new graduate student recruitment event
- Coordinated student volunteers; arranged local travel & activities; interacted with and accommodated recruits; kept the event running smoothly

TEACHING ASSISTANT, EN.540.203: Engineering Thermodynamics

Johns Hopkins University, Baltimore, MD

Spring 2012 & Spring 2013

- Guest lectured, taught weekly recitation sections, held weekly review sessions
- Helped write and graded exams, homework

TEACHING ASSISTANT, CHEM 444: Physical Chemistry II

University of Delaware, Newark, DE

Fall 2009

- Held weekly review sessions
- Graded exams, homework

SKILLS

- Fluency in C++, Python, R, Bash, Excel
- Competency in UNIX environment
- Experienced Rosetta user and developer
- Processing and analysis of large data sets
- Protein homology modeling, focus on antibodies
- Protein–protein docking and complex structure prediction
- Basic machine learning (decision tree, SVM)

HONORS AND AWARDS

- Schwartz Fellowship at Johns Hopkins University, 2011–2016
- Hertz Fellowship Finalist, 2011
- NSF GRFP Honorable Mention, 2011
- Dean's List at University of Delaware, 2007–2011
- 1743 Scholar (academic scholarship) at University of Delaware, 2007–2011