2019

Computational modeling of protein-protein and protein-peptide interactions

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

COMPUTATIONAL MODELING OF PROTEIN-PROTEIN AND PROTEIN-PEPTIDE INTERACTIONS

by

KATHRYN A. PORTER

B.S., University of Maine, 2013
M.Eng., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

Sandor Vajda, Ph.D.
Professor of Biomedical Engineering
Professor of Systems Engineering
Professor of Chemistry

Second Reader

Maxim D. Frank-Kamenetskii, Ph.D., Sc.D.
Professor of Biomedical Engineering
Professor of Materials Science and Engineering

Third Reader

Dmytro Kozakov, Ph.D.
Research Associate Professor of Biomedical Engineering
Boston University, College of Engineering

Assistant Professor of Applied Mathematics & Statistics
Stony Brook University

Fourth Reader

Adrian Whitty, Ph.D.
Associate Professor of Chemistry

Fifth Reader

Dmitri Beglov, Ph.D.
Research Assistant Professor of Biomedical Engineering

“Life, which you so nobly serve, comes from destruction, disorder and chaos.”

- Jean-Baptiste Emanuel Zorg, *The Fifth Element*

ACKNOWLEDGMENTS

I would like to thank everyone at the Structural Bioinformatics Laboratory for all of their support. Having coffee and tea time (and help with the daily crossword puzzle) was a nice break to look forward to every morning, with special thanks to Christine for her many delicious cakes and pastries! And a big thank you to my long-time office buddy, Zhuyezi (Julie) for putting up with all of my antics and helping me troubleshoot various research and server problems.

I'd like to also express my appreciation to former lab members Bing and David for teaching me so much about maintaining our lab servers. My sincerest gratitude to Sandor, Dima, and Dmitri for all of their advice and encouragement. Thanks also to collaborators in the Chemistry Department and the Laboratory of Computational Immunology at Boston University, the Applied BioComputation Group at Stony Brook University, and the Furman Lab at the Hebrew University of Jerusalem. Finally, special thanks to my wonderful family and friends for all of their encouragement and support.

COMPUTATIONAL MODELING OF PROTEIN-PROTEIN

AND PROTEIN-PEPTIDE INTERACTIONS

KATHRYN A. PORTER

Boston University College of Engineering, 2019

Major Professor: Sandor Vajda, Ph.D., Professor of Biomedical Engineering, Professor of Systems Engineering, Professor of Chemistry

ABSTRACT

Protein-protein and protein-peptide interactions play a central role in various aspects of the structural and functional organization of the cell. While the most complete structural characterization is provided by X-ray crystallography, many biological interactions occur in complexes that will not be amenable to direct experimental analysis. Therefore, it is important to develop computational docking methods that start from the structures of component proteins and predict the structure of their complexes, preferably with accuracy close to that provided by X-ray crystallography. This thesis details three applications of computational protein modeling, including the study of antibody maturation mechanisms, and the development of protocols for peptide-protein interaction prediction and template-based modeling of protein complexes.

The first project, a comparative analysis of docking an antigen structure to antibodies across a lineage, reveals insights into antibody maturation mechanisms. A linear relationship between near-native docking results and changes in binding free energy is established, and used to investigate changes in binding affinity following mutation across two antibody-antigen systems: influenza and anthrax. The second project demonstrates that a motif-based search of available protein crystal structures is sufficient

to adequately represent the conformational space sampled by a flexible peptide, compared to that of a rigid globular protein. This observation forms the basis for a global peptide-protein docking protocol that has since been implemented into the Structural Bioinformatics Laboratory's docking web server, ClusPro. Finally, as structure availability remains a roadblock to many studies, researchers turn to homology modeling, in which the desired protein sequence is modeled onto a related structure. This is particularly challenging when the target is a protein complex, further restricting template availability. To address this problem, the third project details the development of a new template-based modeling protocol to be integrated into the ClusPro server. The implementation of a novel template-based search enables users to model both homomeric and heteromeric complexes, greatly expanding ClusPro server functionality.

TABLE OF CONTENTS

ACKNOWLEDGMENTS v

ABSTRACT vi

TABLE OF CONTENTS viii

LIST OF TABLES xi

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvii

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Protein-Protein Docking 2

1.2.1 FFT-Based Rigid Body Sampling 2

1.2.2 Energy Function 4

1.2.3 DARS Potential 5

1.2.4 Clustering 5

1.2.5 Minimization 7

1.3 ClusPro in CAPRI 7

1.4 Contributions 8

2 PEPTIDE-PROTEIN DOCKING 9

2.1 Introduction 9

2.2 PeptiDock Method 11

2.2.1 Input Structures and Motif Selection....13
2.2.2 Clustering of Fragments....14
2.2.3 Docking of Peptide Fragments....15
2.2.4 Selection of Models....15
2.3 Motif-Derived Fragment Sets....16
2.4 PeptiDock Performance....19
2.5 Using ClusPro PeptiDock....22
3 MODELING ANTIBODY MATURATION....24
3.1 Introduction....24
3.2 Methods....28
3.2.1 Mutation of Selected Residues....30
3.2.2 Free Antibody Ensemble Generation....30
3.2.3 Rigid Body "Focused" Docking....32
3.2.4 Connecting Near Native Hits to Free Energy....32
3.2.5 Computational Alanine Scanning....34
3.3 Results and Discussion....35
3.3.1 CASE 1: The influenza HA antibody system....35
3.3.2 CASE 2: The anthrax PA antibody system....41
3.3.3 Interface Assessment....45
3.4 Conclusions....47
4 TEMPLATE-BASED MODELING....50
4.1 Introduction....50

4.2 Methods 55
4.2.1 Template-based Modeling 56
4.2.2 Free Docking 58
4.2.3 Co-minimization via CHARMM 59
4.3 Results and Discussion 59
4.3.1 CASE 1: T152/T1003, a simple homodimer 60
4.3.2 CASE 2: T142/H0974, heterodimer based on homodimer 61
4.3.3 CASE 3: T141/T0976, homodimer based on monomer 62
4.3.4 Future Directions 63
4.4 Conclusions 68
Appendix A: Supplemental Tables for PEPTIDE-PROTEIN DOCKING 70
Appendix B: Supplemental Figures/Tables for ANTIBODY MATURATION 76
BIBLIOGRAPHY 88
CURRICULUM VITAE 96

LIST OF TABLES

Table 2.1: Set of peptide-protein complexes from the PeptiDB v2 set 17

Table 2.2: Overall assessment of the motif-domain docking performance 20

Table 2.3: Use of electrostatic-driven potential improves performance for specific cases 20

Table 3.1: Binding kinetics for UCA, I-2, CH65, and CH67 Fabs, determined by SPR in previously published study.....35

Table 3.2: Binding data for anthrax PA antibodies, determined by SPR in a previous study.....45

Table 4.1: Number of models by template-based docking (A), global free docking (B), and focused free docking (C). 54

Table A.1: Definition of sequence motifs for the extraction of fragments from the PDB for the PeptiDB v2 set.....70

Table A.2: Definition of sequence motifs for the extraction of fragments from the PDB for the "Recent PDB" set 71

Table A.3: Fragments extracted from the PDB using the KXRRL motif for binding of CDC6 derived peptide to cyclin 71

Table A.4: Docking results for representative fragments, based on KXRRL motif, of CDC6 derived peptide binding to cyclin.....73

Table B.1: Near native counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking with HA antigen assuming antibody bound conformation.77

Table B.2: Weighted near native average counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking unbound antibody ensembles.....78

Table B.3: Lowest PIPER energies associated with near native counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking with HA antigen assuming antibody bound conformation. 78

Table B.4: Lowest PIPER energies associated with weighted near native average counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking unbound antibody ensembles 78

Table B.5: Predicted - ΔG values calculated from linear fits between experimental - ΔG values estimated from SPR data to N and PIPER energies from docking CH65-CH67 lineage antibodies with the HA antigen, assuming antibody bound conformation.....79

Table B.6: Predicted - ΔG values calculated from linear fits between experimental - ΔG values estimated from SPR data to weighted N and PIPER energies from docking CH65-CH67 lineage unbound antibody ensembles.....79

Table B.7: Residual errors calculated between experimental - ΔG values and those predicted from linear fits of docking results from CH65-CH67 lineage antibodies (restricted backbone).....80

Table B.8: Residual errors calculated between experimental - ΔG values and those predicted from linear fits of docking results from CH65-CH67 lineage unbound antibody ensembles.....80

Table B.9: Experimental K_D values for the CH65-67 lineage antibodies, estimated from SPR measurements 81

Table B.10: Near native counts (within 10 Å to native interface) for anthrax PA antibodies after docking with PA antigen assuming antibody bound conformation. 81

Table B.11: Weighted near native average counts (within 10 Å to native interface) for anthrax PA antibodies after docking unbound antibody ensembles. 82

Table B.12: Lowest PIPER energies associated with near native counts (within 10 Å to native interface) for anthrax PA antibodies after docking with PA antigen assuming antibody bound conformation. 82

Table B.13: Lowest PIPER energies associated with weighted near native average counts (within 10 Å to native interface) for anthrax PA antibodies after docking unbound antibody ensembles. 82

Table B.14: CHARMM Energy calculations for non-bonded interactions of antibody interface (within 10 Å of HA) residues of FabCH67-HA complex. 83

Table B.15: CHARMM Energy calculations for non-bonded interactions in the interface (within 10 Å of HA) of the FabCH65-HA complex. 84

Table B.16: CHARMM Energy calculations for non-bonded interactions in the interface (within 10 Å of PA) of the HAAb-PA complex. 86

LIST OF FIGURES

Figure 2.1: Overview of the PeptiDock algorithm 12

Figure 2.2: Distribution of distances to the native peptide for motif libraries for a subset of Peptide DB v2 cases 18

Figure 2.3: Modeled protein-peptide complexes from selected PeptiDB v2 set 21

Figure 2.4: Examples for models generated by PeptiDock rigid body docking of peptides to a receptor 22

Figure 3.1: Protocol for docking-based antibody maturation assessment 28

Figure 3.2: Representation of focused docking of the antigen to an antibody structure ..31

Figure 3.3: Overlap of the CH65-CH67 Fab lineage shows the unbound CH67 (PDB 4hkb) overlapping with both bound structures for CH65, CH67 (PDB IDs 4hkx, 3sm5) 36

Figure 3.4: Histogram of H3 alpha carbon RMSD values from the 10X100 ns MD simulations for CH65-CH67 lineage members 37

Figure 3.5: -ΔG predicted by near-native hits and PIPER energies, plotted against experimentally measured -ΔG for CH65-67 Fabs 38

Figure 3.6: If the conformation is fixed, the near-native hits are invariant compared to ensemble docking of MD snapshots 39

Figure 3.7: Fraction of conformers assuming a bound conformation across UCA, I-2, CH65, CH67 populations 41

Figure 3.8: The Kepler group at BUSM provided crystal structures and binding data of three antibodies against the anthrax toxin PA, at low (UCA), medium (1558) and high affinities (1184).....42

Figure 3.9: Histogram of RMSD values between each snapshot and the bound antibody (1184) are shown for each of the MD-generated ensembles starting from the crystallized unbound antibodies 43

Figure 3.10: Comparison between unbound ensemble and assumed-bound conformation docking of anthrax PA Fabs 44

Figure 4.1: General comparison of template-based and free docking methods for an example heterodimer target.....51

Figure 4.2: General outline of the ClusPro template-based modeling (TBM) protocol...55

Figure 4.3: Model of T152/T1003 (green and cyan) overlapped with its homodimeric template (wheat, PDB 2W8T). 61

Figure 4.4: A model of T142/H0974 (green and cyan) overlapped with its homodimeric template (wheat, PDB 4RYK). 62

Figure 4.5: Modeled subunits (green and cyan) of T141/T0976 aligned to different locations on the same chain of the template protein (wheat, PDB 1YT8).....63

Figure 4.6: A model (green, cyan) of T137/T0965 superimposed with the native structure (wheat, PDB 6D2V).....64

Figure 4.7: Two different T75/T0776 models (green-cyan and yellow-pink) aligned to one of the subunits of the target structure (gray, PDB 4Q9A) 66

Figure 4.8: Template based modeling of target T159/H1021 assisted by low-resolution Electron Microscopy data.....67

Figure B.1: Sequences for inferred CH65-CH67 lineage.....76

Figure B.2: Sequences for anthrax PA antibodies.....77

LIST OF ABBREVIATIONS

ABNR ..... Adapted Basic Newton-Raphson

BCR ..... B-cell Receptor

BU.....Boston University

BUSM ..... Boston University School of Medicine

CAPRI..... Critical Assessment of PRediction of Interactions

CASP ..... Critical Assessment of protein Structure Prediction

CDC6 ..... Cell Division Cycle 6

CDR.....Complementarity-determining Region

CHARMM ..... Chemistry at Harvard Macromolecular Mechanics

DARS..... Decoys as the Reference State

Fab ..... Antigen Binding Fragment

FDA ..... Food and Drug Administration

FFT .....Fast Fourier Transform

FT ..... Fourier Transform

GUI .....Graphical User Interface

HA .....Hemagglutinin

IFT .....Inverse Fourier Transform

Ig.....Immunoglobulin

IRMSD..... Interface Root Mean Squared Deviation

mAb .....Monoclonal Antibody

MD..... Molecular Dynamics

N ..... Number of Near-Native Docking Results

PA .....Protective Antigen

PDB ..... Protein Data Bank

PPI .....Protein-Protein Interaction

RMSD .....Root Mean Squared Deviation

SAXS .....Small-angle X-ray Scattering

SLIM.....Short Linear Interacting Motif

SPR ..... Surface Plasmon Resonance

TBM.....Template Based Modeling

UCA..... Unmutated Common Ancestor

1 INTRODUCTION

1.1 Motivation

Protein-protein interactions (PPIs) are involved in a variety of signaling pathways critical for regulating cellular function. The study and characterization of these interactions is an important step in the understanding of many disease mechanisms. While the number of available structures determined through experimental methods such as X-ray crystallography is constantly growing, there are still complexes in which structural elucidation is not possible. Computational methods can prove insightful alternatives for these cases, and can also be used for more efficient screening of potential drug-candidates, via structure-based drug design (Halperin, Ma, Wolfson, & Nussinov, 2002; Ritchie, 2008; G. R. Smith & Sternberg, 2002). For this reason, the development of docking methodologies capable of predicting protein complexes when structures are not available has immediate application to the drug discovery field.

In the last few decades, the use of antibodies and peptides as therapeutic agents has seen a dramatic rise in the pharmaceutical industry. The first antibody therapeutic was approved by the Food and Drug Administration (FDA) in 1986. Since then, improved research and production techniques have led to an increase in the number of antibody drugs approved each year (Awwad & Angkawinitwong, 2018). More recently, peptide therapeutics have seen a surge in popularity. Over 60 drugs have been FDA approved, with hundreds in clinical and pre-clinical trial phases (Erak, Bellmann-Sickert, Els-Heindl, & Beck-Sickinger, 2018). The growing therapeutic market for both peptide and antibody drugs establishes a strong need for structure-based computational methods

that supplement continued studies of these structures. Docking and modeling protocols capable of predicting protein-protein complexes, such as antibody-antigen interactions, and protein-peptide binding may prove advantageous tools for the future.

1.2 Protein-Protein Docking

The first computational protein-protein docking approaches emerged in the late 1970s (Greer & Bush, 1978; Vakser, 2014; Wodak & Janin, 1978) and largely focused on global sampling assuming rigid body motion for component proteins. Along with an increased number of available structures and improvements in computing resources, the introduction of efficient sampling techniques, particularly the fast Fourier transform (FFT) correlation approach proposed by Katchalski-Katizir et al in 1992 (R. Chen, Li, & Weng, 2003; Katchalski-Katzir et al., 1992; Sternberg, 2000; Vakser, 1996) marked a key advancement in the field. Several groups have since incorporated this method into their docking servers, including the BU Structural Bioinformatics lab's protein-protein docking server ClusPro (Kozakov et al., 2013). The ClusPro protocol uses an FFT-Based approach to sample 10^9-10^{10} conformations of the putative complex, followed by clustering of lowest energy conformers, which are then ranked by population and minimized. ClusPro is heavily used; by June 2019 it had 12,000 registered users (registration is not a requirement), and completed over 280,000 docking calculations.

1.2.1 FFT-Based Rigid Body Sampling

The ClusPro server uses the PIPER program (Kozakov, Brenke, Comeau, & Vajda, 2006) to fully explore the conformational space of rigid body orientations between

a given ligand and receptor; this is accomplished through an exhaustive evaluation of an energy function in the discretized space of mutual orientations of the protein and ligand using a FFT correlation approach. The energy-like scoring function *(1.1)* is expressed as a sum of P correlation functions across all possible translations of the ligand relative to the receptor structure: α, β, γ. R_p defines components of the correlation on the receptor, and L_p defines components on the ligand (Kozakov et al., 2006).

$$E(\alpha,\ \beta,\ \gamma) = \sum_{p=1}^{P} \sum_{i,j,k} R_p(i,j,k) L_p(i+\alpha,\ j+\beta,\ k+\gamma) \qquad (1.1)$$

The center of mass of the receptor protein is fixed at the origin of the coordinate system, whereas the ligand is rotated and translated. The translational space is represented as a grid of 1.0 Å displacements of the ligand center of mass, and the rotational space is sampled using 70,000 rotations based on a deterministic layered Sukharev grid sequence, which quasi-uniformly covers the space. The expression is evaluated by using P forward (FT) and one inverse Fast Fourier (IFT) transforms, effectively improving efficiency of the approach from $O(N^6)$ to $O(N^3(\log N^3))$. In the following expressions, $\mathbf{i} = \sqrt{-1}$, with N_1, N_2, and N_3 representing the dimensions the grid along three coordinates. C is equal to $1/(N_1N_2N_3)$.

$$E(\alpha,\ \beta,\ \gamma) = IFT\left\{\sum_{p=1}^{P} FT^*\{R_p\} FT\{L_p\}\right\}(\alpha,\beta,\gamma) \qquad (1.2)$$

$$FT\{F\}(l,m,n) = \sum_{i,j,k} F(i,j,k) e^{-2\pi \mathbf{i}(li/N_1 + mj/N_2 + nk/N_3)} \qquad (1.3)$$

$$IFT\{f\}(i,j,k) = C \sum_{l,m,n} f(l,m,n) e^{2\pi i(li/N_1 + mj/N_2 + nk/N_3)} \qquad (1.4)$$

1.2.2 Energy Function

The energy expression *(1.5)* includes a simplified van der Waals energy E_{vdw} with attractive (E_{attr}) and repulsive (E_{rep}) contributions *(1.6)*, electrostatic interaction energy E_{elec}, and a statistical pairwise potential E_{pair}, representing other solvation effects (Chuang, Kozakov, Brenke, Comeau, & Vajda, 2008).

$$E = E_{vdw} + w_2 E_{elec} + w_3 E_{pair} \qquad (1.5)$$

$$E_{vdw} = E_{attr} + w_1 E_{rep} \qquad (1.6)$$

$$E_{elec} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_1 q_2}{\left(r_{ij}^2 + D^2 exp\left(\frac{-r_{ij}^2}{4D^2}\right)\right)^{\frac{1}{2}}} \qquad (1.7)$$

$$E_{pair} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij} \qquad (1.8)$$

The electrostatic interaction energy is defined in *(1.7)*, where N_R and N_L represent the number of atoms in the receptor and ligand, respectively. D is an atom-type independent approximation of the generalized Born radius, and r the distance between atoms i and j. The weights for each energy term, w_1, w_2, w_3, may be varied to favor different energetic contributions, depending on the type of protein complex being docked.

All energy expressions are defined on the grid. In order to evaluate the energy function E by FFT, it must be written as a sum of correlation functions. The first two terms, E_{vdw} and E_{elec}, satisfy this condition, whereas E_{pair} is written as a sum of a few correlation functions *(1.8)*, using an eigenvalue-eigenvector decomposition (Kozakov et al., 2006).

1.2.3 DARS Potential

The pairwise structure-based potential used in the PIPER energy function is referred to as DARS, or Decoys As the Reference State (Chuang et al., 2008). DARS represents desolvation contributions to the interaction energy in *(1.5).* The statistical potential of two atom types *I* and *J* as ε_{IJ} is expressed as follows:

$$\varepsilon_{IJ} = -kTln(p_{IJ}) \qquad (1.9)$$

k is the Boltzmann constant, *T* is the temperature and p_{IJ} represents the probability that the two atom types will interact. This probability is approximated by *(1.10)* where v_{IJ}^{obs} and v_{IJ}^{ref} represent the number of interacting atom pairs observed and the expected number of interacting pairs in a reference state, respectively.

$$p_{IJ} = \frac{v_{IJ}^{obs}}{v_{IJ}^{ref}} \qquad (1.10)$$

DARS uses docking decoys as the reference set, which allows for discrimination of near-native conformations from other docking results. DARS performs particularly well for enzyme-inhibitor complexes, where it was shown that performance did not heavily depend on complex selection used for decoy generation (Chuang et al., 2008). Since its initial development, an asymmetric implementation has been created specifically for antibody-antigen complexes (Brenke et al., 2012).

1.2.4 Clustering

In ClusPro, the lowest energy structures are clustered using a greedy clustering algorithm. For *k* structures, we calculate the $k \times k$ matrix of pairwise backbone Root

Mean Square Deviations (RMSD). We count the number of neighbors each structure has within a defined cluster radius. The members of the largest cluster are removed from the pool of structures, and the procedure is repeated until no structures remain, resulting in clusters ranked according to their size (Kozakov, Clodfelter, Vajda, & Camacho, 2005). The cluster centers are then minimized (Section 1.2.5) and presented as final results.

Clusters represent isolated, highly populated low energy basins of the energy landscape, and the large clusters are thus more likely to include native structures. The globally sampled conformational space can be considered as a canonical ensemble with the partition function *(1.11)*, where we sum the associated energy E_j over all poses *j*. For the k^{th} cluster, the partition function is given by *(1.12)*, where the sum is restricted to poses within the cluster. Based on these values, the probability of the k^{th} cluster is given by *(1.13)*.

$$Z = \sum_j e^{-E_j/RT} \qquad (1.11)$$

$$Z_k = \sum_{j,\ j\in k} e^{-E_j/RT} \qquad (1.12)$$

$$P_k = Z_k/Z \qquad (1.13)$$

Since the low energy structures are selected from a relatively narrow energy range, and the energy values are calculated with considerable error, it is reasonable to assume that these energies do not differ, *i.e.,* $E_j=E$ for all j in the low energy clusters.

$$Z = Ne^{-E/RT} \qquad (1.14)$$

$$P_k = Z_k/Z = N_k/\mathrm{N} \qquad (1.15)$$

This simplification implies that $P_k = \frac{N_k}{Z} e^{-E/RT}$, and thus the probability P_k is proportional to N_k, where N_k is the number of structures in the k^{th} cluster (Kozakov et al., 2013).

1.2.5 Minimization

For each of final structures retained after clustering, minimization is performed using the polar hydrogen PARAM19 like forcefield with CHARMM (Brooks et al., 2009). The protocol consists of 300 steps (with fixed backbone) using only the van der Waals term of the CHARMM potential, removing steric overlap with only small conformational changes in the protein complex (Kozakov et al., 2017).

1.3 ClusPro in CAPRI

Community-wide assessments CAPRI (Critical Assessment of Predicted Interactions) and CASP (Critical Assessment of protein Structure Prediction) are both important platforms used to evaluate protein-protein docking and protein structure prediction methodologies, respectively. CAPRI, which was modeled on CASP, was launched in 2001, with the aim of providing structures to participating groups who would then submit blind docking predictions for evaluation (Janin et al., 2003). Since then, the competition has evolved, with the addition of certain data-assisted targets, and scoring categories.

This first joint CASP-CAPRI round was held in 2014 (CASP11-CAPRI), and has continued on a bi-annual basis (Lensink et al., 2018). Sequences are typically provided to CASP participants, who will then provide structural predictions which will serve as

inputs for complex prediction by CAPRI groups. However, more recently, CAPRI groups have adopted template-based modeling techniques to produce their own component structures.

The initial version of ClusPro participated in CAPRI assessments beginning in 2003 (Comeau et al., 2007), after which PIPER with DARS was added into the protocol. This latest iteration of ClusPro has been a CAPRI participant since Round 13 (Round 47 is the most recent, in 2019), including all joint CASP-CAPRI rounds, and has repeatedly ranked among the top servers (Kozakov et al., 2013; Kozakov et al., 2010; Vajda et al., 2017). The results of the latest CASP-CAPRI assessment will be discussed in Chapter 4.

1.4 Contributions

Bing Xia performed peptide weight optimization docking runs and helped to add the peptide docking feature (Chapter 2) into the ClusPro web server. The work in Chapter 3 was a collaboration between myself and Zhuyezi Sun. Istvan Kolossvary performed all MD simulations for the study. The Chapter 4 docking runs for the CASP11-12 homodimer evaluation were performed by Israel Desta. Dzmitry Padhorny helped to automate extensive template searches used in ClusPro TBM.

2 PEPTIDE-PROTEIN DOCKING

2.1 Introduction

While ClusPro is designed to dock two proteins that are relatively rigid, i.e., their backbones do not substantially change conformations upon binding, many ClusPro users attempt to dock flexible linear peptides; the rigid body docking algorithm currently in ClusPro is not suitable for docking such flexible peptides. In spite of this shortcoming, many users pre-generate peptide conformations and attempt the docking. The demand for peptide-protein docking is high because a significant fraction of proteins actually bind to flexible peptide-like regions, called short linear interacting motifs (SLIMs) of the partner protein (Stein & Aloy, 2008). In addition, peptide docking is frequently required in the process of drug discovery. Peptide-mediated interactions are involved in a number of critical cellular processes, prompting the emergence of peptide-like compounds for favorable drug targeting. Of the publicly accessible peptide-protein docking servers, few globally sample the receptor structure, i.e. they require information on the approximate location of the binding site. Those that do use computationally less efficient Monte Carlo and Molecular Dynamics (MD) approaches (Ben-Shimon & Niv, 2015; Blaszczyk et al., 2016; Dagliyan, Proctor, D'Auria, Ding, & Dokholyan, 2011; Kurcinski, Jamroz, Blaszczyk, Kolinski, & Kmiecik, 2015; Lee, Heo, Lee, & Seok, 2015; London, Raveh, Cohen, Fathi, & Schueler-Furman, 2011; Schindler, de Vries, & Zacharias, 2015). It is clear that peptide docking is an unmet need, and the development of a fairly accurate algorithm and its addition to ClusPro provides an important tool to the biomedical research community.

As described in Section 1.2, the FFT-based approach implemented in the current version of the ClusPro server was developed for docking relatively rigid proteins, and was not applicable to flexible systems, such as peptide-protein interactions. Existing peptide docking methods are local, i.e. require information on the approximate location of the binding site, or use computationally less efficient Monte Carlo and MD. In spite of performing local rather than global docking, such methods are too resource demanding for server implementation. However, the peptides in SLIMS frequently have conserved sequence motifs which makes the peptide-protein problem amenable to an FFT-based approach. For example, cyclins recognize the RXL motif, where R (arginine) and L (leucine) are fixed as amino acids, and X represents any amino acid. It has been shown that motif regions are limited to a small ensemble of potential conformations within a protein environment (Wang et al., 2013). The key idea of our approach is that we generate this ensemble by searching for proteins containing the motif of interest in the Protein Data Bank (PDB) and extract the regions that match the motif. Conceptually similar approaches of mining fragments from the PDB have previously been used for protein structure prediction in the Rosetta program (Simons, Kooperberg, Huang, & Baker, 1997). While Rosetta was based on sequence and secondary structure similarity, our method uses motif-based fragment extraction to focus on a relevant subset of conformations, taking advantage of the specific features of the motif docking challenge. By combining this resulting fragment library with systematic FFT grid-based sampling using accurate molecular mechanics potentials (Chuang et al., 2008; Kozakov et al., 2006), we efficiently sample and discriminate near native docked peptide models with success rates similar to

protein docking, despite significant peptide flexibility. The method was validated by application to the PeptiDB v2 dataset for which the interacting motif had been reported (Lavi et al., 2013). To prevent inherent bias, structures of proteins with more than 30% sequence identity to the target were excluded from the search for motif backbones, and only information about sequence motifs available before the publication of the solved complex structure were included.

2.2 PeptiDock Method

Our motif-based approach is composed of four main steps, and hinges on the observation that extracting representative fragments based on a known motif will adequately sample the peptide conformational space. Figure 2.1 demonstrates the steps of our protocol on an example application, the interaction between cyclin (structure of free cyclin, PDB ID 1H1R) and a peptide derived from CDC6 (HTLKGRRLVFDN) (Cheng et al., 2006) (structure of the complex, PDB ID 2CCH). The RxL peptide motif (Arginine, followed by any amino acid, and Leucine) was defined based on a literature search (Cheng et al., 2006). We use a set of rules (further described in Section 2.2.1) to extend and refine this initial motif, until a search in the PDB results in a comprehensive set of fragment hits (note that for protocol validation homologs of complex structures are, of course, excluded). In this example (Figure 2.1A), the peptide sequence covering the initial motif (i.e. RRL) was first extended in the N-terminal direction towards a polar residue, to yield a pentamer sequence KGRRL (applying rule **Sp**). Since this motif is found only nine times in the PDB, we made it more general by introducing a wild-card at the smallest residue, G, to obtain KXRRL (rule **E**). This motif is found frequently enough

to proceed to the next step (in 472 PDB structures, homologs of the solved CDC6-cyclin structure (PDB ID 2CCH) excluded). Once the docking motif is defined, we extract the matching fragments from the PDB (1051 fragments) and cluster them with a 0.5 Å RMSD cluster threshold (resulting in 40 clusters for this example, see Figure 2.1B and Table A.3). Representatives from the top 25 largest clusters are then each docked to the receptor structure (Figure 2.1C). All solutions are pooled and clustered with a 3.5 Å RMSD threshold, and representatives of each cluster are further minimized to produce the final 100 models (Figure 2.1D and Table A.4). In the cyclin-CDC6 peptide example, the third ranked model lies within 1.9 Å RMSD (Figure 2.1E).

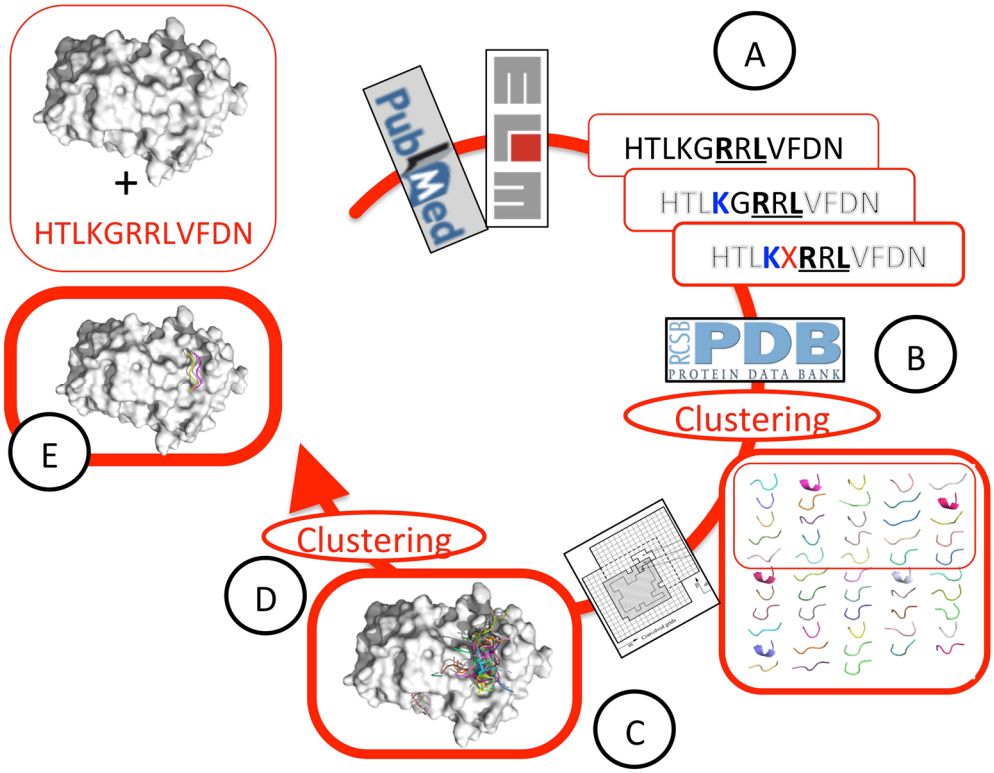


Figure 2.1: Overview of the PeptiDock algorithm. **A)** Given a peptide complex of interest and a reported binding motif, the motif is expanded until a sufficient number of fragments can be extracted from the PDB. **B)** The large pool of fragments is clustered together and ranked by population. **C)** The top 25 cluster centers are docked to the receptor structure. **D)** The lowest 250 structures from each are retained and clustered a final time. **E)** Final peptide conformers are minimized.

2.2.1 Input Structures and Motif Selection

The structure of the free receptor is represented as an independent binding unit that is defined as either a single domain, or repeated, non-decomposable domains (Lavi et al., 2013). Unstructured terminal tails are removed. On the peptide side we start with a peptide sequence that covers the initial motif, and expand the sequence if the original motif is too short (less than 5 residues). We further generalize it by introducing wildcards based on motif information, or restrict it by further expansion. The generalization protocol is iterative, based on available PDB information, to ensure reasonable structural coverage (*i.e.*, a library of 100-1000 conformations). Tables A.1 and A.2 show the motif building process for systems in both the PeptiDB v2 data set and the "Recent PDB" set used for validating the protocol.

Successful definition of a good motif for peptide fragment selection is the critical step of our protocol: A general, non-biased protocol should define a motif that is both loose enough to provide good coverage, and informative enough to enrich for relevant conformations. We start from the peptide sequence of interest and a known motif, and apply the following rules: (1) Start (**S**): Start with a peptide sequence of minimal length of 5 residues (to allow for a motif of 4 and more residues and one or more wild cards if necessary). This sequence should cover the initial motif, and if needed be extended by including additional positions in the peptide sequence. The preferential direction of extension is defined based on the type of residues, according to the following priority: (**Sp**) Polar, (**Sa**) Aromatic, and (**So**) other residues. Small amino acids (GSTA) are not considered for extension, except as a bridge to the next extended residue (e.g. extension

of PXQ motif to PQQATD for the peptide PQQATDD, leading to a 6 residue long starting motif), or if this is the only possible option to extend the motif to the minimal length. This initial sequence will usually result in very few fragment hits in the PDB, and we therefore expand the motif in the following step(s). (2) Expand (**E**): Insert wildcards back (X, or redundant positions of the motif), starting with the smallest residues. Refrain from introducing adjacent wild cards if possible, and do not introduce X at the termini of the peptide. (3) Large (**L**): If more than 1000 hits to PDB structures are found, introduce specific residues back into the motif, starting with the largest residues. If this does not help, try to extend, if possible. (4) Stop when there are between 100 and 1000 hits to PDB structures (or more if further extension of motif is not possible). (5) Complement F/Y (**F**): F & Y show very similar conformational preferences in the backbone dependent rotamer libraries (Ting et al., 2010). Once the motif has been designed and the set of fragments has been extracted, the amino acid sequence is changed back to the actual peptide sequence (using a backbone-dependent rotamer library) (Dunbrack & Karplus, 1993).

2.2.2 Clustering of Fragments

The extracted peptide fragments are clustered using a greedy algorithm and a stringent clustering radius of 0.5 Å (a 2.0 Å cutoff was found to result in too few clusters). The cluster center of each of the top 25 clusters is selected for docking in the next step. This selection method is similar to the 25 top-scoring fragments used in Rosetta *ab initio* modeling (Simons et al., 1997).

2.2.3 Docking of Peptide Fragments

Each of the (up to 25) fragments is docked to the receptor structure using an FFT-based sampling protocol. Two sets of weights for the energy expression (Section 1.2.2) are used: the original set ($w_1 = 1.3$, $w_2 = 160$ and $w_3 = 2.6$) and a set of weights that was recently shown to improve performance for polar-dominated interactions ($w_1 = 4$, $w_2 = 600$, $w_3 = 0$) the pairwise potential is omitted, and consequently the relative electrostatic contribution is increased). For consistency, we use the original set of weights as our default, and report on improvement for polar interactions in Table 2.3.

2.2.4 Selection of Models

The 250 lowest energy poses are retained from each of the 25 separate fragment-docking runs. The resulting 6,250 docking solutions are then clustered based on backbone atoms, using a clustering radius of 3.5 Å, which represents the assumed radius of attraction for peptide-protein docking. Representatives of the top-ranking selected clusters are further locally minimized using the polar hydrogen PARAM19 like forcefield with CHARMM (Brooks et al., 2009). The protocol consists of 500 steps of unconstrained Adapted Basic Newton-Raphson (ABNR) minimization, where both protein and peptide are free to move, followed by the restoration of crystal protein coordinates, and 1000 steps of ABNR minimization of the peptide with the fixed protein. Any final solutions that overlap with domain-domain interfaces are removed.

2.3 Motif-Derived Fragment Sets

The PeptiDock protocol was tested and optimized on a set of 16 complexes, selected from the larger PeptiDB v2 dataset (Lavi et al., 2013). Of the non-redundant complexes, we identified cases for which there was both an unbound receptor structure and a bound complex present in the PDB, and for which there was an interacting motif reported in literature (Table 2.1).

Our motif search is based on the hypothesis that motif information limits the structural configuration space of the peptide conformers, and that some of those conformers are close enough to the native pose to produce a productive docking encounter complex. To illustrate the importance of motif information to our approach, we compare our motif-derived fragment sets to equivalently sized fragment sets selected from random sequences. We then produce histograms of the calculated fitted RMSD between all fragments of a particular set to the native peptide pose. For example, in the cyclin interaction described in Section 2.2, the motif-based fragment set consists of 1051 structures extracted from the PDB using the motif KXRRL (note that the reported motif RXL is found too frequently, so the set of rules discussed in Section 2.2.1 was developed to extend and refine the motifs until a search in the PDB resulted in a comprehensive set of fragment hits). To build the comparison fragment set, we randomly sampled 1051 fragments, 5 residues in length, from all possible 5 residue fragments in the PDB. We extracted the bound KGRRL peptide from the CDC6-cyclin structure (PDB ID 2CCH) for our alpha carbon RMSD calculations.

Table 2.1: Set of peptide-protein complexes from the PeptiDB v2 set. We model a diverse set of 16 domain-motif interactions. The docking protocol was validated on a set of 5 motif-domain complexes recently published in the PDB. For each complex, a bound and free receptor structure is available in the PDB, and an interaction motif has been reported.

	Bound [a]	Free [b]	Peptide [c]	Motif reported [d]
PeptiDB v2 set				
sh2a1 (SH2)	1D4TA	1D1ZA	KSL**TIY**AQ**V**QK	TIYXX[VI] (Poy et al., 1999)
lsb3 sla1 (SH3)	1SSHA	1OOTA	GPP**P**AM**P**A**R**PT	PXXPX[R/K] (Hou, Li, Li, & Wang, 2012)
erbB2 (PDZ)	1MFGA	2H3LA	EYLGLD**VPV**	VXV' (Jaulin-Bastard et al., 2001)
wdr5 (WD40)	2H9MA	2H14A	**ART**KQT	□δR□ (Schuetz et al., 2006)
usp7 (MATH)	2FOJA	2F1WA	GAR**AHSS**	[PA]XXS (Sheng et al., 2006)
p97 N-glycanase (PUB)	2HPLA	2HPJA	DDL**YG**	ϕYX' (D. M. Smith et al., 2007)
traf2 (TRAF)	1CZYA	1CA4A	ace-**PQQAT**DD	PxQ (Devergne et al., 1996)
i-ap1 (BIR)	1JD5A	1JD4B	**AIAY**FIPD	A[VTI][AP][FY] (Srinivasula et al., 2001)
ap2 (appendage domains)	2VJ0A	1B9KA_2	**FEDNF**VP	DXF (Brett, Traub, & Fremont, 2002)
ap2	2VJ0A	1B9KA_1	PKG**W**VT**F**E	WXX[F/W] (Olesen et al., 2008)
pim1 kinase (transferase domain)	2C3IA	2J2IB_2	K**RRRHPS**G	RXRHXS (Bullock, Debreczeni, Amos, Knapp, & Turk, 2005)
cdk2 cyclin	2CCHB	1H1RB	HTLKG**RRL**VFDN	RXL(Cheng et al., 2006)
dystrophin (WW)	1EG4A	1EG3A_1	NMTPYRS**PPPY**VP	PPXY (H. I. Chen & Sudol, 1995)
pcna	1RXZA	1RWZA	KST**Q**AT**L**ER**WF**	QXXϕXXρρ (Warbrick, 1998)
endothiapepsin	1ER8E	4APEA	PFH**LLV**Y	ϕϕ (E.C.3.4.23.22 [e])
gga1 (VHS)	1JWGAC	1JWFA	**DEDLL**HI	DXXLL' (H. J. Chen, Yuan, & Lobel, 1997)
"Recent PDB" set				
G3BP1 (NF2-like domain)	4FCMA	4FCJB	SG**FSF**	FXFG (Clarkson, Kent, & Stewart, 1996)
KEAP1 (Kelch)	3ZGCA	3ZGDA	G**DEETGE**	DXETGE (Kobayashi et al., 2002)
Rev1 (C-terminal domain)	4GK5E	4GK0E	S**FF**DKKRS	FF (Ohashi et al., 2009)
DNAK (C-terminal domain)	4R5IA	4R5JA	NR**LLL**T	LLL (Rudiger, Germeroth, Schneider-Mergener, & Bukau, 1997)
COPI (WD40) [f]	2YNNA	2YNOB	CTF**KTK**TN	KXKXX' (Jackson, Nilsson, & Peterson, 1990)

[a] PDB id of receptor-peptide complex structure; [b] PDB id of free receptor structure, including chain, and number of domain in multi-domain proteins (according to CATH); [c] Region underlined is part of the motif; defined amino acids in the motif are in bold; [d] Motif definitions: ' - c-terminal; δ- small (A,G); γ- no bulky side chains; ϕ – hydrophobic side chain; ρ- aromatic side chain; [e] Flanking cleavage site: Enzyme Nomenclature EC number (http://www.chem.qmul.ac.uk/iubmb/enzyme/); [f] Same peptide binding domain (WD40) as in PeptiDB v2 set, but different peptide motif.

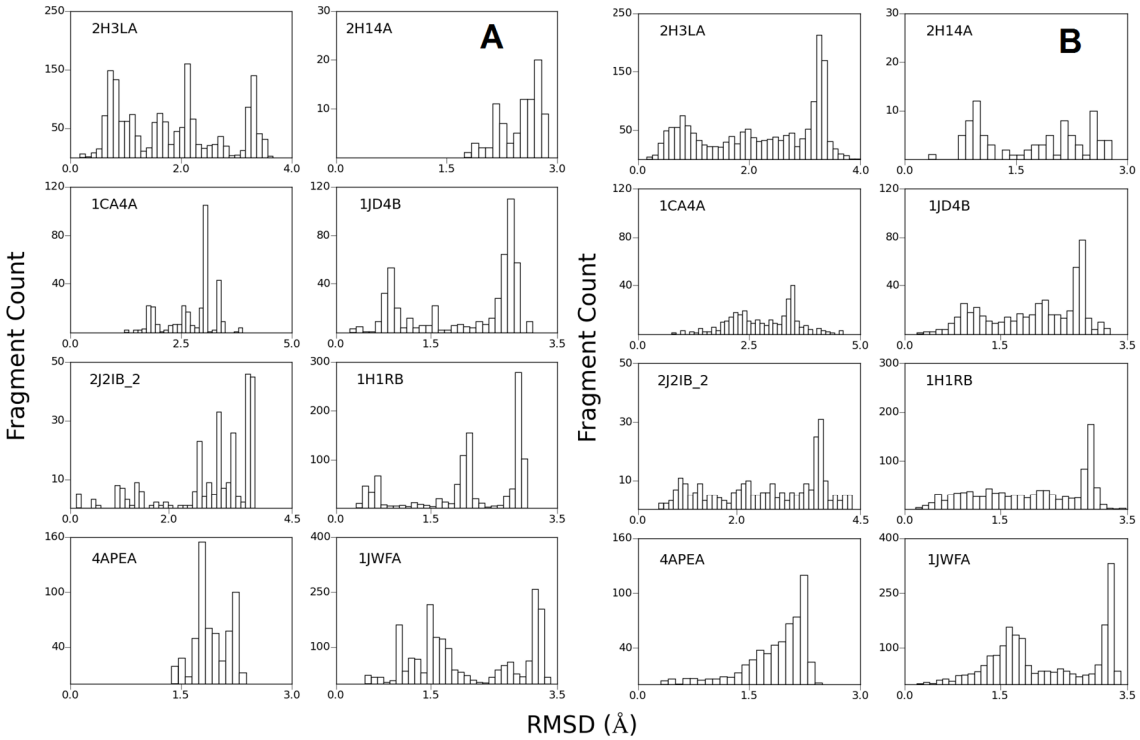


Figure 2.2: Distribution of distances to the native peptide for motif libraries for a subset of Peptide DB v2 cases. **A)** Distributions calculated for motif-based fragment sets. **B)** Distributions for randomly sampled fragment sets.

Figure 2.2 shows the resulting distributions for a subset of the calibration set. Figure 2.2A depicts the RMSD distribution of the motif-based fragment sets, while Figure 2.2B shows the corresponding randomly sampled sets. Two main observations were made from these distributions. The first is that there is a natural clustering of structures in the motif-derived sets. Secondly, in the majority of the tested calibration cases, there is a larger population of structures in a lower RMSD range (i.e. more similar to the native pose) in the motif-derived sets. Both observations lend support to the concept that structure is encoded in the sequence; the motif-based search produces more informative fragment sets.

2.4 PeptiDock Performance

Using a CAPRI-inspired threshold for success, namely defining a near-native conformation if the peptide lies within 4.0 Å backbone RMSD of the native peptide bound to the receptor (*i.e.,* the CAPRI criterion for an acceptable peptide-protein docking prediction), a near-native peptide conformation is found among the top 10 PeptiDock predictions for 11 of the 16 complexes, and all apart from two cases are identified among the top 20 clusters. Similar performance is obtained for the additional validation set: for 4 out of 5 complexes, a conformation similar to bound is extracted using the motif-based search, and for 3 out of the 5 cases a near-native structure is ranked first. The overall detailed assessment of PeptiDock performance is provided in Table 2.2**,** and comparison of docked poses to crystal conformations are shown in Figure 2.3. Predicted poses are shown in yellow. Blue ligands are the native poses. The predicted pose for 1EG3A_1 is shown in pink, as it did not meet the desired 4.0 Å cutoff. The green predicted pose (1B9KA_1) represents a case in in which we used an alternative weight set in our energy expression.

For the 1B9A_1 case and for one case in the validation set (2YNOB, not shown), when the original weight set was used no predicted poses were within 6 Å of the bound peptide. Instead, when we used a set of weights recently shown to improve performance for polar-dominated interactions (the pairwise potential is omitted, and consequently the relative electrostatic contribution is increased), we obtained a near native pose in the 2rd and 3nd predictions, for 2YNOB and 1B9KA_1 respectively (Table 2.3).

Table 2.2: Overall assessment of the motif-domain docking performance. Global docking of motifs identifies for most cases near-native peptide conformations (within 4.0 Å peptide backbone RMSD) among the top-ranking predictions.

Bound [a]	Free [b]	Motif scanned in PDB [c]	Rank [d]	RMSD [e] (Å)
PeptiDB v2 set				
1D4TA	1D1ZA	**TI**[**Y**F]XX[**VI**]	**5**	**3.7**
1SSHA	1OOTA	**P**XM**P**X**R**	**8**	**3.4**
1MFGA	2H3LA	LD**V**X**V**	**3**	**3.9**
2H9MA	2H14A	A**R**[TS]KQ	12	**3.8**
2FOJA	2F1WA	R**[PA]**HX**S**	18	**1.7**
2HPLA	2HPJA	DXL[**Y**F]G	**1**	**3.5**
1CZYA	1CA4A	**P**X**Q**XXDD	**4**	**3.3**
1JD5A	1JD4B	**A[VTI][AP][YF]**[YF]	**2**	**3.5**
2VJ0A	1B9KA_1	**W**XX[**F**Y]E	-	>6.0
2VJ0A	1B9KA_2	[FY]X**D**N[**F**Y]	**5**	**2.4**
2C3IB	2J2IB_2	**R**X**RH**X**S**	**8**	**4.0**
2CCHB	1H1RB	KX**RR**L	**3**	**1.9**
1EG4A	1EG3A_1	RX**PP**X[**Y**F]	**10**	4.1
1RXZA	1RWZA	**Q**XX**[LVI]**XX**W[FY]**	**3**	**3.5**
1ER8E	4APEA	H**[LVI][LVI][LVI]**[YF]	**10**	**2.9**
1JWGAC	1JWFA	**D**XD**LL**	22	**4.0**
"Recent PDB" set				
4FCMB	4FCJB	SX[**F**Y]S[**F**Y]	36	**4.0**
3ZGCA	3ZGDA	**DXETGE**	**10**	**3.9**
4GK5E	4GK0E	**[F**Y][**F**Y]DXK	**2**	**1.7**
4R5IA	4R5JA	NR**LLL**	**2**	**3.8**
2YNNA	2YNOB	**K**T**K**XN	-	>6.0

[a] PDB id of receptor-peptide complex structure; [b] PDB ID of free receptor structure, including chain, and number of domain in multi-domain proteins (according to CATH); [c] Region underlined is part of the motif; defined amino acids in the motif are in bold; [d] Best rank of model within 4.0 Å RMSD; ranks 1-10 in bold; [e] Peptide backbone RMSD; successful predictions (<= 4.0 Å RMSD) are in bold.

Table 2.3: Use of electrostatic-driven potential improves performance for specific cases. Since no near-native structures were sampled for two cases (PeptiDB v2: 2VJ0 & Recent PDB: 2YNN) using the 'Normal' energy function weight set, the cases were re-docked using an electrostatic driven potential.

Bound [a]	Free [b]	Motif used for scanning PDB [c]	Energy function weight	Rank [d]	RMSD [e] (Å)
2VJ0A	1B9KA_1	**W**XX[**F**Y]E	Normal	-	>6.0
			Electrostatic	**3**	**3.9**
2YNNA	2YNOB	**K**T**K**XN	Normal	-	>6.0
			Electrostatic	**2**	**3.9**

[a] PDB id of receptor-peptide complex structure; [b] PDB ID of free receptor structure, including chain, and number of domain in multi-domain proteins (according to CATH); [c] Region underlined is part of the motif; defined amino acids in the motif are in bold; [d] Best rank of model within 4.0 Å RMSD; ranks 1-10 in bold; [e] Peptide backbone RMSD; successful predictions (<= 4.0 Å RMSD) are in bold.

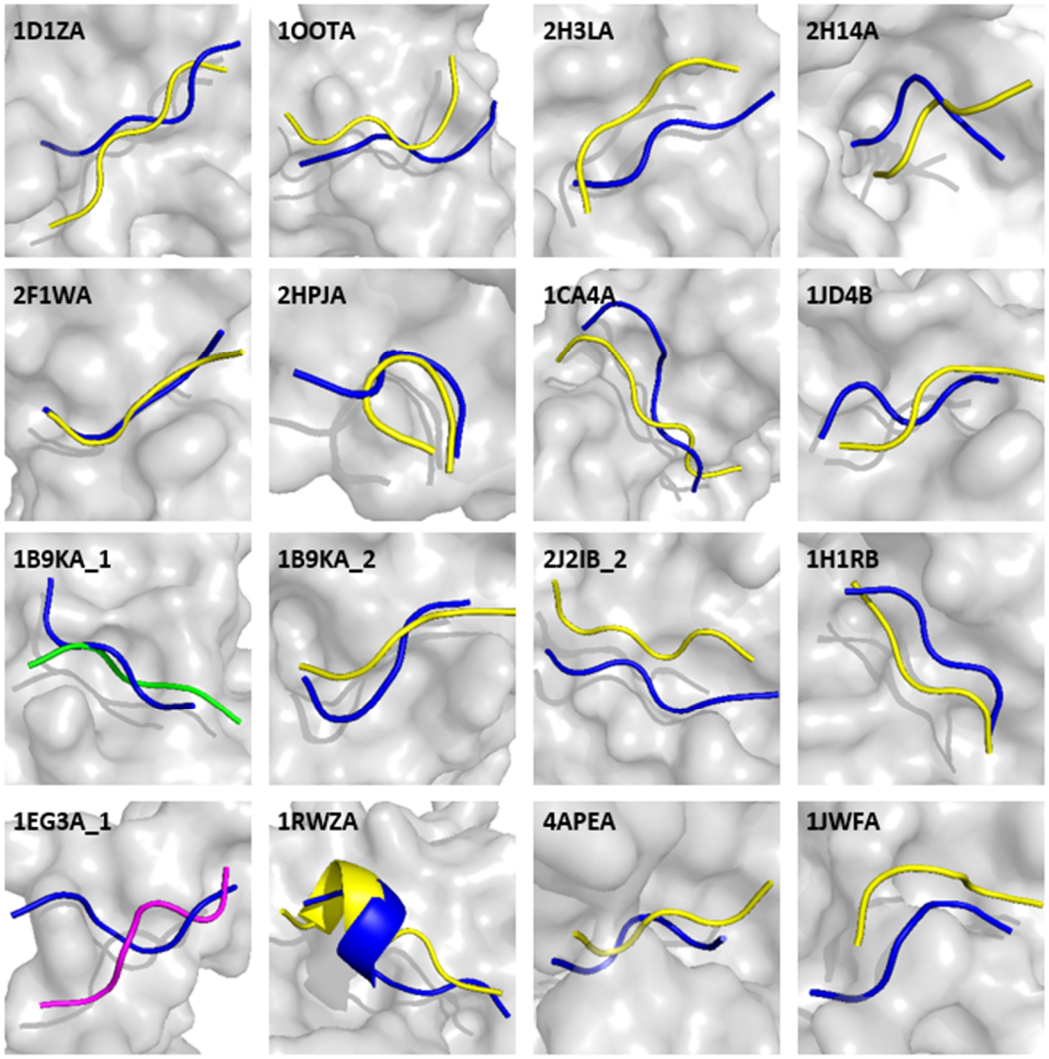


Figure 2.3: Modeled protein-peptide complexes from selected PeptiDB v2 set. Blue is the crystal, native pose, yellow is the acceptable accuracy model. Pink shows the closest non-acceptable accuracy model produced by the approach. Green depicts the acceptable accuracy model for a case (PDB ID 1B9K) in which only the electrostatics coefficient set gave a strong result.

Figure 2.4 shows an example of a successful case and a challenging case. For the latter, the native complex forms hydrogen bonds between the peptide backbone and protein side chains, but lacks strong hydrophobic interactions with the aromatic side chains. The hydrophobic valine points into the solvent (forming crystal contacts with a symmetry mate in the solved structure). Interestingly, in this and in the one additional case for which no near-native structure was sampled (2YNOB), considerable improvement was

obtained by using an electrostatic-driven potential (Table 2.3), indicating that scoring rather than sampling limits performance in these interactions that are dominated by electrostatic attraction (and crystal contacts).

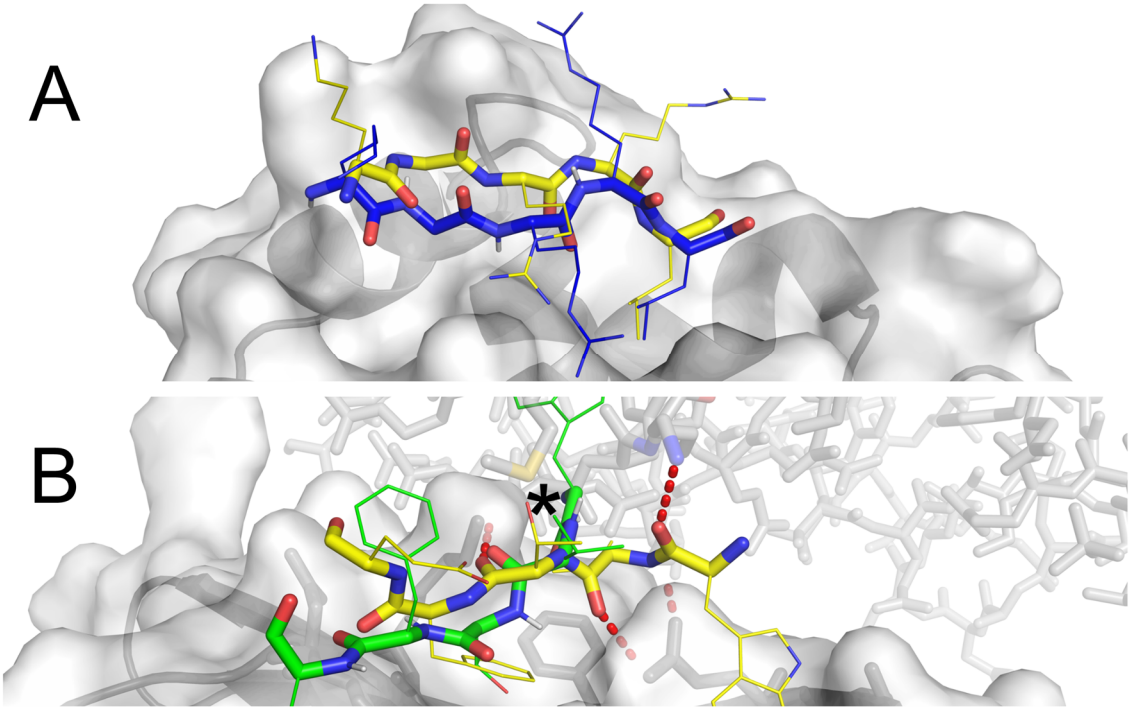


Figure 2.4: Examples for models generated by PeptiDock rigid body docking of peptides to a receptor. Receptor structures are shown in light grey. Yellow structures represent the native peptide pose. **A)** A peptide derived from CDC6 with the sequence motif KGRRL is successfully docked to cycle. The third ranked prediction (dark blue) produces an acceptable accuracy results (1.9 Å backbone RMSD; apo/holo PDB Is: 1H1R/2CCH). **B)** No near native structure is sampled using the standard energy function and weight set when a peptide derived from synaptojanin is docked to the ap2 adaptor (apo/holo PDB IDs: 1B9K/2VJ0). Nevertheless, a 4.0 Å RMSD model (green) is produced when an electrostatics-favored coefficient set is used. This can be explained by the fact that this interaction is dominated by several hydrogen bonds of the peptide backbone (dotted line) in the native complex, but lacks strong hydrophobic interactions with the aromatic side chains, as well as by crystal contacts in the bound conformation. The hydrophobic V6 points into the "solvent", but actually contacts the symmetry mate (interaction is marked with *).

2.5 Using ClusPro PeptiDock

Peptide docking has been implemented as an option of the ClusPro server at https://peptidock.cluspro.org/. Users may also access PeptiDock by clicking the 'Peptide

Docking' tab from the ClusPro home page (https://cluspro.org). To submit a peptide docking job, users may either upload a PDB file as the receptor or indicate the PDB ID and chain to use. For the peptide input, users have two options. The first option, which is the standard usage, involves inputting both a peptide sequence and peptide motif. Alternatively, users may upload their own fragment set (up to 25 fragments), which will bypass the motif-based search. If a user specifies a peptide sequence and motif, they may also list PDB structures to exclude from the fragment search. After the user presses the 'Dock' button at the end of the page, the job will be submitted. If the user has a ClusPro account, they will receive an email upon job completion or if the job has failed to run.

As part of a pre-processing step, it is possible that the given motif will not result in a sufficient number of fragments to proceed. If this happens, the job will fail early, and the user will receive an error that suggests either further specifying or generalizing the motif for re-submission. An additional feature was added to ClusPro to assist users in expanding their motifs. If the user supplies the full peptide sequence and a starting motif, they may use the 'Build Motif' button, which will use the motif extension rules outlined in Section 2.2.1 to iteratively search the PDB until the motif generates between 100 and 1000 hits.

Upon job completion, the user can visit the results page based on their job id. Up to 100 models will be available for download for each of the two weight sets used during docking ('Peptide Balanced' and 'Peptide VdW+Elec'). Under the 'View Model Scores' cluster populations and CHARMM energies are visible for each output model.

3 MODELING ANTIBODY MATURATION

3.1 Introduction

Immunoglobulin (Ig) molecules are key components of both the recognition and effector arms of the adaptive immune system. B cells produce Ig in the form of surface-expressed B-cell receptors (BCR) which bind a specific antigen, therefore signaling B cell activation. Once activated, B cells secrete soluble Ig molecules, antibodies, which bind the same antigens (Alberts, 2002). Antibody molecules are extremely diverse, assembled combinatorially from gene segments. In the presence of the eliciting antigen during the immune response, B cell development continues via affinity maturation, the direct result of somatic hypermutation. Ig genes are mutated and B cells bearing mutated BCRs that have acquired higher affinity are favored for survival, coming to dominate the humoral response and becoming the long-lived plasma cells and memory cells that confer protection in subsequent exposures (Haynes, Kelsoe, Harrison, & Kepler, 2012).

Complementarity determining regions (CDR) on the variable domains of antibodies are responsible for modulating antibody affinity and specificity (Regep, Georges, Shi, Popovic, & Deane, 2017). The CDR H3 loop has proven to be of particular importance as it has been shown to form the most contacts on average with the antigen, while also demonstrating highest structural variation (Clark, Ganesan, Papp, & van Vlijmen, 2006; MacCallum, Martin, & Thornton, 1996). Amino acids altered by somatic hypermutation may drive affinity maturation through different observed mechanisms. For example, mutations that increase shape complementarity of the interface, improve electrostatic interactions, hydrogen bonding, and even promote increased burial of

hydrophobic regions in the interface all improve binding by enthalpic means. Decreasing entropic penalties associated with complex formation is another method in which maturation may improve binding. While somewhat counterintuitive, the formation of a protein complex imposes more order on the system, and therefore binding itself is associated with a decrease in entropy. Typically an increase in enthalpy will counteract this loss, resulting in an overall increase in free energy upon binding. By restricting the flexibility of a structure prior to binding, entropic loss due to binding may be further reduced (Kepler & Wiehe, 2017).

Many challenges are currently faced in antibody design and therapeutic development, particularly the ability to predict antibodies with properties that confer high enough affinity and specificity for their desired target. While widespread screening techniques can be used to assess and optimize a number of different variables, such techniques are in some cases impractical. Instead, computational methods can be used to additionally reduce the search space (Tiller & Tessier, 2015). One study, from 2013, demonstrated that protein-protein docking could be used to generate antibody-antigen structures of high enough quality to be useful in identifying affinity-enhancing residues in a cross-reactive neutralizing antibody to dengue virus (Tharakaraman et al., 2013). Another key study, which focused on antibodies binding influenza hemagluttanin (HA), demonstrated that structural analysis, binding kinetics, and long time-scale MD could be used to study influenza virus antibody evolution in a subject immunized with the 2007 trivalent vaccine. They showed that increased affinity in later antibodies can mostly be attributed to the rigidification of the initially flexible heavy chain CDR (Schmidt et al.,

2013). Along with experimental measures, computational methods may be used to offer additional insight into antibody-antigen interactions that may prove necessary to future antibody design and antibody therapeutic development.

Here, we show that analysis of computational docking results may be used to further understand the driving components behind changes in binding free energy observed throughout antibody maturation. The underlying idea is that, given a protein complex, if we dock the two parts by sampling the conformational space, a higher affinity complex will yield more docked structures close to the native state than one with lower affinity. A similar approach has been used for discriminating between biological and crystallographic dimers with success (Yueh et al., 2017). The complexity introduced by studying multiple structures across a lineage necessitates additional sampling, therefore FFT-based docking is coupled with multiple MD simulations to gain insight into the maturation mechanisms of two different antibody-antigen systems. We focus on previous experimental and computational results, with the goal of exploring the changes in free energy throughout the process of antibody maturation from the UCA (unmated common ancestor) to high affinity binders.

The two systems investigated in this study were selected on the basis of the availability for both bound and unbound antibody structures and the availability of affinity measurements. Structures for influenza HA antibodies were previously deposited in the PDB while structures for the anthrax system were provided by the Kepler group at the Boston University School of Medicine (BUSM). In both cases, the highest affinity structure was available both separately crystallized and as part of the antibody-antigen

complex. Starting from the bound complex, we separated the antigen and used this conformer to establish a theoretical upper limit to the number of near native hits (N) possible with the protein backbone held in place. We then used molecular dynamics (MD) to generate ensembles for the unbound antibody structure, clustered the conformations, calculated cluster probabilities, docked the cluster center, and calculated the average N values, weighted by the probabilities of the MD-generated clusters. We show that influenza antibody maturation is primarily guided by a reduced loss of entropy upon binding, an observation supported by the changes in the on-rate as well as by a previous computational study (Schmidt et al., 2013). Results from applying our protocol to anthrax antibodies appear dramatically different, leading to our conclusion that maturation is primarily enthalpy change driven, a theory supported by the experimental data.

3.2 Methods

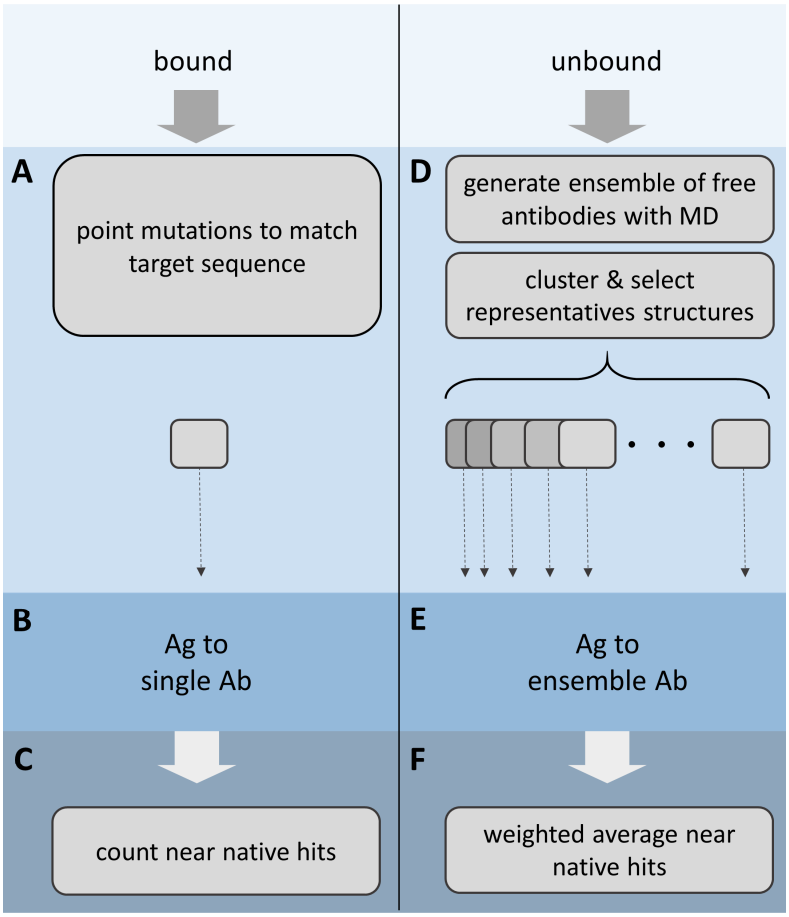


Figure 3.1: Protocol for docking-based antibody maturation assessment. Left panel: Approach for assessing maturation-based contributions by restricting the antibody backbone to that of the bound conformation. **A)** Starting from the bound structure, models of the lower affinity structures are prepared by mutating residues to match the desired sequence. **B)** The resulting structure is then docked with the antigen, using restricted sampling about the antigen center of mass. **C)** Output poses from docking are compared back to the crystal bound and "N", the number of near-native predictions counted. Right panel: Approach for assessing entropic contributions by sampling unbound antibody structures. **D)** Beginning with the crystallized free antibody structure, an ensemble of structures is generated via MD. Trajectory snapshots are clustered and ranked. **E)** Representative conformers are each docked with the antigen in separate docking events. **F)** Predicted poses are compared back to the bound complex, and a weighted average, based on cluster population, of N is calculated.

Our docking-based approach for investigating antibody maturation relies on comparing docking results from two different scenarios. 1) "perfect" binding in which the H3 loop is already fixed in its bound conformation and 2) a more realistic representation

of the flexible H3 loop, based on perturbing the unbound structures by MD simulations. The first approach (Figure 3.1A-3.1C) is based on starting from a high affinity bound antibody structure and establishes a theoretical limit for near native hits resulting from docking the antigen to the antibody structure. Beginning from the crystal complex of the antigen and the highest affinity antibody, we separate and introduce point mutations to the antibody structure to match the sequence of a lower affinity antibody in the lineage (Figure 3.1A). The antigen is then docked to this modeled antibody structure (Figure 3.1B). Since the antibody-antigen interface is known, we restrict sampling about the antigen center of mass. The resulting 1000 lowest energy poses are retained and compared back to the antigen from the known complex (Figure 3.1C). The interface Root Mean Square Deviation (iRMSD) is calculated between the two structures, and considered to be 'near-native' if <= 10 Å from the bound interface. Because the antibody backbone is kept relatively fixed in its bound position, the impact of the point mutations can be better assessed without additional contributions of larger conformational changes that may otherwise be present, as in the case for the influenza virus antibodies studied. The protocol for evaluating unbound antibody structures is outlined in Figure 3.1D-3.1F. Starting with a crystallized free antibody, an ensemble of structures is generated using multiple MD simulations. The resulting trajectory is clustered and the cluster centers are selected for ensemble docking (Figure 3.1D). The antigen from the bound complex is then docked to each antibody (Figure 3.1E) and docking results are compared back to the bound complex. Since multiple structures, represented by cluster centers, are considered for the free antibody, cluster populations are used to calculate a weighted average for N across

dockings: $N = \sum_{i=1}^{K} p_i N_i$, where p_i is the probability of the ith cluster and N_i is the number of near-native docked structures obtained by docking the center of the ith cluster (Figure 3.1F). By comparing the docking results between the modeled bound and unbound structures, we can estimate the importance of changes in enthalpic and entropic contributions to binding free energy across structures from an antibody lineage. By restricting the conformation of the antibody to the bound form, we establish an upper limit for N resulting from docking, while the second approach accounts for larger conformational changes in the structure. The resulting comparison and analysis from these two approaches for two antibody-antigen systems is discussed in Section 3.3.

3.2.1 Mutation of Selected Residues

Starting from the bound antibody structures, residue mutations were introduced with SCWRL4.0, using a backbone-dependent rotamer library for side chain replacement. Only residues which differed between the high and low affinity antibodies were selected for replacement, with the constant residues serving as steric boundaries (Krivov, Shapovalov, & Dunbrack, 2009).

3.2.2 Free Antibody Ensemble Generation

Ten 100 ns simulations, each starting with a random initial velocity, were implemented for each single unbound antibody structure using the 2016-4 GPU version Desmond (Bowers, 2006). Simulations were performed using SPC water and the AMBER99SB-ILDN force field. The Desmond relaxation protocol, as defined in the Maestro GUI, was used at the beginning of each run. After merging the output

trajectories into a single 1 μs aggregate, antibody conformers extracted from trajectory snapshots were clustered based on alpha carbons from the H3 loop, as defined by the Kabat numbering scheme (Wu & Kabat, 1970). The clustering radii varied slightly between trajectories, typically between 0.5 and 1.5 Å, as they were selected with respect to each pairwise RMSD distribution based on previous findings (Kozakov et al., 2005). Clustering was accomplished using the same greedy clustering algorithm implemented in ClusPro (Kozakov et al., 2017), and cluster centers ranked by cluster population.

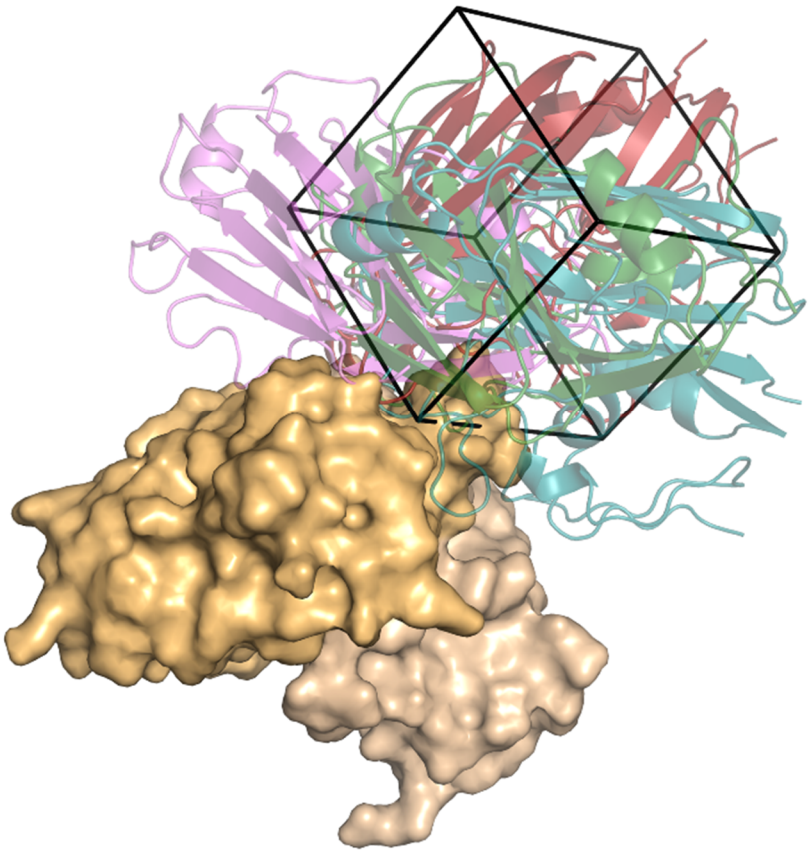


Figure 3.2: Representation of focused docking of the antigen to an antibody structure. The outlined box depicts the restriction of antigen poses (various colors) that are sampled about the antibody structure (wheat).

3.2.3 Rigid Body "Focused" Docking

PIPER (Kozakov et al., 2006) was used to sample potential poses of the antigen about the antibody structure. Since the binding site was known, FFT-based sampling was restricted about the antigen-binding site through the use of a sampling box. The size of the box, with sides equal to 75% of the antigen diameter, was centered on the antigen center of mass (Figure 3.2). PIPER energies were calculated as previously described (Kozakov et al., 2017), using established antibody DARS (Brenke et al., 2012).

3.2.4 Connecting Near Native Hits to Free Energy

The near native count obtained by analyzing lowest energy docking results can be connected to changes in free energy using the following derivation.

$$F = -RTlnQ \tag{3.1}$$

$$Q = \sum_j e^{-E_j/RT} \tag{3.2}$$

$$Q \approx Ne^{\frac{-E}{RT}} \tag{3.3}$$

$$F \approx -RT\ ln\ N + E \tag{3.4}$$

$$\Delta G = \Delta F \approx -RT\ ln\frac{N}{N_{ref}} + (E - E_{ref}) \tag{3.5}$$

$$\Delta G \approx \alpha\ ln\ N + \beta E + \gamma \tag{3.6}$$

We start with the free energy of a complex, given by *(3.1)*, where Q denotes the partition function of the system shown by *(3.2)*. Since the antibody-antigen structure shows only local conformational variations around the native state, we assume that the far-from-native

energy minima represent transitional states or simply false positives due to the particular scoring function used in the docking calculation, and hence restrict consideration to structures that have less than 10 Å RMSD from the X-ray structure of the complex. The dominant part of partition function Q is provided by low energy conformations in this region, and these structures will be used to approximate Q. Furthermore, since the low energy structures are from an energy range that is very narrow relative to the overall energy variation, and the energy values are calculated with considerable error that is comparable to this energy range, it is reasonable to neglect the small differences, thus to assume that $E_j \approx E$ for all j poses. This implies that Q is approximated as shown in *(3.3)*, where E is the average energy in the low energy region and N is the number of structures in this region. Although neglecting the energy differences among the low energy structures seems to be arbitrary, we employ this approximation in our docking server ClusPro with success. Thus, the approximation seems to be adequate (Kozakov et al., 2013). Substituting this approximation into the free energy expression *(3.1)* yields the expression *(3.4)*. However, the calculated free energy values are relative to an unspecified reference state, and the difference in binding free energy between an antibody and an arbitrary reference state can be approximated by *(3.5)*. Note that $\Delta G = \Delta F$, because the volume of the liquid phase system is unchanged. Based on this relationship, we simply dock the antibody structure with its antigen, and use this approximation to estimate the expected changes in the binding free energy. We write *(3.6)* using two unspecified coefficients and a constant, where N is the number of docked structures in a neighborhood of the native state. Using the unspecified α coefficient rather than $-RT$ represents the

uncertainty associated with the volume of the docking box and the criterion used to define near-native structures. While these factors clearly affect the value of N, they are the same for all antibody structures and hence can be used when evaluating the changes in the lnN values. The second coefficient, β, in *(3.6)* is introduced to account for the fact that the energy value E depends on the selected energy scale. Here we use the PIPER energy, which is very useful for generating docked conformations and thus for comparing different structures, but it is not scaled to any measured binding energy. The third coefficient $\gamma = RT\ ln\ N_{ref} - E_{ref}$ acts as a constant, encompassing the terms from the reference state. A similar expression has recently been proposed by the Dill group using a different derivation (Morrone, Perez, MacCallum, & Dill, 2017).

3.2.5 Computational Alanine Scanning

CHARMM (Brooks et al., 2009) was used to calculate the interaction energy of the antibody-antigen complex, and to also estimate the impact of each antibody residue after systematically replacing each residue with an alanine residue. First the antibody-antigen complex was minimized using 1000 steps of unconstrained Adapted Basic Newton-Raphson (ABNR) minimization, using a polar hydrogen PARAM19 like forcefield. A constant dielectric (setting EPS to 20.0) and a distance cutoff of 15 Å for non-bonding interactions was used for the interaction energy calculation. The contribution of each antibody residue was determined by looping through each residue, deleting all atoms except the backbone and beta carbon atoms, and then re-calculating the interaction energy of the complex. The difference between electrostatic and van der Waals terms before and after residue modification was used to calculate ΔE.

3.3 Results and Discussion

3.3.1 CASE 1: The influenza HA antibody system

The crystal structures used for our study of influenza HA antibodies include two bound complexes (CH67: PDB ID 4hkx, CH65: PDB ID 3sm5) and three unbound antibodies (UCA: PDB ID 4hk0, I-2: PDB ID 4hk3, CH67: PDB ID 4hkb). The inferred CH65-CH67 lineage tree is shown in Figure 3.3, along with an overlap of the H3 loops from each structure. Unlike the UCA and I-2 loops, the unbound CH67 loop already aligns with the bound loop conformers of both CH65 and CH67. We refer to both CH65 and CH67 as high affinity binders, as there is a three order of magnitude difference between their K_D values and those of the low affinity binders, UCA and intermediate I-2 (Table 3.1). As previously reported (Schmidt et al., 2013), Surface Plasmon Resonance (SPR) measurements were performed using a Biacore 3000 with the HA head immobilized on the CM5 sensor chip. Purified Fab were injected over the chip at a flow rate of 30 μL/min and mature antibody binding kinetics were fit using a single site binding model. Although a double exponential model was explored for UCA and I-2, the bulk of binding can be attributed to the kinetics described in Table 3.1.

Table 3.1: Binding kinetics for UCA, I-2, CH65, and CH67 Fabs, determined by SPR in previously published study (Schmidt et al., 2013).

Fab	**k_{on} ($M^{-1}s^{-1}$)**	**k_{off} (s^{-1})**	**K_D (µM)**
UCA	$4.4 \pm 0.3 \times 10^3$	0.51 ± 0.03	118 ± 14
I-2	$4.0 \pm 0.5 \times 10^3$	0.56 ± 0.01	142 ± 15
CH65	$1.33 \pm 0.05 \times 10^5$	0.064 ± 0.011	0.49 ± 0.10
CH67	$2.37 \pm 0.14 \times 10^5$	0.086 ± 0.012	0.36 ± 0.04

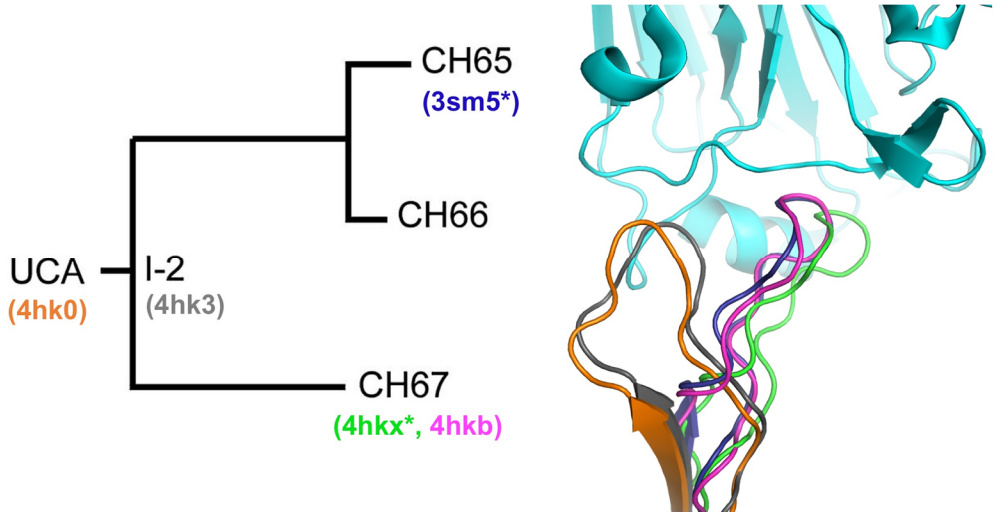


Figure 3.3: Overlap of the CH65-CH67 Fab lineage shows the unbound CH67 (PDB 4hkb) overlapping with both bound structures for CH65, CH67 (PDB IDs 4hkx, 3sm5). The HA antigen is depicted in cyan, based on its placement in the CH67 crystal structure. The H3 loops for lower affinity free antibodies I-2, UCA (PDB IDs 4hk3, 4hk0) are far from the bound pose (Schmidt et al., 2013).

To generate representative ensembles for each antibody, we ran MD simulations starting from the free structure when available. We note that as no unbound structure was available for the CH65 Fab, we begin this simulation using the antibody from the bound CH65-HA complex (PDB ID 3sm5). Since the unbound and bound backbone of CH67 overlap, we assume the same behavior for CH65. Additional preparation was needed for the unbound CH67 Fab (PDB ID 4hkb), which is missing CDR L2; after requesting trajectories from the previous study (Schmidt et al., 2013), we grafted the equivalent loop from their CH67 starting structure. The H3 loop across the trajectories was assessed by calculating the RMSD from each trajectory snapshot back to the H3 loop of the bound CH67 Fab. It is clear from the resulting histograms (Figure 3.4) that there is a large difference in loop distribution when comparing UCA and I-2 ensembles to those generated for structures in which the H3 loop has already converged to the bound conformation (CH65 and CH67).

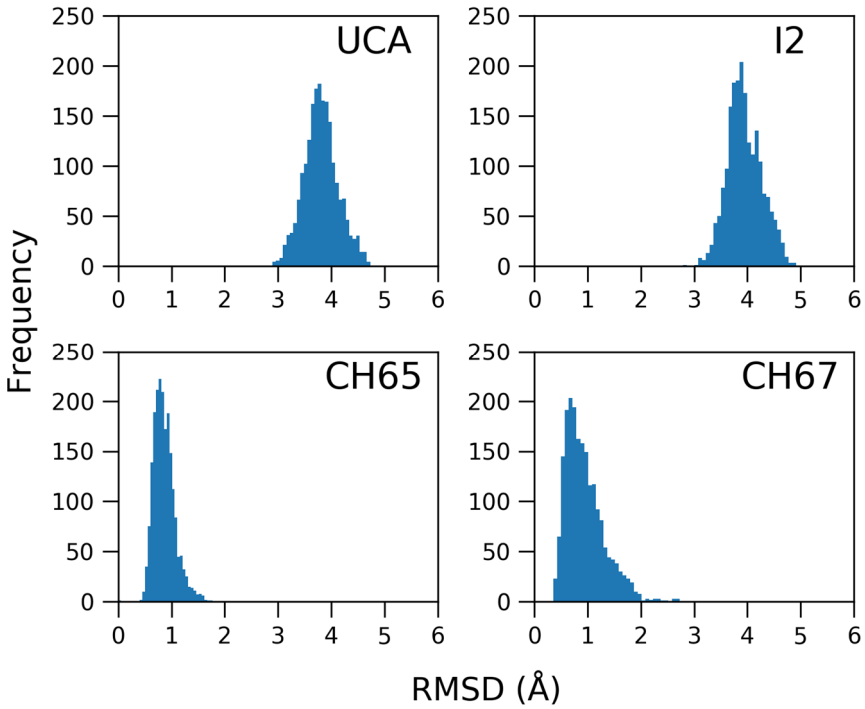


Figure 3.4: Histogram of H3 alpha carbon RMSD values from the 10X100 ns MD simulations for CH65-CH67 lineage members. Each snapshot H3 loop is compared back to the bound H3 loop (PDB ID 4hkx).

After clustering the trajectories and docking the top cluster centers, we used multiple linear regression to fit the weighted near native counts (Table B.2) and PIPER energies (Table B.4) with the ΔG values calculated from experimental K_D values (Table 3.1). We plot predicted ΔG values compared to the experimental ΔG values in Figure 3.5, confirming a strong linear relationship between docking results and ΔG. Predictions were consistent across a broad range of docking parameters. Estimation errors for ΔG values were comparable to experimental errors (Tables B.7–B.9). However, it is clear that the measured binding free energies essentially represent only two points, and thus the observed linearity is fully expected.

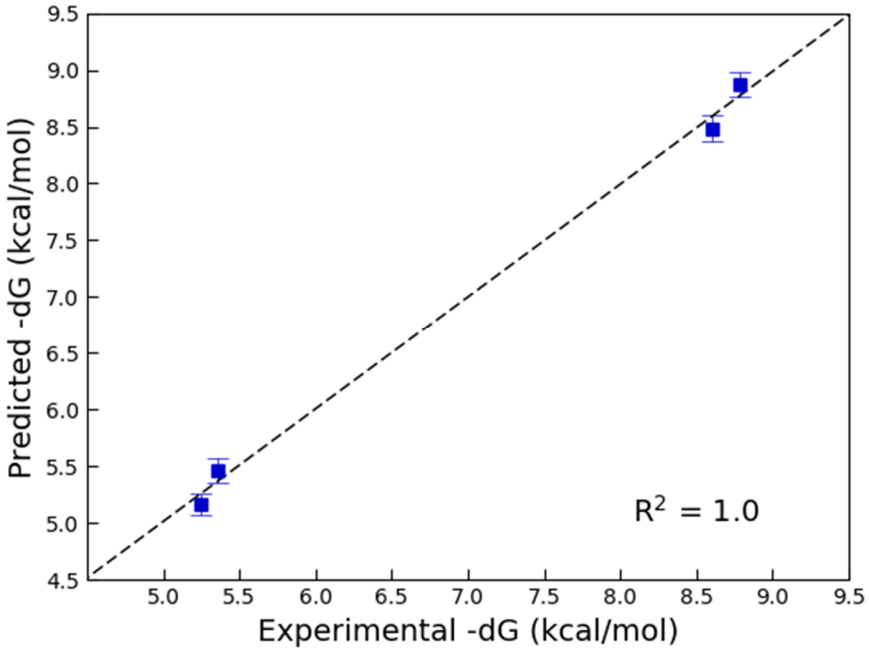


Figure 3.5: -ΔG predicted by near-native hits and PIPER energies, plotted against experimentally measured -ΔG for CH65-67 Fabs. Error bars display standard deviation across 10 docking parameter sets.

It is instructive to study the relationship between –ΔG and only the lnN term, which shows that in the case of influenza hemagglutinin the latter on its own is a good predictor of the free energy change. On Figure 3.6 the upper (horizontal) line is based on the results of docking the different antibodies but fixing the structures in the antigen-bound conformation. By comparing the number of native hits resulting from both methods outlined in Section 3.2, it is immediately noticeable that by allowing for larger conformational changes, as represented by the free antibody ensembles, the result is a distinctly stronger relationship between lnN and –ΔG (Figure 3.6). The larger slope fitting the ensemble data points indicates that a large portion of the change in binding free energy upon antibody maturation is due to conformational changes in the structure. Comparatively, the direct impact from mutations alone, established through our

'theoretical' limit for near native hits, is likely smaller for this system. This figure suggests that the number of docked structures, lnN, remains essentially unchanged if we introduce the mutations observed in the process of maturation but fix the conformations in that of the antigen-bound structure. In contrast, considering the structural ensembles around the X-ray structures, the lnN values substantially drop as we move from the conformations of the high affinity structures, CH65 and CH67, toward the much lower affinity structures UCA and I-2. In fact, the change in the lnN values correlates well with the observed ΔG.

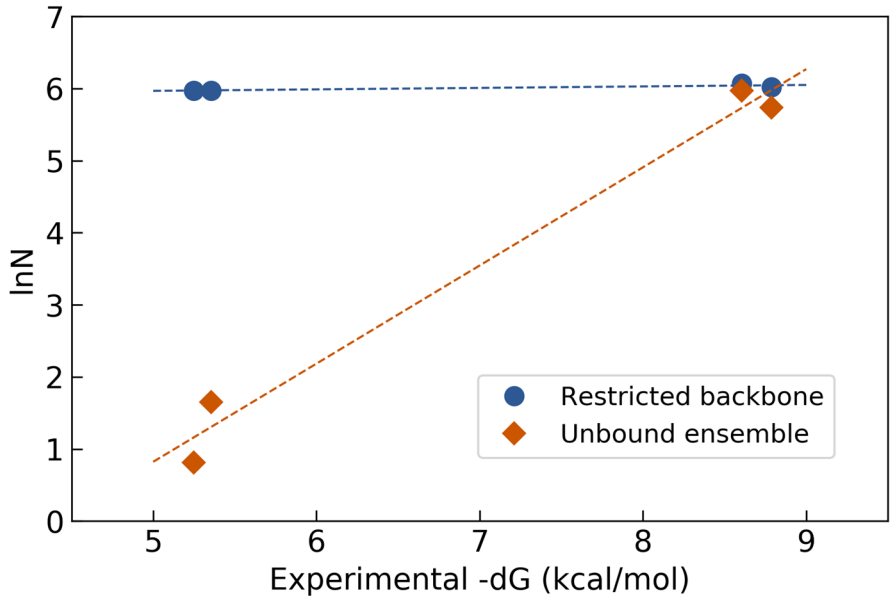


Figure 3.6: If the conformation is fixed, the near-native hits are invariant compared to ensemble docking of MD snapshots. Results from docking Fabs modeled after the bound conformation are shown in blue, with average results from unbound ensemble docking shown in orange.

The two lines in Figure 3.6 suggest that most changes in ΔG are due to changes in the lnN term, and the impact of direct energy changes due to the mutations is much smaller. Since the change in the binding free energy primarily depends on the lnN term, we

conclude that the antibody maturation in this case is entropy driven, whereas the changes in the binding energy have a much smaller impact. This observation is in good agreement with the fact that during maturation the on-rate, k_{on}, increases by two orders of magnitude, whereas the off-rate, k_{off}, reduces only by one order of magnitude, together resulting in the three orders of magnitude change in the binding free energy (Table 3.1). The dominance of the on-rate also suggests that most changes in binding free energy are of entropic (or conformational) origin. The preconfiguration of the H3 loop offers strong evidence that improved binding across the CH65-CH67 lineage is driven by the reduction of entropy loss upon binding by reducing the flexibility of the system.

It also follows that the N values on the lower line in Figure 3.6 represent N_{prod}, the number of "productive" antibody conformations that result in near-native docked structures. Since the structures fixed in the "bound" conformation represent the maximum number of "productive" conformations (N), $N_{prod} / N = p_{prod}$ is the probability of the free antibody being in a "productive" conformation (Figure 3.7B). Notice that Schmidt et al. (2013) determined such probabilities by long direct MD simulations of the unbound antibody structures (Figure 3.7A). Both UCA and I-2 populations have lower probabilities of the loop assuming a bound, or "productive" conformation, in contrast to CH65 and CH67, whose populations are predominantly found with already rigidified loops.

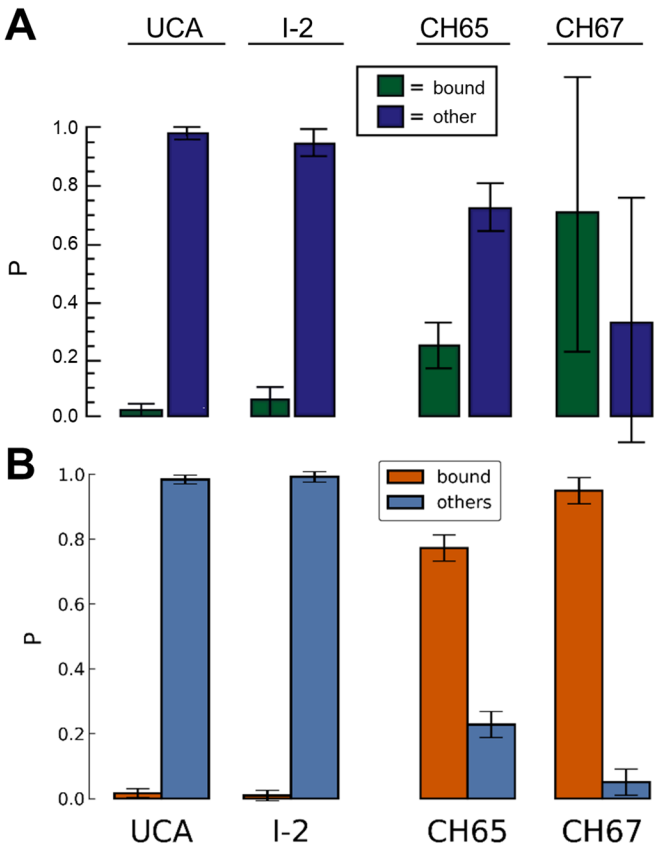


Figure 3.7: Fraction of conformers assuming a bound conformation across UCA, I-2, CH65, CH67 populations. **A)** Probability that CDR H3 loop assumes bound (opposed to any other) conformation, according to lengthy free Fab simulations (Schmidt et al., 2013). **B)** Probability of forming a productive conformation, according to docking analysis. Error bars display standard deviation across 10 docking parameter sets.

3.3.2 CASE 2: The anthrax PA antibody system

Following our results from analyzing the influenza-HA antibody lineage, we were interested to see if the same type of analysis would translate to another system. Colleagues at BUSM provided us with four anthrax PA antibody structures, displaying low (UCA), medium (1558), and high (1184) binding affinities. An unbound structure was available for all three Fabs with a bound structure only provided for the highest affinity antibody (Fab 1184). Figure 3.8 shows the structures aligned to the PA-Fab1184

complex. The loops of the high affinity structure are mostly unchanged between bound and unbound structures. While there do appear to be differences in the H3 loops of the medium affinity and UCA structures, the loop movement is not as dramatic as that seen in the case of the influenza HA antibodies discussed in the previous section.

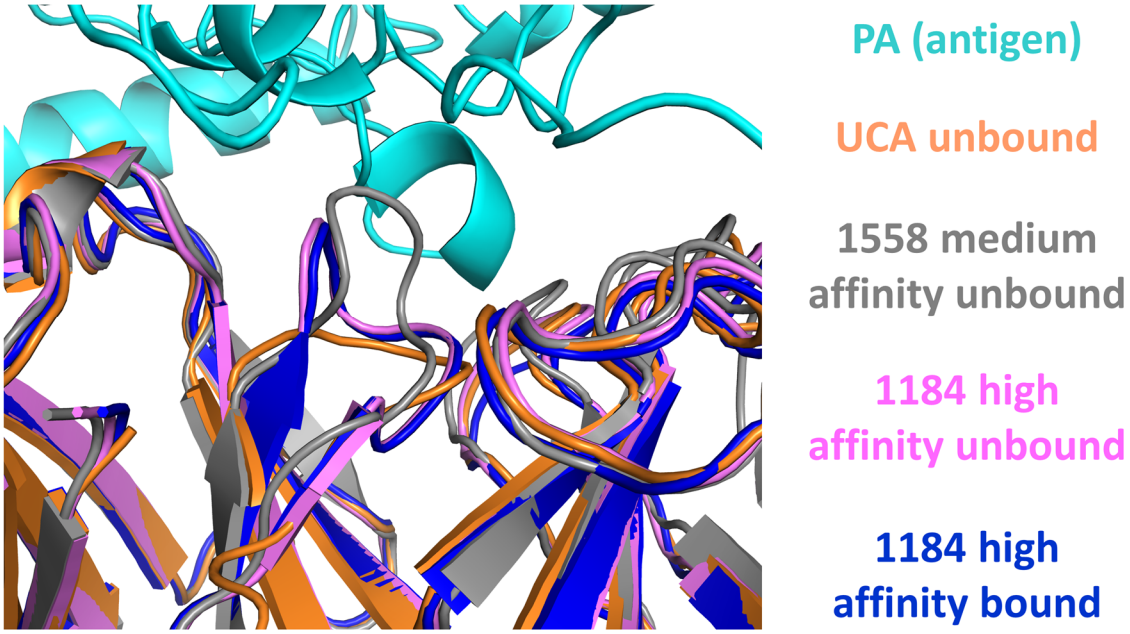


Figure 3.8: The Kepler group at BUSM provided crystal structures and binding data of three antibodies against the anthrax toxin PA, at low (UCA), medium (1558) and high affinities (1184).

When the 1184 bound antibody is used as a starting template and mutations introduced to match first the Fab 1558 sequence and then the UCA sequence (See Figure B.2 for mutations), the near native hits resulting from docking these three structures are unsurprisingly similar (Table B.10). The backbone has remained rigid, and many of the mutations introduced during structure preparation are not within the antibody-antigen interface. Figure 3.9 shows the histograms of H3 loop variation from the MD-generated ensembles of the unbound antibody structures.

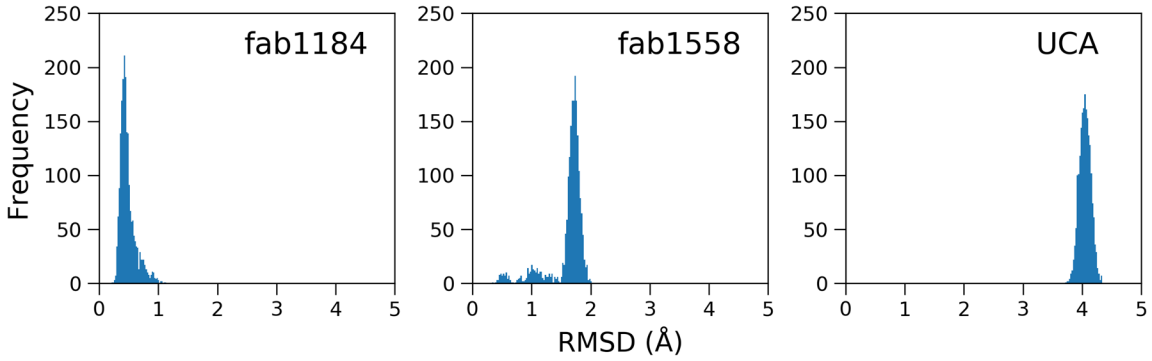


Figure 3.9: Histogram of RMSD values between each snapshot and the bound antibody (1184) are shown for each of the MD-generated ensembles starting from the crystallized unbound antibodies. RMSD was calculated on alpha carbons from the H3 loop.

The distributions show a gradual change in H3 loop RMSD compared to the native, with the medium affinity Fab 1558 having a majority of structures around 2 Å from the bound H3 loop and the UCA starting at roughly 4 Å from the bound. After docking representative structures from the unbound ensembles, we once again compared the weighted N values (Table B.11) to their counterparts from docking with an assumed bound-like conformation. In Figure 3.10, it can be seen that lnN remains essentially unchanged in spite of the substantial change in the binding free energy ΔG and in loop conformations upon antibody maturation, with a small shift between the two docking methods. This suggests that conformational changes in the structure do not play a dominating role in this particular binding mechanism. Since there appears to be little difference between the docking results, the fraction of productive conformations is almost invariant.

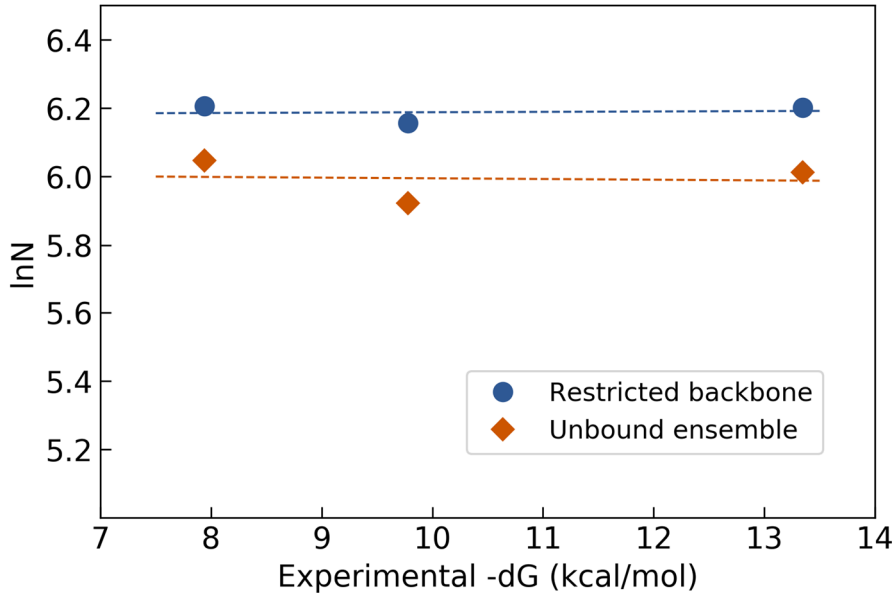


Figure 3.10: Comparison between unbound ensemble and assumed-bound conformation docking of anthrax PA Fabs. Near-native hits alone cannot account for free energy changes in this anti-anthrax system.

Contrary to the influenza HA Fabs, here it appears that changes in binding free energy cannot be explained by loop rigidification. As there appears to be little to no impact on the number of near native hits when fluctuations in the unbound structure are permitted through MD simulations, it seems likely that the driving component for maturation is entirely different. Binding affinity measures, provided by BUSM and determined by SPR, support the notion that maturation is not accomplished by the same mechanism at play for the influenza Fabs. SPR experiments were performed with whole synthesized antibodies immobilized to a COOH2 chip, followed by the injection of recombinant anthrax PA over the chip at a flow rate of 40 μl/min. Such a setup ensures monovalent binding with subsequent analysis carried out using SPRDesign (developed by BUSM collaborators) and a 4-state binding model (Ataca, 2018). Across the three Fabs crystallized, binding affinity data shows a four order of magnitude increase, however, in this case k_{on} shows

only minor changes, within the same order of magnitude across Fabs (Ataca, 2018). Instead, it is k_{off} which displays a three order of magnitude increase between the UCA and Fab 1884 (Table 3.2). For this system, the number of docked antibody-antigen complex structures remains essentially independent of the structure of the loops, instead relying on fine-tuning of antibody residues throughout maturation so as to decrease dissociation after binding. Anthrax PA antibodies are therefore k_{off} driven and can be viewed as more enthalpically driven.

Table 3.2: Binding data for anthrax PA antibodies, determined by SPR in a previous study (Ataca, 2018).

Fab	**k_{on} ($M^{-1}s^{-1}$)**	**k_{off} (s^{-1})**	**K_D (μM)**
UCA	5.99×10^{4}	9.06×10^{-2}	1.5
1558	1.58×10^{4}	1.06×10^{-3}	0.0674
1184	9.75×10^{4}	1.59×10^{-5}	0.000163

3.3.3 Interface Assessment

Computational alanine scanning was used to further investigate the role of particular residues throughout maturation. As described in Section 3.2.5, CHARMM was used to systematically remove all but the backbone of each residue in a bound antibody complex, followed by binding energy calculations. Observations made regarding interface residue interaction energy are used to further investigate the factors at play in these particular interactions.

Tables B.14 and B.15 show the calculated ΔE values after all but the backbone atoms are removed from each residue in Fab CH67 and Fab CH65 complexes. Residues showing the largest positive ΔE values upon removal include VAL 106, ASP 107, and TYR 109, all present in the H3 loop. These residues are consistent across the UCA, I-2, CH65 and CH67 structures analyzed. ASP 95, in CDR L3 is also present in all structures, and results in ΔE values of 2.71 and 3.89 kcal/mol in Fabs CH67 and CH65, respectively. ASP 93, a light chain residue specific to CH67, unsurprisingly shows a high ΔE of 6.38 kcal/mol, half of which is attributed to electrostatic interactions. Mutations that are restricted to mature antibodies (CH65, 66, 67), including ARG 29 of the light chain CDR1 and ASP 31 of the heavy chain H1 loop also display relatively large changes in ΔE, mostly due to forming van der Waals contacts with the antigen. Removing mutations restricted to CH65 (and CH66), including light chain CDR1 ASP 26 and heavy chain CDR2 ASP 57, result in ΔE values of 2.89 and 1.68 respectively.

CHARMM ΔE values calculated for the Fab1184-PA complex are shown in Table B.16. The highest ΔE values were associated with residues engaged in cation-π formation, salt-bridges, hydrogen bonding and critical van der Waals interactions which are made through direct contact the PA antigen. Interface residues such as GLY 50 and MET 35 of the heavy chain, along with PRO 95 of the light chain make only minor contributions according to CHARMM calculations, however their importance to optimizing loop conformation makes them key residues (Ataca, 2018). Both GLY 99 and GLN 98 showed slightly negative ΔE values which suggests each residue on its own may have a minor unfavorable impact on the interface. However, by comparing the antibodies

in the lineage, we see that these two residues work together to allow the antibody loop to adopt a conformation which enables salt bridge formation between the heavy chain ASP 106 and ARG 178 of the antigen (Ataca, 2018).

3.4 Conclusions

Using two antibody-antigen systems, we have shown that protein-protein docking may be used to study antibody maturation mechanisms, provided structures of unbound and bound antibodies are available. For each antibody of a lineage, we assess docking results across two modalities, the first which requires the antibody be modeled on a high-affinity bound structure and the second, which uses an ensemble of free antibody conformations generated by MD simulations. To connect successful docking conformations, known as near native results (N) to changes in binding free energy, we applied a statistical mechanics-based approach, in which we derive a linear relationship between -ΔG, lnN, and PIPER energies. Fitting the data resulting from each docking approach allows us to analyze the impact that conformational changes may have on the system. By first restricting the antibody backbone to its bound conformation, changes in N can only be attributed to the difference in residue side chains, imposed by mutations across a lineage. When the unbound antibody is used for docking, changes in N may be a reflection of larger conformational changes. Comparing the fits of this relationship to data from both unbound and bound docking reveal clear differences between the influenza and anthrax antibodies studied.

Results from four antibodies of the influenza CH65-CH67 lineage demonstrated

the impact that large changes in loop conformations may have on docking; the coefficient for lnN in the linear fit between ΔG and near-native counts dramatically increased when unbound antibody ensembles were docked as opposed to docking the structures restricted to a bound-like backbone. From these results, it is clear that maturation is largely driven by larger loop changes in the structure. A ratio of the two N values per antibody was used to estimate a fraction of productive conformations in the populations. Although an entirely different approach, using significantly longer 25 μs MD simulations, the results from a 2013 study strongly supported our own predictions. Authors had also commented that preconfiguration of the loop acted as a means of reducing binding loss due to entropy (Schmidt et al., 2013). When the same approach was applied to three antibodies from an anthrax PA antibody lineage, the resulting fits were largely unchanged between the two docking methodologies. Generating unbound free antibody ensembles yielded similar near native hits to their rigid counterparts, suggesting that this lineage undergoes maturation through a different mechanism than that of influenza antibodies. Inspection of the structures also suggests that loops, while undergoing some fluctuation, do not modulate the interaction in the same way. Retrospectively, the computational results ultimately came down to the inherent difference in the interface of the flu and anthrax antibodies. The epitope on HA is the sialic acid binding site, which is a relatively small and narrow pocket, and therefore it is selective of the conformation of the H3 loop upon binding. In contrast, the anthrax antigen and antibody interface is comparatively flat, and hence it is less sensitive to changes in the conformation of the hypervariable loop.

The structural evolution of influenza hemagglutinin antibodies suggests that reducing entropy loss upon binding is the main factor driving antibody maturation. This result agrees with the accepted hypothesis that maturation rigidifies the CDR loops, and is supported by the observation that the mutations increase the on-rate by two orders of magnitude, while the off-rate decreases only by an order of magnitude. Anthrax PA antibodies undergo maturation that relies less on reducing entropic loss upon binding, and is instead driven by enthalpic contributions. SPR data from a prior study confirms that this maturation is driven by the off-rate which displays a three order of magnitude increase, with only a twofold change in the on-rate (Ataca, 2018). This analysis successfully demonstrates that protein-protein docking, in conjunction with MD simulations, is capable of providing researchers with useful insights into the maturation driving factors for different antibody systems. While restricted by the availability of structures, results imply that docking derived near-native hits may be useful for the identification of the main factors contributing to changes in the binding free energy upon antibody maturation. As demonstrated, the two antibody-antigen systems studied here represent two very different mechanisms of maturation.

4 TEMPLATE-BASED MODELING

4.1 Introduction

Computing the structures of protein complexes has been one of the central but challenging problems in computational structural biology (Nussinov, Papin, & Vakser, 2017). Even for relatively rigid proteins it is difficult to explore the 6D rotational-conformational space of mutual orientations potentially sampled by a pair of proteins as they interact through complementary patches on their surfaces. Predicting the association of proteins is further complicated by flexibility. Proteins are not static objects; they constantly interconvert between conformers of varying energies (Nussinov et al., 2017).

In spite of the complexity of the problem, a variety of docking methods, including some easy-to-use servers, are currently available for predicting the structures of protein-protein complexes. The choice of the method used depends on the nature of the docking problem. “Free” docking methods can be used if X-ray structures are available for all proteins to be docked or for their very close homologs. However, the number of structures of protein complexes has been increased in the PDB. Knowledge of complex structures makes prediction of related protein complexes amenable to template-based and homology modeling methods, even when the structures of component proteins are not available (Figure 4.1A).

The prediction of protein complexes remains an active and challenging field. A relatively small number of heteromeric complexes are available in the PDB compared to their individually crystallized components. Due to low complex availability, docking servers and modeling tools are often employed to predict such interactions. While the

number of structures deposited in the PDB continues to grow, reportedly at a yearly rate of ~10% (Rose et al., 2017), there is a continued need for docking and modeling tools that have the capability to handle larger structures and the ability to account for more complicated experimental data (Carroni & Saibil, 2016).

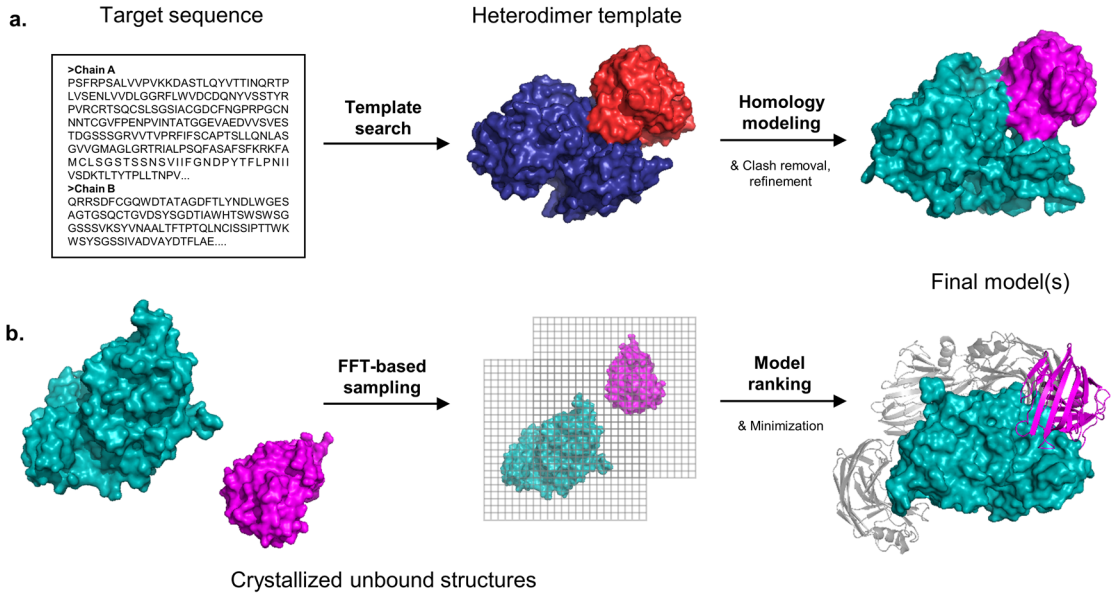


Figure 4.1: General comparison of template-based and free docking methods for an example heterodimer target. **A)** A template-based method begins with the target sequences, using a template search to identify an existing heterodimer template from the PDB. Homology modeling is used to map the target sequence onto the template structure. The model is refined to its final form. **B)** In free docking methods, the two components must be individually crystallized. Billions of protein conformations are evaluated, often through the use of an FFT-based algorithm. Final models of the heterodimer are ranked and minimized.

Strategies for predicting protein complexes typically fall into two categories: free docking and template-based modeling. Free docking techniques take structural inputs, sample potential orientations and rotations of the two structures, and often filter or rank resulting poses using a scoring function (Figure 4.1B). Template-based modeling uses the protein sequence to search available databases for related proteins to use as structural templates (Figure 4.1A). Template availability of complexes is often considered a

limiting factor in this approach. However, it has been shown that nearly all known protein-protein complexes can be modeled, provided there are strong homologs deposited in the PDB for each of their components (Kundrotas, Zhu, Janin, & Vakser, 2012).

Community-wide assessments such as CASP (Critical Assessment of protein Structure Prediction) and CAPRI (Critical Assessment of Predicted Interactions) serve as important platforms to not only evaluate the performance of current structural prediction servers, but to also challenge participants with unique targets and encourage advances in server methodologies. ClusPro v2, a participant in CAPRI since Round 13, including all joint CASP-CAPRI rounds, has repeatedly ranked among the top servers (Lensink et al., 2018; Lensink, Velankar, & Wodak, 2017; Lensink & Wodak, 2013). The ClusPro server performs three main steps: (1) Fast Fourier Transform (FFT)-based rigid-body sampling, (2) ranking via cluster population, and (3) energy minimization to remove steric clashes. This algorithm has proven itself to be an effective method for a variety of targets. Features added to ClusPro, including the ability to account for Small Angle X-ray Scattering (SAXS) profiles (Ignatov, Kazennov, & Kozakov, 2018; Xia et al., 2015) and pairwise distance restraints (Xia, Vajda, & Kozakov, 2016), have been motivated by specific CAPRI targets, where this information was made available to predictors. More recently, CASP-CAPRI targets inspired the addition of a tool for the discrimination between biological and crystallographic dimers (Vajda et al., 2017).

While early CAPRI targets presented participants with crystal structures for one or even both complex subunits, later rounds, including those combined with CASP rounds, have required participants to use homology models as representative subunit

structures. ClusPro, as a free docking server, has not previously incorporated homology modeling into its automated protocol, instead either relying on structural predictions from CASP participants or on homology models generated by the HHPred web server (Zimmermann et al., 2018). These models were then submitted for ClusPro docking in hopes of producing a near-native interface. This method was employed in CASP12, and produced an acceptable or better solution within the top 10 submitted models in 7 of 10 targets, 3 of which were of medium quality (Lensink et al., 2018). However, compared to other servers employing template-based modeling approaches, ClusPro generally produced fewer high accuracy results.

In a retrospective study (Porter, Desta, Kozakov, & Vajda, 2019) on 15 validated homodimers from CASP11-CAPRI and CASP12-CAPRI assessments, it was shown that template-based modeling greatly increased the reliability of predictions for the 12 designated easy targets. When templates were available, higher quality predictions were produced via template-based modeling alone (Table 4.1). Interestingly, for one of the three difficult targets (T72/T0770, T86/T0815, and T116/T0893) that did not have suitable templates, global docking yielded an acceptable model, whereas the template-based method produced none. These findings further support the need for ClusPro to incorporate both free docking and template-based searches into its predictive strategy.

Table 4.1: Number of models by template-based docking **(A)**, global free docking **(B)**, and focused free docking **(C)**.

CAPRI ID	PDB ID	A	B	C
T69	4Q34	1*	1*	1*
T72	4Q69	0	1*	0
T75	4Q9A	3*/2**	2*	3*/2**
T79	5A49	2*/2**	2*	2*/2**
T80	4PIW	10*/6**	10*/1**	10*/3**
T85	4WJI	8*/5**	8*/3**	8*/4**
T86	4U13	0	0	0
T87	4WBT	9*/4**	9*	9*/2**
T90	4XAU	10*/4**	10*/3**	8*/3**
T91	4URJ	6*/2**	4*	5*/1**
T92	4W66	2*	3*	10*
T93	4XRR	8*/5**	9*/2**	8*/3**
T94	4W9R	1*	0	1*
T116	5IDJ	0	0	0
T119	5YVS	9*/2**	9*/1**	8*/2**
TOTAL		69*/32**	68*/10**	73*/22**

* Acceptable or better predictions
** Medium or better predictions

Here we present the template-based modeling feature of ClusPro; this protocol will be discussed in reference to T152/T1003, a homodimer, as an example of a straightforward case where many good templates are available, followed by the discussion of targets T142/H0974 and T141/T0976 that required more complicated modeling steps. We will discuss plans to further expand ClusPro TBM by incorporating new template selection techniques, modeling/docking decision making, and inputs for experimental data such as SAXS profiles and Electron Microscopy (EM) density maps to help guide the modeling process.

4.2 Methods

For each target, we first attempted to perform template search/homology modeling using the novel ClusPro template-based modeling functionality. Whenever templates were available, the resulting models were submitted as target predictions. When no templates were identified, we used ClusPro free docking capabilities to generate the models.

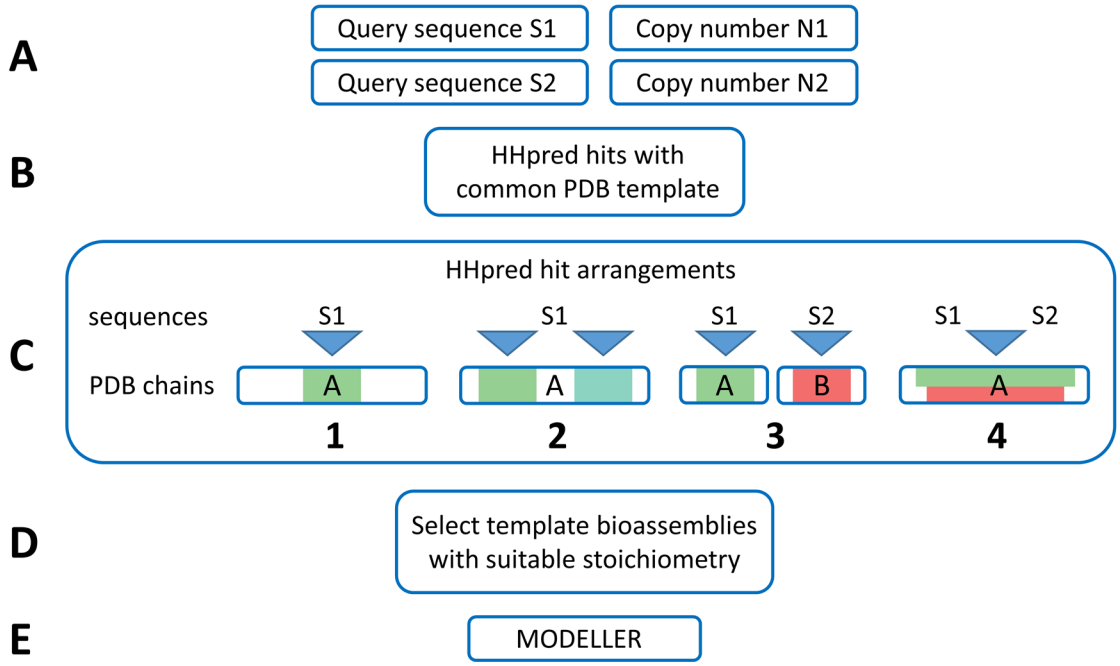


Figure 4.2: General outline of the ClusPro template-based modeling (TBM) protocol**. A)** ClusPro TBM takes component sequences and their corresponding copy numbers in the modeled assembly as inputs. **B)** HHpred is used to find potential templates for each query sequence and HHpred hits sharing a common structural template are identified. **C)** The HHpred hits are combined to obtain potential assembly-generating arrangements for each template structure (arrangements in the figure are examples, and are not necessarily generated as a part of a single server job). **D)** The arrangements are evaluated on their ability to produce a model with user-specified stoichiometry based on biological assemblies specified in the template PDB file. **E)** Arrangements passing the stoichiometry filter are used to construct the assembly models using MODELLER.

4.2.1 Template-based Modeling

As inputs, the ClusPro template-based modeling module requires a set of sequences in FASTA format and the stoichiometry of the assembly to be modeled (Figure 4.2A). Potential structural templates for each query sequence are identified using a local installation of HHpred (Soding, Biegert, & Lupas, 2005), which runs HHblits and HHsearch using default settings and searches through the latest versions of uniprot20 and pdb70 databases, respectively. HHpred results are filtered by HHpred probability (>90%) and query sequence coverage (>50%), after which PDB structures that have at least one HHpred hit for each of the unique query sequences are identified (Figure 4.2B). We term such a "shared" PDB file and a set of HHpred results pointing to it a "common template". It should be noted that a single PDB template can accommodate several hits of the same query sequence in different positions and chains, and these multiple hits are included in the "common template".

Since a single "common template" can have multiple HHpred hits from a single query sequence, various combinations of hits can be used to construct different "hit arrangements", potentially leading to different assembly models. We combinatorially generate all such possible "hit arrangements", with the requirement that at least one hit for each query sequence is present in the arrangement. The resultant arrangements can represent a variety of query-template relationships (Figure 4.2C).

For each generated arrangement, we iterate through all biological assemblies specified in the shared template PDB structure and check whether this template assembly can be used to produce a model of user-specified stoichiometry given a particular "hit

arrangement" (Figure 4.2C,D). Figure 4.2C provides examples of some representative HHpred hit arrangements for homo and heterooligomeric multimers that were present as targets in CASP13. The leftmost in Figure 4.2C-1 depicts the most straightforward homooligomeric case, where a single query sequence is aligned to a separate chain in the template PDB. If the target in this case is an A2 complex, the required template stoichiometry for this arrangement is also A2. Figure 4.2C-2 shows a more complicated case, in which a single query sequence aligns to multiple regions within the same chain of a template PDB. This relationship is likely to occur when the template protein is a result of gene duplication and fusion as seen in T141/T0976 (see Section 4.3.3). If the target is a dimer in this case, the template should be a simple monomer. Figure 4.2C-3, represents the simplest heterooligomeric case, where two query sequences align to two different template chains. Similar to the simplest homooligomeric case, the template stoichiometry here should match the stoichiometry of the modeled assembly. For example, a simple heterodimer target needs a heterodimer template to be modeled correctly. Finally, Figure 4.2C-4 shows a case where different query sequences align to the same region within the same template chain. Such an outcome is likely to happen if query sequences are related to each other (like in T142/H0974). The template in such a case needs to be a homomultimer with the number of subunits equal to the combined number of S1 and S2 subunits of the target assembly. For instance, the template needs to be a homodimer if the target is a heterodimer.

If the assembly template matches the stoichiometry requirements, it is used to construct an assembly model. For each query sequence, the top-ranking HHpred hit from

the arrangement is used to construct the model of the monomer. The target sequence is modeled onto a single chain of the homo-multimeric template using MODELLER (Webb & Sali, 2016). Regions of the target which are not aligned to a template sequence are removed to avoid the addition of unstructured loops into the model, while aligned portions of the target are built with the same backbone. Once produced, the monomer model is copied and aligned to other locations of the multimer template based on HHpred hits for the given query sequence present in the "hit arrangement", followed by interface minimization. These models are ranked based on the averaged ranks of HHpred hits used to build the models of the monomers. The server also provides an advanced option for manually curated homology modeling, allowing the users to upload their own templates and alignments.

4.2.2 Free Docking

When assembly templates were not available, we used ClusPro free docking capabilities to generate the predictions. Monomer models were constructed using the HHpred server (Zimmermann et al., 2018). When the HHpred server did not produce any models, we used monomer models as predicted by the CASP servers. The free docking pipeline was as described previously (Kozakov et al., 2017). Briefly, the FFT-based PIPER protein docking program (Kozakov et al., 2006) is used to generate 1000 low-energy poses which are then clustered together using a 9 Å clustering radius. Clusters are ranked by their populations and cluster centers are treated as complex models. These models are subjected to local energy minimization by CHARMM (Brooks et al., 2009) and returned as final server predictions.

4.2.3 Co-minimization via CHARMM

Both for template-based and free docking models, CHARMM was used to co-minimize the modeled interface using the PARAM19 force field with polar hydrogens only. ClusPro TBM complexes were first minimized using 1000 steps of Adapted Basic Newton-Raphson (ABNR) minimization, with harmonic restraints set on the alpha carbons, to remove larger clashes that would otherwise occur in the interface. The harmonic restraints were then removed, followed by 1000 steps of unconstrained ABNR minimization. A constant dielectric was used during the energy calculations, and a distance cutoff of 15 Å when considering non-bonding interactions.

4.3 Results and Discussion

Similarly to previous CASP-CAPRI rounds, the majority of targets in CASP13-CAPRI were homomeric complexes. As the majority of biological assemblies in the PDB are also homomeric, the targets of this type, compared to heteromeric targets, are much more likely to have structural templates readily available. Additionally, our experience with previous CASP-CAPRI rounds suggests that structural templates, when present, enable the construction of higher accuracy models than those generated using free docking approaches.

Motivated by these observations, we utilized a template-first pipeline to predict the structures of target complexes, in which we performed template-based modeling whenever assembly templates were available, and used free docking otherwise. Here we

describe several representative cases from the last CASP-CAPRI round that highlight the new server functionality.

4.3.1 CASE 1: T152/T1003, a simple homodimer

Target T152/T1003, 5'-Aminolevulinate Synthase 2, serves as an ideal template-based modeling case, having A2 stoichiometry and an abundance of available templates with high sequence similarity (~up to 48% sequence identity). Within the first 11 structures suggested by an HHpred template search, 10 were available as dimerized biological assemblies. All were listed with reported HHpred probabilities of 100, which exceeds the value (0.95) considered high enough to indicate certain homology between query and template sequences (Zimmermann et al., 2018). As our protocol describes, for each of these templates, MODELLER produced monomer models, which were then copied onto each unit in the corresponding template assembly. The first five models submitted by ClusPro TBM were based on templates 2W8T, 2X8Y, 5TXR, 3TQX, and 2BWN, and were all evaluated as medium quality models. Figure 4.3 shows a representative model produced by ClusPro TBM.

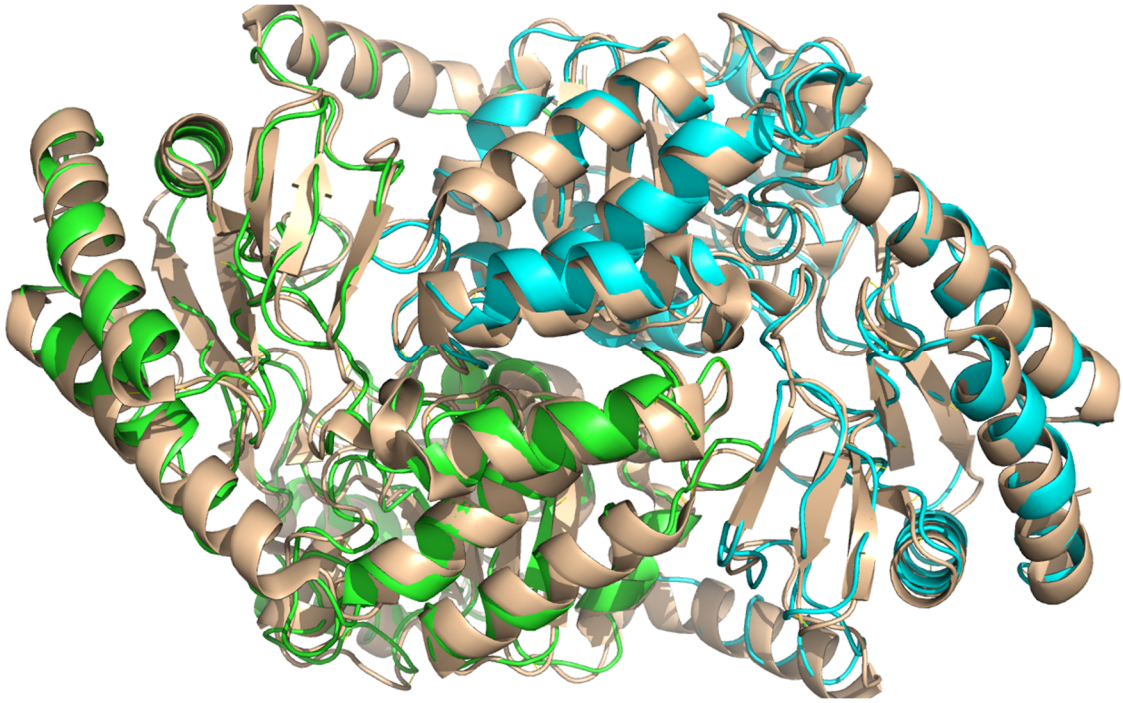


Figure 4.3: Model of T152/T1003 (green and cyan) overlapped with its homodimeric template (wheat, PDB 2W8T).

4.3.2 CASE 2: T142/H0974, heterodimer based on homodimer

Target T142/H0974 with A1B1 stoichiometry is an example of successful modeling of a heterodimer using homodimeric complexes of remote homologs as templates. The target represented a heterocomplex of DNA binding proteins, and the sequences of the target subunits were homologous to each other. Predictably, the templates identified by HHpred were predominantly homodimeric complexes, and HHpred hits for target subunits were usually mapped to the same chain of the template structure. While handling of such templates is trivial when done manually, it is less straightforward in the automatic regime. The arrangement procedure implemented in ClusPro TBM was successful in automatically determining the homomeric templates as

having suitable stoichiometry and correctly aligning the monomer models onto different chains of the template assembly, producing three medium quality and one acceptable models. Figure 4.4 shows an example homomeric template (PDB 4RYK) together with the predicted complex model based on it.

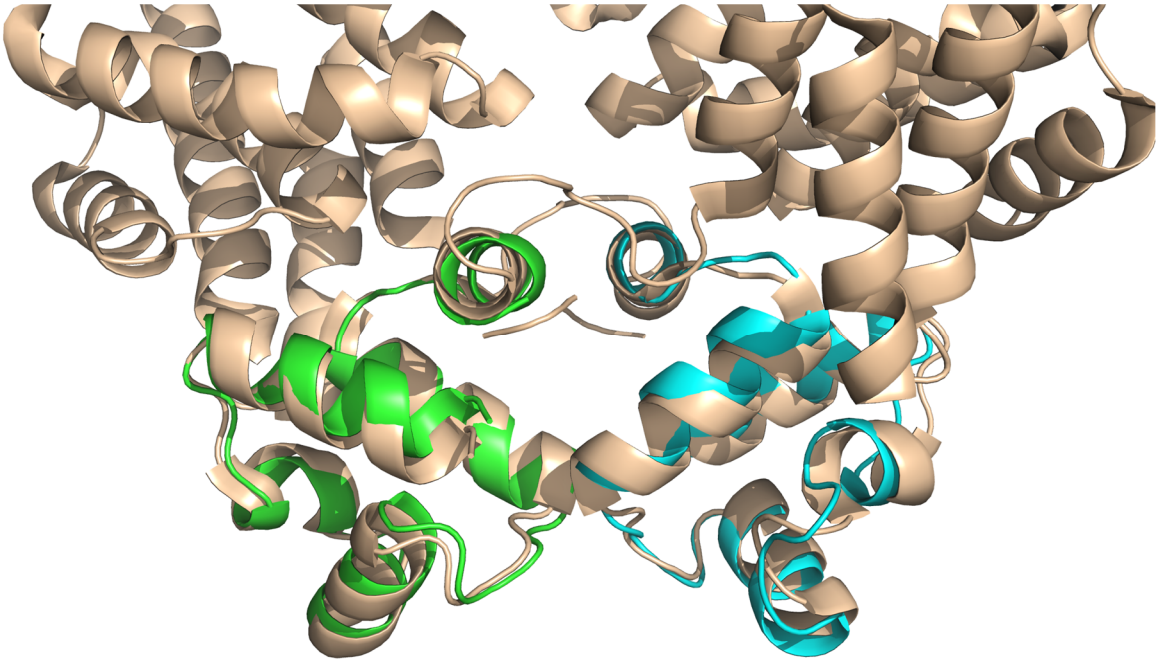


Figure 4.4: A model of T142/H0974 (green and cyan) overlapped with its homodimeric template (wheat, PDB 4RYK).

4.3.3 CASE 3: T141/T0976, homodimer based on monomer

Another notable docking target was a homodimer formed by the Rhodanese-like family protein SCHU S4. The only productive template identified by HHpred was, in fact, a monomeric fusion protein that had appeared in three different HHpred hits. Since the hit arrangement procedure of ClusPro TBM allows for model construction from multiple HHpred hits, two of these HHpred hits using non-overlapping regions of the

template chain were used by the server as monomer alignment sites to construct an acceptable quality model (Figure 4.5).

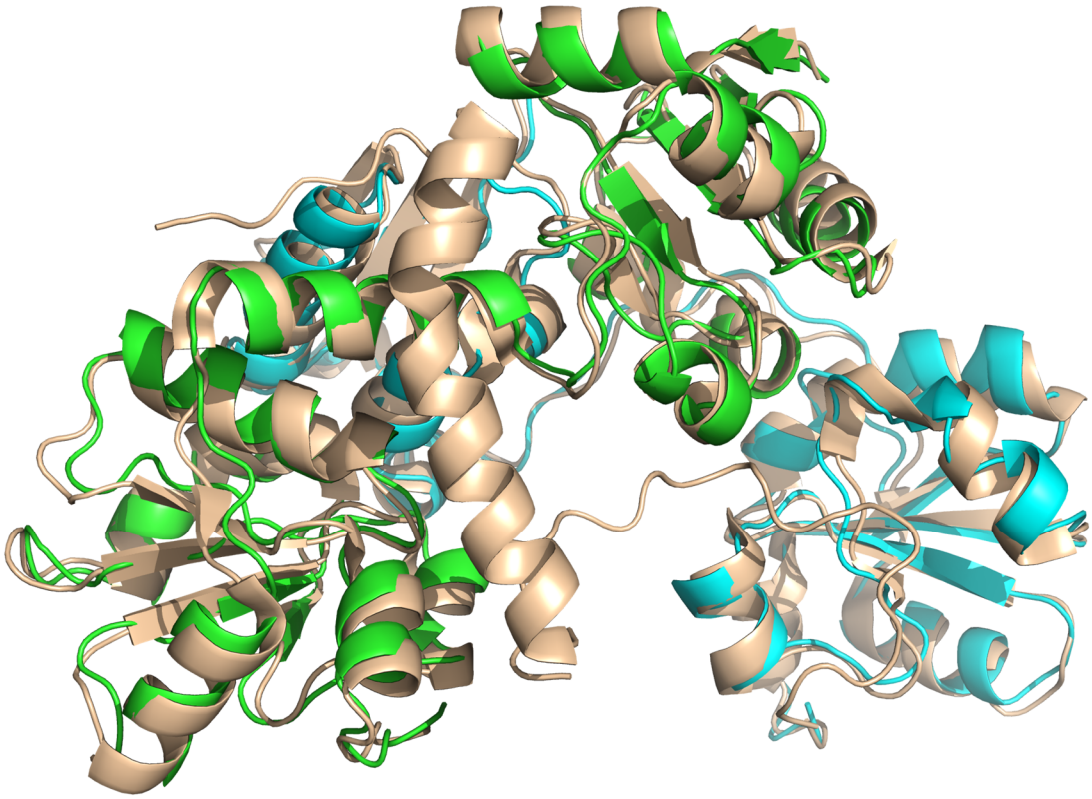


Figure 4.5: Modeled subunits (green and cyan) of T141/T0976 aligned to different locations on the same chain of the template protein (wheat, PDB 1YT8).

4.3.4 Future Directions

Our template-based modeling demonstrated promising results in CASP13, however, it can be further improved by implementing more sophisticated template searches, adding follow-up free docking steps, and incorporating experimental data. In the following sections, we discuss existing limitations in the methodology and propose potential enhancements to the server.

As demonstrated by target T137/T0965, routine template-based modeling may sometimes lead to low-quality models. Following an HHpred search of the provided sequence, numerous high-probability (>0.95) homodimer templates are given. The target appears an easy one, with a noticeable agreement between the top ten template interfaces. However, none of the models produced by ClusPro were evaluated as acceptable or better compared to the crystal complex (PDB 6D2V). The template complex for this target has correct contact location, however, one of the subunits is 130 degrees rotated with respect to the target structure (see Figure 4.6).

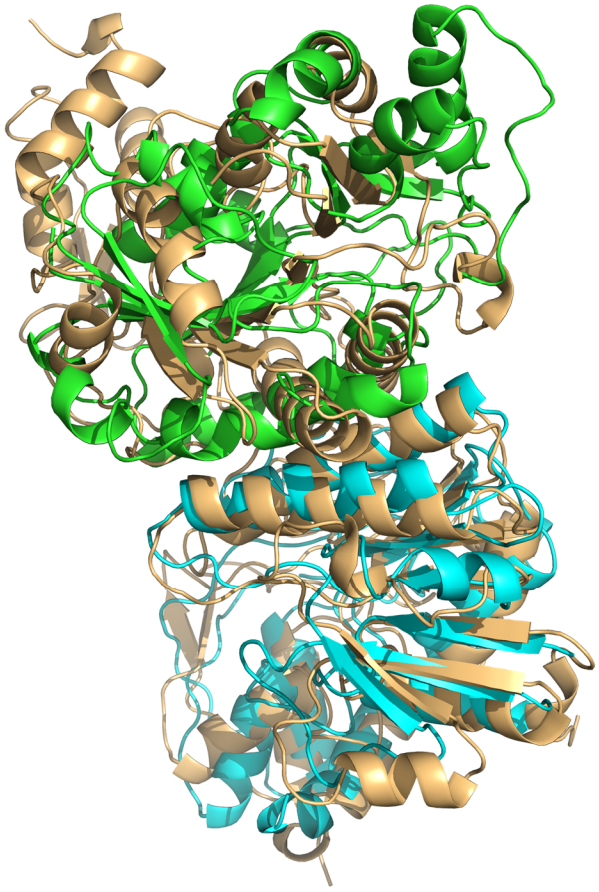


Figure 4.6: A model (green, cyan) of T137/T0965 superimposed with the native structure (wheat, PDB 6D2V). The green subunit is visibly rotated, compared to the native interface.

This case demonstrates the need for a merger between template-based and free docking methodologies. While our template-based method alone failed, it was later shown that the complex could be solved by applying free docking with restricted sampling (i.e. “focused” docking) about the modeled interface. Thus a criterion should be developed which can effectively distinguish deceptive templates and switch the protocol from template-based mode to free docking.

Ranking of the different templates is another issue of the template-based method which could be resolved by free docking. We tested an approach based on re-docking the separated subunits of the models to be evaluated. The expectation was that the more correct models would be more frequently reconstructed. This approach was inspired by the problem of discriminating between biological and crystal contacts in X-ray structures (Yueh et al., 2017). For targets with several different template models, this strategy might be applied to rank and prioritize them by the number of low energy docking solutions discovered in the neighborhood, which can be a good indicator of a low free energy state. The successful example of this approach was target T75/T0776 (PDB 4Q9A) from the previous CASP11-CAPRI challenge, for which 2 significantly different templates were available (see Figure 4.7). The number of docking poses near the correct template model was about twice the number of poses near the incorrect one.

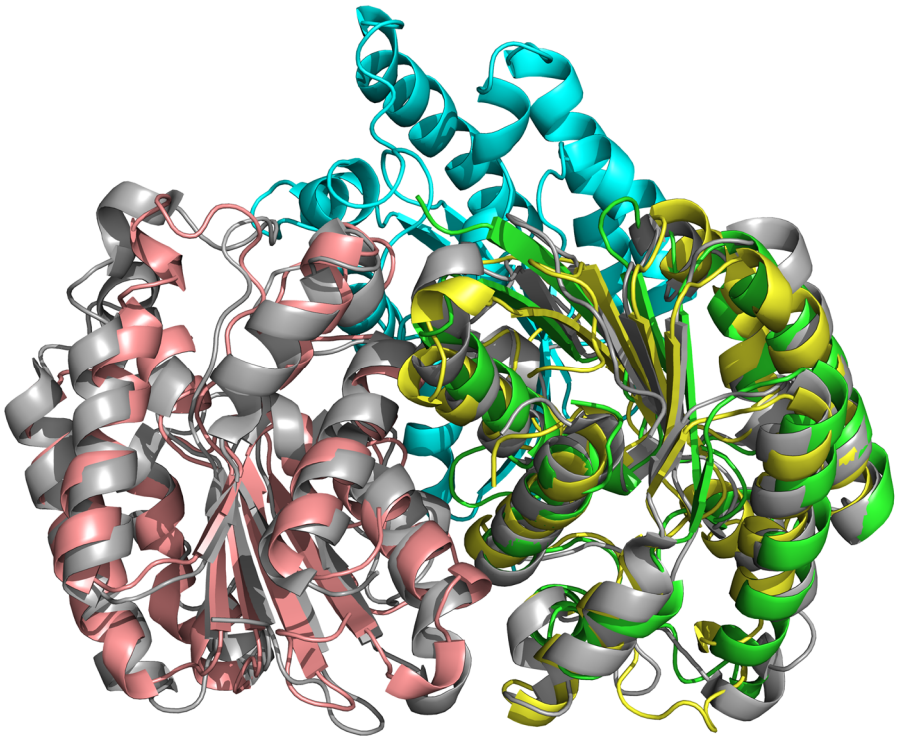


Figure 4.7: Two different T75/T0776 models (green-cyan and yellow-pink) aligned to one of the subunits of the target structure (gray, PDB 4Q9A). When the subunits are re-docked to each other, the number of near-native hits produced by the yellow-pink model was nearly twice that of the green-cyan model.

Over the years the ClusPro free docking procedure has been enhanced with a variety of tools for incorporating experimental data, including options for using SAXS and arbitrary restraints. At this point, however, these tools are not available as a part of the template-based modeling pipeline, which is a definite flaw of the current version of the server. In addition to the need for making the existing tools available through the TBM interface, the latest CASP-CAPRI round has demonstrated that ClusPro needs to be enhanced with tools for handling EM data, which is currently rapidly growing in availability. One particular example where EM was used in a human submission was target T159/H1021, representing a portion of a contractile insertion system and

possessing A6B6C6 stoichiometry. For this target, a low-resolution EM map (EMDB-2419) was available at the time the target was made open. Also, while there was no template for the assembly as a whole, partial templates were available (for instance, 1J9Q for the A6B6 portion and 1J2M for the B6C6 portion). Thus, for our submission as a human group, we individually fitted these partial templates into the EM map using a new version of the fast manifold Fourier transform (FMFT) software (Padhorny et al., 2016).

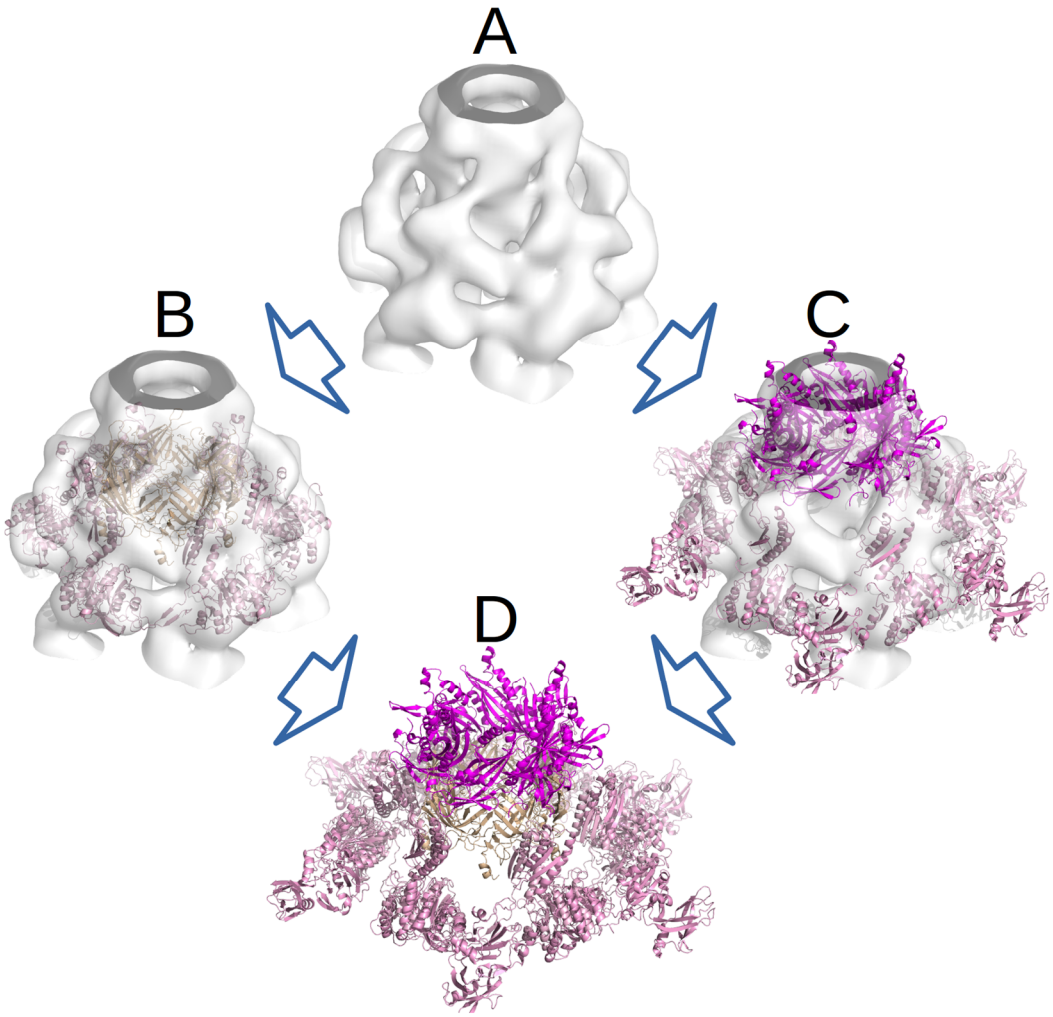


Figure 4.8: Template based modeling of target T159/H1021 assisted by low-resolution Electron Microscopy data. **A)** EM density map (EMDB-2419). **B)** Partial template for subunits A and B aligned to the EM map **C)** Partial template for subunits B and C aligned to the EM map. **D)** Mutual arrangement of the templates induced by the EM map.

During the fitting procedure, the EM density grid was correlated with the steric density grid of the template being fitted, and 350 best-scoring conformations were clustered to produce the final fitting poses. These poses were combined to build the global assembly template (see Figure 4.8), which we then used for homology modeling with MODELLER. The resulting models recapitulated the global geometry of the assembly, and had several interfaces evaluated as acceptable or medium quality. However, working with EM data is not yet implemented in ClusPro.

4.4 Conclusions

With the latest template modeling addition to ClusPro, the server is now able to submit models produced by template-based modeling or free docking. The protocol has been successful for a variety of cases; for example, we have shown that ClusPro TBM is well suited for modeling straightforward homooligomer targets (such as T152/T1003), as well as cases requiring less-conventional models based on a combination of HHpred hits. For T142/H0974, successful ClusPro models were produced by modeling the target, which had A1B1 stoichiometry, on homodimer templates. We also describe a case, T141/T0976, where another modeling mode was explored, in which the predicted structures of the A2 target are modeled on different regions of monomer templates.

This test of the ClusPro TBM module has been very promising as predictions were successful across the majority of the assessment targets. However, an important caveat is that most CAPRI targets were homomers, and hence the new module needs substantial further testing on heterodimers. Nevertheless, the targets of the current

CAPRI rounds already inspired several new avenues to server improvement. Experimental data, like the EM maps used in H1021 prior to docking, may prove useful for future complex prediction challenges, either as a modeling guidance tool or perhaps even as a scoring method for template-based models. Template discrimination is another important aspect of our modeling approach which will require future work, but whose success would improve the efficiency of the ClusPro TBM protocol. An even more challenging problem arose in target T137. Despite a strong agreement between template structures, the target complex can only be reproduced when focused docking is applied. Integration of free docking and template-based modeling into one pipeline may help to expand the number of difficult targets that can be tackled by ClusPro.

Not all targets in the latest assessment were well suited for ClusPro TBM. If no good templates were available, free docking of subunit models was used to generate predictions. Unfortunately for the few targets where this was the case, there were no acceptable or better predictions, which may be attributed to the low quality of the templates used. In fact, the side chain positions and loop conformations are usually less accurate in the homology models than in the X-ray structures of the separately crystallized constituent proteins of a complex. It appears that the methods and parameters developed for docking X-ray structures are less than optimal for docking such homology models, and there is a well-defined need for adjusting the methodology. A beta version of ClusPro TBM is available at https://tbm.cluspro.org/.

Appendix A: Supplemental Tables for PEPTIDE-PROTEIN DOCKING

Table A.1: Definition of sequence motifs for the extraction of fragments from the PDB for the PeptiDB v2 set**.** For each peptide sequence, docking motif selection is shown in a step by step fashion, following the motif building rules detailed in Section 2.2.1.

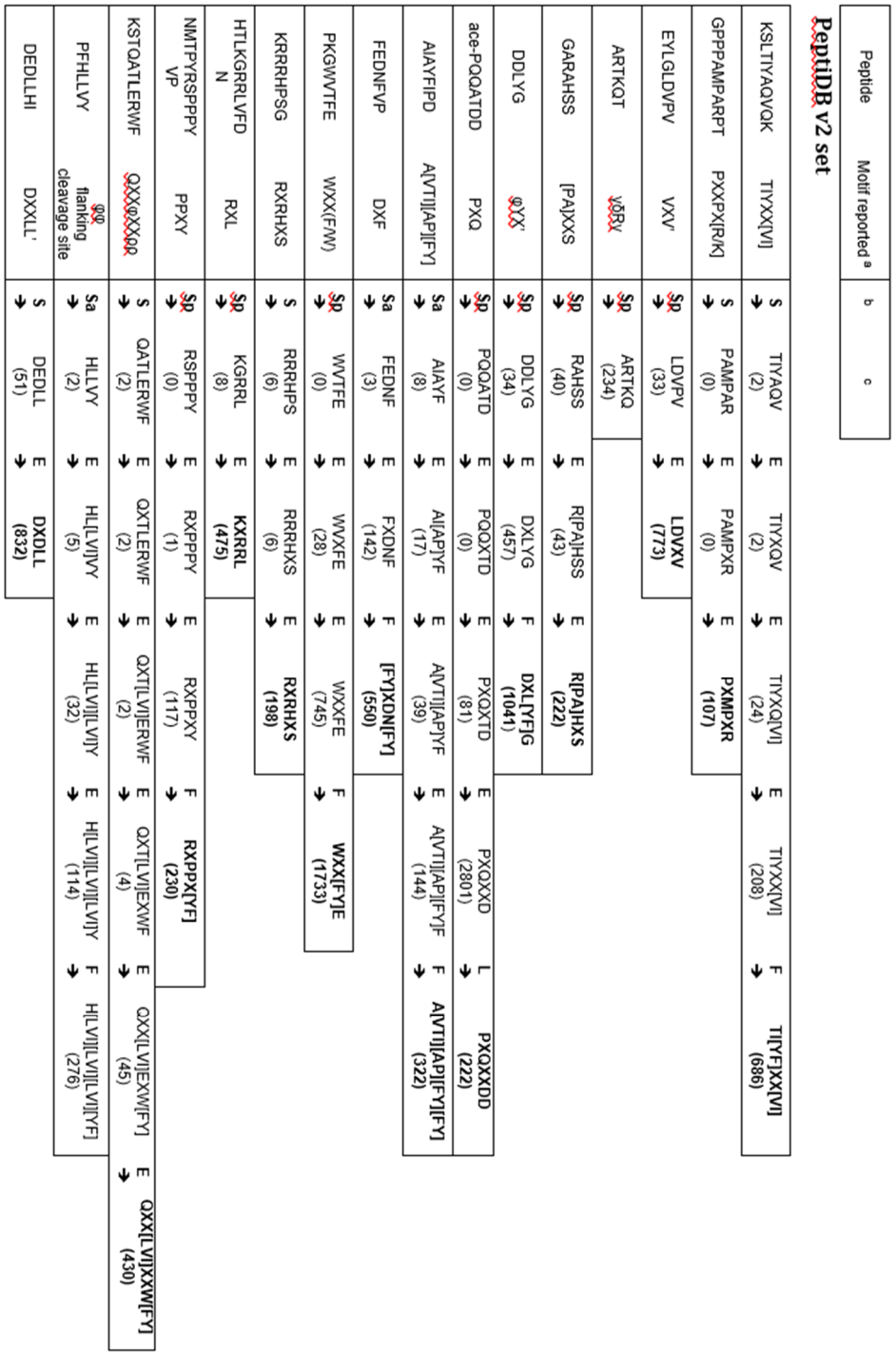


Table A.2: Definition of sequence motifs for the extraction of fragments from the PDB for the "Recent PDB" set. For each peptide sequence, docking motif selection is shown in a step by step fashion, following the motif building rules detailed in Section 2.2.1.

Peptide	Motif reported [a]	b	c
"Recent PDB" set			
SGFSF	FXFG	S → SGFSF (36)	E → **SXFSF (293)** → F → **SX[FY]S[FY] (1221)**
GDEETGE	DXETGE	S → DEETGE (6)	E → **DXETGE (173)**
SFFDKKRS	FF	Sp → FFDKK (9)	E → **FFDXK (142)** → F → **[FY][FY]DXK (597)**
NRLLLT	LLL	Sp → **NRLLL (145)**	
CTFKTKTN	KXK	Sp → KTKTN (12)	E → **KTKXN (230)**

Table A.3: Fragments extracted from the PDB using the KXRRL motif for binding of CDC6 derived peptide to cyclin. Fragments were clustered according to a 0.5Å RMSD cutoff, and ranked according to cluster size (1051 fragments were clustered into 40 clusters; the top 25 were used for docking). The source PDB of the cluster center, as well as its RMSD to the native peptide conformation in the complex are indicated in the 3rd and 4th column, respectively.

Cluster	Cluster Size	Source PDB	RMSD – C□ (Å)
1	417	2B8P	2.91
2	256	3P50	2.04
3	87	1G3I	0.44
4	63	3UKX	0.85
5	26	2CQS	1.98
6	19	1AGI	2.69
7	18	1JVB	1.89
8	16	1BL9	1.65
9	14	3VZB	2.17
10	9	1A25	1.38
11	8	2HPI	1.63
12	7	2Z11	1.85
13	6	1PML	2.72
14	6	1YEW	2.35
15	6	1WN1	0.44
16	6	3TFH	2.64
17	5	3A5Z	1.05

18	5	4M59	1.32
19	4	1JKG	2.03
20	4	2YN9	2.11
21	4	3HM0	0.85
22	4	4GQY	2.01
23	3	4LQS	2.05
24	2	3KH5	1.8
25	2	1YRP	1.68
26	2	2E61	0.6
27	2	2FEF	1.44
28	2	2H1E	2.53
29	2	3A32	0.72
30	2	3C1A	1.39
31	2	3IL0	1.76
32	2	3KTW	0.98
33	2	3N05	1.77
34	2	3QWU	1.53
35	2	3Q6S	1.06
36	2	3SL7	1.89
37	2	4B0R	0.91
38	2	4GQV	1.72
39	2	4M4W	2.89
40	1	1G4A	0.65

Table A.4: Docking results for representative fragments, based on KXRRL motif, of CDC6 derived peptide binding to cyclin. Models were clustered using a 3.5Å RMSD threshold into 100 clusters, and ranked according to cluster size. RMSD values to the native conformation are given both for the structure before and after local minimization in the 3rd and 4th column, respectively. The 3rd ranking cluster is 1.9Å RMSD away from native structure.

Cluster Center	Cluster Size	RMSD (Å)	RMSD after Minimization (Å)
1	714	8.8	9.0
2	559	4.6	4.7
3	424	2.7	1.9
4	274	5.0	4.8
5	181	8.0	8.8
6	171	10.0	9.5
7	169	10.0	10.5
8	169	35.7	36.1
9	145	4.4	3.3
10	145	4.9	6.3
11	122	7.7	8.4
12	117	4.3	5.6
13	112	5.4	5.0
14	110	4.5	5.3
15	105	35.0	34.5
16	104	5.0	4.4
17	101	8.0	9.3
18	101	7.8	7.8
19	94	5.3	4.8
20	86	4.6	5.0
21	83	9.7	9.8
22	83	34.1	34.5
23	75	9.8	9.8
24	72	10.8	10.0
25	68	7.0	4.5
26	68	5.1	6.1
27	61	4.2	4.2
28	60	7.4	8.4
29	59	5.2	5.4
30	58	5.1	4.6
31	54	9.3	9.5
32	49	33.9	35.3

33	48	36.3	36.3
34	47	10.2	9.8
35	46	7.7	6.7
36	43	37.6	37.8
37	43	8.4	8.3
38	42	10.2	9.3
39	37	9.2	9.4
40	35	3.6	2.8
41	35	32.1	32.7
42	35	31.2	31.8
43	32	5.8	5.9
44	32	9.8	10.7
45	30	9.1	8.7
46	29	35.9	36.3
47	29	12.5	12.3
48	28	9.2	9.2
49	28	34.4	35.8
50	27	34.7	35.4
51	26	35.4	36.0
52	25	7.5	7.8
53	25	5.5	6.4
54	25	31.8	33.9
55	24	6.3	6.6
56	23	5.4	4.1
57	23	8.2	8.4
58	22	8.5	8.8
59	22	11.7	11.3
60	21	7.3	7.3
61	21	25.4	26.2
62	21	38.8	37.1
63	20	7.9	9.0
64	20	36.0	35.4
65	20	38.8	37.2
66	19	5.5	5.7
67	19	8.0	8.6
68	18	33.5	35.4
69	18	9.1	8.7

70	17	7.8	8.0
71	17	35.7	36.6
72	16	7.4	6.2
73	15	8.9	8.7
74	14	6.9	6.3
75	14	8.0	7.9
76	13	9.1	10.5
77	13	35.7	36.6
78	13	5.7	8.0
79	12	4.7	4.3
80	10	7.9	8.1
81	10	36.3	35.3
82	9	5.2	3.8
83	9	7.8	7.4
84	9	33.2	35.2
85	9	8.5	8.2
86	9	8.0	8.0
87	8	8.9	8.4
88	7	35.9	36.3
89	7	13.0	12.8
90	6	5.9	5.6
91	5	35.1	35.7
92	4	10.0	9.5
93	4	33.3	34.5
94	4	5.1	5.8
95	4	33.0	34.1
96	4	9.9	10.6
97	4	6.2	6.3
98	3	9.4	11.0
99	3	8.9	9.6
100	1	5.2	5.9

Appendix B: Supplemental Figures/Tables for ANTIBODY MATURATION

Figure B.1: Sequences for inferred CH65-CH67 lineage (Schmidt et al. 2013). Beginning with the highest affinity mature antibody (CH67), the necessary mutations to model the lower affinity antibodies are shown in each line. No structure was crystallized for CH66.

[1] Parameter sets are distinguished by rotation ('deg') and weight ('coeff') set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

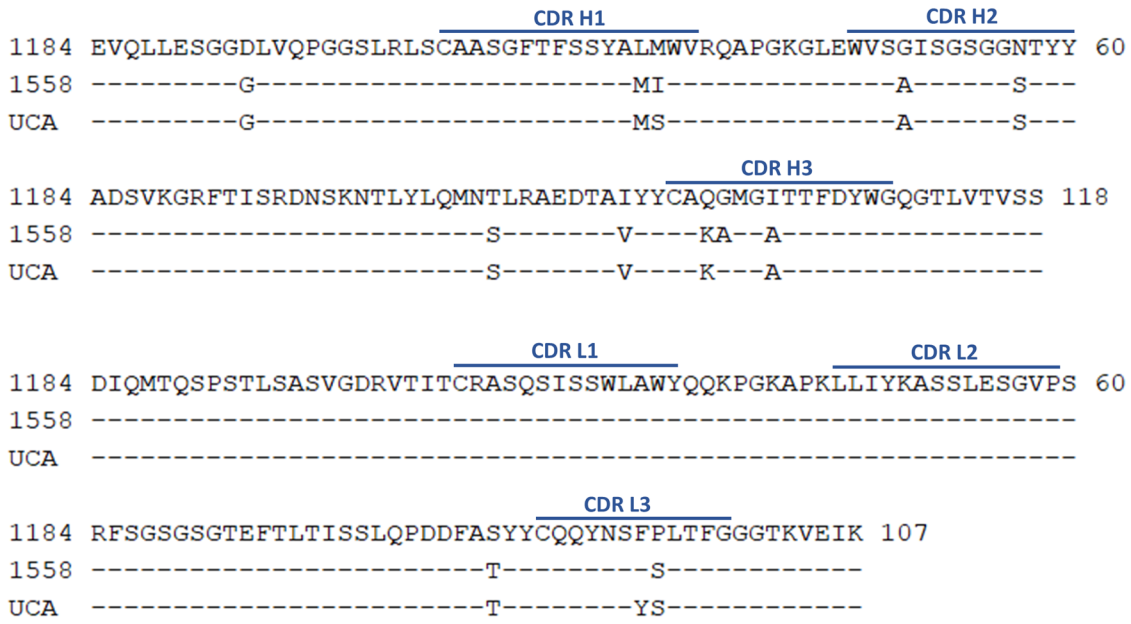


Figure B.2: Sequences for anthrax PA antibodies (Ataca 2018). Beginning with the highest affinity mature antibody (1184), the necessary mutations to model the lower affinity antibodies are shown in each line.

Table B.1: Near native counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking with HA antigen assuming antibody bound conformation.

Parameter Set[1]	CH67	CH65	I-2	UCA
20deg_coeff71	433	411	393	393
20deg_coeff107	433	408	388	390
30deg_coeff5	719	701	677	661
30deg_coeff50	746	749	683	670
30deg_coeff53	748	754	683	663
30deg_coeff86	741	736	689	668
30deg_coeff98	776	769	693	681
40deg_coeff51	793	825	661	522
40deg_coeff96	752	750	649	534
40deg_coeff98	888	955	863	793

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function
[2] Residues in red text identified as being mutated through affinity maturation

Table B.2: Weighted near native average counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking unbound antibody ensembles.

Parameter Set[1]	CH67	CH65	I-2	UCA
20deg_coeff71	393.21	310.27	2.25	5.23
20deg_coeff107	392.42	306.72	2.25	6.01
30deg_coeff5	708.40	542.59	0.06	5.76
30deg_coeff50	712.15	595.67	0.31	5.18
30deg_coeff53	722.34	588.01	0.27	5.52
30deg_coeff86	707.94	587.76	0.48	5.23
30deg_coeff98	724.59	620.72	0.73	5.80
40deg_coeff51	752.67	583.51	11.98	19.67
40deg_coeff96	735.31	612.62	33.54	23.70
40deg_coeff98	820.63	722.95	2.79	6.01

Table B.3: Lowest PIPER energies associated with near native counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking with HA antigen assuming antibody bound conformation.

Parameter Set[1]	CH67	CH65	I-2	UCA
20deg_coeff71	-767.92	-721.25	-821.89	-810.26
20deg_coeff107	-735.41	-686.41	-787.19	-778.76
30deg_coeff5	-488.05	-458.85	-581.17	-591.03
30deg_coeff50	-592.77	-573.80	-678.37	-680.22
30deg_coeff53	-606.19	-586.75	-691.67	-694.73
30deg_coeff86	-566.64	-552.60	-649.95	-647.48
30deg_coeff98	-713.87	-701.70	-799.77	-790.96
40deg_coeff51	-467.22	-484.58	-466.35	-453.07
40deg_coeff96	-579.92	-605.39	-580.55	-563.83
40deg_coeff98	-713.87	-701.70	-820.18	-803.38

Table B.4: Lowest PIPER energies associated with weighted near native average counts (within 10 Å to native interface) for CH65-CH67 lineage antibodies after docking unbound antibody ensembles

Parameter Set[1]	CH67	CH65	I-2	UCA
20deg_coeff71	-725.27	-626.01	-34.66	-152.41
20deg_coeff107	-686.19	-592.81	-32.94	-159.08
30deg_coeff5	-536.68	-428.95	-4.98	-13.87
30deg_coeff50	-627.91	-525.00	-28.01	-28.31
30deg_coeff53	-639.90	-534.90	-28.01	-61.32
30deg_coeff86	-592.36	-503.85	-26.46	-26.99
30deg_coeff98	-735.51	-635.09	-34.90	-73.45
40deg_coeff51	-448.45	-429.68	-143.55	-203.22
40deg_coeff96	-549.42	-538.38	-312.21	-269.46
40deg_coeff98	-772.86	-671.81	-85.19	-38.34

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.5: Predicted - ΔG values calculated from linear fits between experimental - ΔG values estimated from SPR data to N and PIPER energies from docking CH65-CH67 lineage antibodies with the HA antigen, assuming antibody bound conformation.

Predicted - ΔG (kcal/mol)				
Parameter Set[1]	**CH67**	**CH65**	**I-2**	**UCA**
20deg_coeff71	8.79	8.59	5.17	5.44
20deg_coeff107	8.79	8.60	5.15	5.46
30deg_coeff5	8.58	8.73	5.65	5.03
30deg_coeff50	8.48	8.85	5.52	5.14
30deg_coeff53	8.37	8.94	5.49	5.20
30deg_coeff86	8.50	8.85	5.43	5.22
30deg_coeff98	8.65	8.73	5.36	5.25
40deg_coeff51	8.51	8.15	6.62	4.72
40deg_coeff96	8.37	8.16	6.74	4.72
40deg_coeff98	8.45	8.88	5.10	5.57

Table B.6: Predicted - ΔG values calculated from linear fits between experimental - ΔG values estimated from SPR data to weighted N and PIPER energies from docking CH65-CH67 lineage unbound antibody ensembles.

Predicted - ΔG (kcal/mol)				
Parameter Set[1]	**CH67**	**CH65**	**I-2**	**UCA**
20deg_coeff71	8.72	8.62	5.02	5.62
20deg_coeff107	8.71	8.61	4.98	5.69
30deg_coeff5	9.01	8.32	5.20	5.46
30deg_coeff50	8.95	8.40	5.20	5.44
30deg_coeff53	8.98	8.36	5.19	5.46
30deg_coeff86	8.93	8.43	5.20	5.42
30deg_coeff98	8.94	8.42	5.19	5.44
40deg_coeff51	8.80	8.59	5.25	5.36
40deg_coeff96	8.84	8.54	5.26	5.34
40deg_coeff98	8.83	8.55	5.22	5.39

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.7: Residual errors calculated between experimental - ΔG values and those predicted from linear fits of docking results from CH65-CH67 lineage antibodies (restricted backbone).

Residual Errors from Estimated - ΔG (kcal/mol)					
Parameter Set[1]	**CH67**	**CH65**	**I-2**	**UCA**	**Average**
20deg_coeff71	5.87E-03	1.27E-02	7.59E-02	8.28E-02	4.43E-02
20deg_coeff107	1.17E-03	8.12E-03	9.50E-02	1.04E-01	5.21E-02
30deg_coeff5	2.04E-01	1.29E-01	4.00E-01	3.25E-01	2.65E-01
30deg_coeff50	3.03E-01	2.44E-01	2.78E-01	2.19E-01	2.61E-01
30deg_coeff53	4.15E-01	3.33E-01	2.41E-01	1.60E-01	2.87E-01
30deg_coeff86	2.83E-01	2.46E-01	1.78E-01	1.41E-01	2.12E-01
30deg_coeff98	1.33E-01	1.26E-01	1.15E-01	1.09E-01	1.21E-01
40deg_coeff51	2.79E-01	4.54E-01	1.37E+00	6.41E-01	6.87E-01
40deg_coeff96	4.15E-01	4.42E-01	1.50E+00	6.41E-01	7.49E-01
40deg_coeff98	3.38E-01	2.73E-01	1.48E-01	2.13E-01	2.43E-01

Table B.8: Residual errors calculated between experimental - ΔG values and those predicted from linear fits of docking results from CH65-CH67 lineage unbound antibody ensembles.

Predicted - ΔG (kcal/mol)					
Parameter Set[1]	**CH67**	**CH65**	**I-2**	**UCA**	**Average**
20deg_coeff71	6.22E-02	1.97E-02	2.23E-01	2.66E-01	1.43E-01
20deg_coeff107	7.26E-02	1.03E-02	2.68E-01	3.30E-01	1.70E-01
30deg_coeff5	2.28E-01	2.88E-01	4.70E-02	1.07E-01	1.68E-01
30deg_coeff50	1.65E-01	1.99E-01	4.73E-02	8.15E-02	1.23E-01
30deg_coeff53	1.95E-01	2.43E-01	5.98E-02	1.07E-01	1.51E-01
30deg_coeff86	1.43E-01	1.70E-01	4.16E-02	6.83E-02	1.06E-01
30deg_coeff98	1.50E-01	1.80E-01	5.79E-02	8.86E-02	1.19E-01
40deg_coeff51	1.73E-02	1.85E-02	1.15E-03	1.33E-05	9.25E-03
40deg_coeff96	5.70E-02	6.20E-02	1.64E-02	1.15E-02	3.67E-02
40deg_coeff98	4.50E-02	5.03E-02	2.56E-02	3.08E-02	3.79E-02

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.9: Experimental K_D values for the CH65-67 lineage antibodies, estimated from SPR measurements (Schmidt et al. 2013). **A)** Ranges are shown for K_D values using reported experimental errors associated with each measurement. **B)** Errors are propagated to - ΔG values. **C)** "Residual" error is calculated between the minimum and maximum - ΔG value and the value of - ΔG estimated from the reported K_D value.

	CH67	**CH65**	**I-2**	**UCA**
A. Experimental K_D (M)				
reported	3.60E-07	4.90E-07	1.42E-04	1.18E-04
−	3.20E-07	3.90E-07	1.27E-04	1.04E-04
+	4.00E-07	5.90E-07	1.57E-04	1.32E-04
B. Derived - ΔG (kcal/mol)				
estimated	8.79	8.60	5.25	5.36
−	8.86	8.74	5.31	5.43
+	8.72	8.49	5.19	5.29
C. Error propagated through calculation (kcal/mol)				
−	6.98E-02	1.35E-01	6.61E-02	7.48E-02
+	6.24E-02	1.10E-01	5.95E-02	6.64E-02

Table B.10: Near native counts (within 10 Å to native interface) for anthrax PA antibodies after docking with PA antigen assuming antibody bound conformation.

Parameter Set[1]	**Fab1184**	**Fab1558**	**UCA**
20deg_coeff71	494	472	496
20deg_coeff107	500	484	500
30deg_coeff5	932	937	962
30deg_coeff50	884	853	881
30deg_coeff53	908	888	895
30deg_coeff86	893	879	895
30deg_coeff98	843	793	811
40deg_coeff51	763	853	842
40deg_coeff96	625	693	683
40deg_coeff98	808	774	785

[1] Parameter sets are distinguished by rotation ('deg') and weight ('coeff') set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.11: Weighted near native average counts (within 10 Å to native interface) for anthrax PA antibodies after docking unbound antibody ensembles.

Parameter Set[1]	Fab1184	Fab1558	UCA
20deg_coeff71	408.81	373.46	423.06
20deg_coeff107	415.17	364.38	433.52
30deg_coeff5	754.99	608.86	756.85
30deg_coeff50	725.00	581.41	669.75
30deg_coeff53	745.15	600.76	688.96
30deg_coeff86	733.31	576.16	681.05
30deg_coeff98	693.58	558.46	620.86
40deg_coeff51	602.32	431.64	506.58
40deg_coeff96	534.77	359.55	471.05
40deg_coeff98	614.10	492.93	553.89

Table B.12: Lowest PIPER energies associated with near native counts (within 10 Å to native interface) for anthrax PA antibodies after docking with PA antigen assuming antibody bound conformation.

Parameter Set[1]	Fab1184	Fab1558	UCA
20deg_coeff71	-922.37	-991.82	-1084.18
20deg_coeff107	-890.23	-944.83	-1035.08
30deg_coeff5	-548.34	-587.85	-673.31
30deg_coeff50	-692.44	-714.06	-796.83
30deg_coeff53	-705.17	-736.95	-825.51
30deg_coeff86	-669.84	-689.16	-757.33
30deg_coeff98	-852.04	-863.16	-949.03
40deg_coeff51	-571.10	-596.90	-620.54
40deg_coeff96	-713.06	-720.40	-744.56
40deg_coeff98	-852.04	-863.16	-949.03

Table B.13: Lowest PIPER energies associated with weighted near native average counts (within 10 Å to native interface) for anthrax PA antibodies after docking unbound antibody ensembles.

Parameter Set[1]	Fab1184	Fab1558	UCA
20deg_coeff71	-804.22	-889.77	-966.78
20deg_coeff107	-771.49	-828.58	-904.33
30deg_coeff5	-503.74	-552.42	-634.06
30deg_coeff50	-601.40	-646.73	-721.05
30deg_coeff53	-618.16	-669.19	-744.42
30deg_coeff86	-573.49	-610.17	-674.59
30deg_coeff98	-726.21	-768.03	-837.96
40deg_coeff51	-489.56	-491.37	-519.14
40deg_coeff96	-605.42	-609.60	-625.26
40deg_coeff98	-726.83	-768.56	-839.60

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.14: CHARMM Energy calculations for non-bonded interactions of antibody interface (within 10 Å of HA) residues of FabCH67-HA complex.

CH:RES_NUM[2]	**delta-ENER** (kcal/mol)	**delta-VDW** (kcal/mol)	**delta-ELEC** (kcal/mol)
H:TYR_109	7.705	6.899	0.80616
H:VAL_106	6.481	5.864	0.616
L:ASP_93	6.385	2.882	3.504
L:ARG_29	5.662	4.174	1.489
H:ASP_107	5.364	4.991	0.373
L:TRP_90	3.111	2.901	0.210
L:ASP_95	2.710	2.204	0.506
H:HIS_52	2.660	2.463	0.197
L:SER_92	2.490	2.298	0.192
H:ASN_54	2.466	2.359	0.107
H:ASP_31	2.439	1.237	1.202
H:TYR_33	1.795	1.670	0.126
L:ASN_26	1.742	1.526	0.217
L:ASP_91	1.740	0.099	1.640
H:ARG_104	1.433	0.481	0.952
H:GLU_102	1.370	0.282	1.088
H:TYR_111	1.105	0.997	0.108
H:GLU_65	0.970	0.012	0.957
H:TYR_108	0.670	0.086	0.584
H:TRP_50	0.563	0.753	-0.190
H:ARG_72	0.548	0.038	0.509
H:SER_105	0.539	1.024	-0.486
H:PHE_110	0.504	0.097	0.407
L:LYS_30	0.415	0.105	0.310
H:SER_55	0.400	0.370	0.030
L:ARG_31	0.376	0.002	0.374
H:PRO_103	0.284	0.210	0.073
H:ASN_32	0.268	0.190	0.077
L:SER_94	0.166	0.034	0.133
H:GLY_56	0.155	0.006	0.149
H:ALA_57	0.142	0.005	0.137
L:GLY_67	0.113	0.000	0.113
H:GLN_62	0.105	0.002	0.102
L:THR_68	0.088	0.003	0.085
H:THR_58	0.069	0.004	0.065
L:GLY_28	0.062	0.001	0.061
H:HIS_35	0.058	0.006	0.052

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

H:GLY_100	0.046	0.000	0.046
H:THR_30	0.043	0.140	-0.097
H:TRP_47	0.041	0.021	0.020
H:LEU_101	0.026	0.000	0.026
H:PHE_29	0.023	0.004	0.019
H:ILE_34	0.022	0.002	0.020
H:TYR_27	0.022	0.050	-0.029
H:ALA_99	0.021	0.000	0.021
L:ILE_27	0.009	0.002	0.008
L:SER_66	0.007	0.000	0.007
H:THR_74	0.001	0.003	-0.001
L:VAL_32	-0.000	0.000	-0.000
H:ILE_51	-0.003	0.003	-0.005
H:GLY_49	-0.008	0.000	-0.008
H:GLY_112	-0.026	0.000	-0.027
H:THR_28	-0.028	0.042	-0.070
H:ALA_61	-0.123	0.000	-0.123
L:VAL_97	-0.174	0.006	-0.180
H:TYR_60	-0.239	0.004	-0.243
L:ASN_25	-0.300	0.052	-0.353
H:LYS_59	-0.431	1.373	-1.804
L:HIS_96	-0.453	0.007	-0.460

Table B.15: CHARMM Energy calculations for non-bonded interactions in the interface (within 10 Å of HA) of the FabCH65-HA complex.

CH:RES_NUM[2]	**delta-ENER** (kcal/mol)	**delta-VDW** (kcal/mol)	**delta-ELEC** (kcal/mol)
H:TYR_109	8.345	7.776	0.568
H:ASP_107	7.761	4.681	3.079
L:ARG_29	7.304	7.870	-0.567
H:VAL_106	4.995	4.952	0.042
L:ASP_95	3.897	2.101	1.796
L:TRP_90	3.214	3.061	0.153
L:ASP_26	2.896	0.979	1.917
H:HIS_52	2.546	2.326	0.220
H:ASP_31	2.411	1.342	1.069
H:ASN_54	2.407	2.171	0.236
H:GLU_102	1.937	0.304	1.632
H:TYR_111	1.711	1.706	0.005
L:SER_92	1.707	1.441	0.266
H:ASP_57	1.678	1.734	-0.057

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

H:TRP_50	1.010	1.114	-0.103
L:ASP_91	0.983	0.160	0.823
H:HIS_33	0.914	0.739	0.176
H:SER_105	0.886	1.172	-0.286
H:ARG_104	0.808	0.123	0.686
H:SER_55	0.720	0.685	0.035
H:TYR_32	0.696	0.557	0.139
L:SER_93	0.631	0.292	0.339
H:ASN_59	0.440	0.364	0.076
H:TYR_110	0.429	0.108	0.322
H:ARG_72	0.403	0.041	0.362
H:PRO_103	0.340	0.330	0.010
H:ARG_98	0.198	0.007	0.192
H:GLY_56	0.198	0.010	0.188
L:SER_94	0.186	0.024	0.162
H:TYR_108	0.158	0.064	0.094
H:THR_28	0.155	0.035	0.121
H:PRO_53	0.109	0.069	0.040
H:THR_58	0.095	0.010	0.085
H:PHE_29	0.066	0.003	0.064
H:GLY_100	0.043	0.000	0.043
H:TYR_27	0.023	0.024	-0.002
H:TRP_47	0.018	0.011	0.007
H:THR_30	0.013	0.091	-0.078
H:GLY_99	0.002	0.000	0.002
H:ILE_51	-0.005	0.007	-0.012
H:GLN_65	-0.008	0.017	-0.025
L:GLY_67	-0.040	0.000	-0.041
H:ALA_61	-0.045	0.000	-0.045
L:ILE_27	-0.090	0.004	-0.094
L:LYS_30	-0.120	0.164	-0.284
L:SER_31	-0.141	0.001	-0.142
L:GLY_28	-0.144	0.002	-0.145
L:ASN_68	-0.170	0.007	-0.177
L:ASP_52	-0.212	0.002	-0.214
H:TYR_60	-0.250	0.003	-0.255
L:ASN_25	-0.254	0.002	-0.256
L:HIS_96	-0.490	0.034	-0.524

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

Table B.16: CHARMM Energy calculations for non-bonded interactions in the interface (within 10 Å of PA) of the Fab1184-PA complex.

CH:RES_NUM[2]	**delta-ENER** (kcal/mol)	**delta-VDW** (kcal/mol)	**delta-ELEC** (kcal/mol)
L:LYS_50	21.427	7.328	14.099
L:PHE_94	5.732	5.319	0.413
H:TYR_32	5.593	5.326	0.267
L:TRP_32	4.841	4.388	0.453
L:TYR_49	4.085	4.189	-0.104
H:MET_100	3.960	3.657	0.304
H:TYR_59	3.600	3.520	0.081
H:ILE_102	2.802	1.719	1.083
H:ASP_106	2.686	1.305	1.381
L:TYR_91	2.604	1.478	1.126
H:ASN_57	2.251	2.010	0.241
H:SER_31	2.043	1.597	0.445
H:THR_104	1.546	1.513	0.033
L:GLU_55	1.484	1.014	0.471
H:ARG_72	1.391	0.023	1.368
H:SER_54	1.254	1.113	0.141
H:THR_103	1.185	2.143	-0.957
H:THR_28	1.114	1.039	0.075
L:ASN_92	0.819	0.310	0.509
L:LEU_96	0.773	0.589	0.185
L:SER_56	0.707	0.429	0.278
H:GLY_101	0.674	0.034	0.640
H:PHE_27	0.506	0.042	0.464
L:LEU_33	0.478	0.003	0.475
H:PHE_29	0.426	0.002	0.424
H:SER_30	0.418	0.027	0.391
H:THR_58	0.401	0.024	0.377
L:ALA_51	0.399	0.002	0.396
H:LEU_34	0.389	0.001	0.388
H:ALA_33	0.376	0.010	0.365
H:GLY_53	0.371	0.028	0.343
H:ILE_51	0.351	0.020	0.331
L:GLN_89	0.324	0.077	0.247
H:SER_52	0.319	0.172	0.148
H:GLY_56	0.271	0.020	0.251
H:LYS_65	0.256	0.007	0.250
L:LEU_46	0.221	0.641	-0.420

[1] Parameter sets are distinguished by rotation ('deg') and weight ('coeff') set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

L:ALA_34	0.197	0.000	0.197
H:GLY_26	0.170	0.000	0.170
L:ILE_2	0.148	0.037	0.110
L:SER_93	0.143	0.146	-0.003
H:GLY_50	0.141	0.000	0.140
H:GLY_55	0.140	0.015	0.126
H:TRP_47	0.132	0.131	0.001
L:PRO_95	0.090	0.034	0.056
H:MET_35	0.083	0.065	0.017
L:ILE_29	0.081	0.016	0.065
H:VAL_2	0.058	0.006	0.052
H:TYR_60	0.057	0.008	0.050
L:SER_60	0.055	0.033	0.022
L:ARG_61	0.055	0.000	0.055
L:PRO_59	0.050	0.001	0.049
L:PHE_62	0.034	0.000	0.034
L:THR_97	0.016	0.005	0.011
L:TYR_36	0.015	0.012	0.003
L:SER_63	0.014	0.000	0.014
L:SER_31	0.003	0.005	-0.002
L:GLY_57	-0.001	0.005	-0.006
H:PHE_105	-0.002	0.041	-0.043
L:TRP_35	-0.024	0.001	-0.025
L:LEU_54	-0.041	0.166	-0.207
L:VAL_58	-0.072	0.009	-0.081
H:GLU_1	-0.095	0.003	-0.098
L:ASP_1	-0.098	0.002	-0.100
H:ASN_74	-0.102	0.008	-0.109
H:GLY_99	-0.110	0.001	-0.112
L:GLN_90	-0.141	0.122	-0.263
L:ILE_48	-0.164	0.017	-0.181
H:TRP_108	-0.203	0.002	-0.205
H:TYR_107	-0.262	0.098	-0.360
L:SER_30	-0.280	0.008	-0.288
L:LEU_47	-0.288	0.001	-0.289
L:LYS_45	-0.304	0.001	-0.304
H:GLN_98	-0.339	0.071	-0.410
L:SER_52	-0.367	0.042	-0.410
L:SER_53	-0.615	0.064	-0.679

[1] Parameter sets are distinguished by rotation (‘deg’) and weight (‘coeff’) set for the PIPER energy function

[2] Residues in red text identified as being mutated through affinity maturation

BIBLIOGRAPHY

Alberts, B. J., A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. (2002). B Cells and Antibodies. In *Molecular Biology of the Cell* (4th ed.). New York: Garland Science.

Ataca, S. (2018). Biophysical characterization of affinity maturation in the human response to anthrax vaccine (Doctoral dissertation). Retrieved from OpenBU: https://open.bu.edu/handle/2144/32958

Awwad, S., & Angkawinitwong, U. (2018). Overview of Antibody Drug Delivery. *Pharmaceutics, 10*(3), e83.

Ben-Shimon, A., & Niv, M. Y. (2015). AnchorDock: Blind and Flexible Anchor-Driven Peptide Docking. *Structure, 23*(5), 929-940.

Blaszczyk, M., Kurcinski, M., Kouza, M., Wieteska, L., Debinski, A., Kolinski, A., & Kmiecik, S. (2016). Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods, 93*, 72-83.

Bowers, K. J. C., E.; Xu, H.; Drof, R. O.; Eastwood, M. P.; Gregerson, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. (2006). *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*. Paper presented at the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida.

Brenke, R., Hall, D. R., Chuang, G. Y., Comeau, S. R., Bohnuud, T., Beglov, D., . . . Kozakov, D. (2012). Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics, 28*(20), 2608-2614.

Brett, T. J., Traub, L. M., & Fremont, D. H. (2002). Accessory protein recruitment motifs in clathrin-mediated endocytosis. *Structure, 10*(6), 797-809.

Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., . . . Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry, 30*(10), 1545-1614.

Bullock, A. N., Debreczeni, J., Amos, A. L., Knapp, S., & Turk, B. E. (2005). Structure and substrate specificity of the Pim-1 kinase. *Journal of Biological Chemistry, 280*(50), 41675-41682.

Carroni, M., & Saibil, H. R. (2016). Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods, 95*, 78-85.

Chen, H. I., & Sudol, M. (1995). The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. *Proceedings of the National Academy of Sciences of the United States of America, 92*(17), 7819-7823.

Chen, H. J., Yuan, J., & Lobel, P. (1997). Systematic mutational analysis of the cation-independent mannose 6-phosphate/insulin-like growth factor II receptor cytoplasmic domain. An acidic cluster containing a key aspartate is important for function in lysosomal enzyme sorting. *Journal of Biological Chemistry, 272*(11), 7003-7012.

Chen, R., Li, L., & Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins, 52*(1), 80-87.

Cheng, K. Y., Noble, M. E., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., . . . Johnson, L. N. (2006). The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition. *Journal of Biological Chemistry, 281*(32), 23167-23179.

Chuang, G. Y., Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2008). DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophysical Journal, 95*(9), 4217-4227.

Clark, L. A., Ganesan, S., Papp, S., & van Vlijmen, H. W. (2006). Trends in antibody sequence changes during the somatic hypermutation process. *Journal of Immunology, 177*(1), 333-340.

Clarkson, W. D., Kent, H. M., & Stewart, M. (1996). Separate binding sites on nuclear transport factor 2 (NTF2) for GDP-Ran and the phenylalanine-rich repeat regions of nucleoporins p62 and Nsp1p. *Journal of Molecular Biology, 263*(4), 517-524.

Comeau, S. R., Kozakov, D., Brenke, R., Shen, Y., Beglov, D., & Vajda, S. (2007). ClusPro: performance in CAPRI rounds 6-11 and the new server. *Proteins, 69*(4), 781-785.

Dagliyan, O., Proctor, E. A., D'Auria, K. M., Ding, F., & Dokholyan, N. V. (2011). Structural and dynamic determinants of protein-peptide recognition. *Structure, 19*(12), 1837-1845.

Devergne, O., Hatzivassiliou, E., Izumi, K. M., Kaye, K. M., Kleijnen, M. F., Kieff, E., & Mosialos, G. (1996). Association of TRAF1, TRAF2, and TRAF3 with an Epstein-Barr virus LMP1 domain important for B-lymphocyte transformation: role in NF-kappaB activation. *Molecular and Cellular Biology, 16*(12), 7098-7108.

Dunbrack, R. L., Jr., & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *Journal of Molecular Biology, 230*(2), 543-574.

Erak, M., Bellmann-Sickert, K., Els-Heindl, S., & Beck-Sickinger, A. G. (2018). Peptide chemistry toolbox - Transforming natural peptides into peptide therapeutics. *Bioorganic and Medicinal Chemistry, 26*(10), 2759-2765.

Greer, J., & Bush, B. L. (1978). Macromolecular shape and surface maps by solvent exclusion. *Proceedings of the National Academy of Sciences of the United States of America, 75*(1), 303-307.

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins, 47*(4), 409-443.

Haynes, B. F., Kelsoe, G., Harrison, S. C., & Kepler, T. B. (2012). B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature Biotechnology, 30*(5), 423-433.

Hou, T., Li, N., Li, Y., & Wang, W. (2012). Characterization of domain-peptide interaction interface: prediction of SH3 domain-mediated protein-protein interaction network in yeast by generic structure-based models. *Journal of Proteome Research, 11*(5), 2982-2995.

Ignatov, M., Kazennov, A., & Kozakov, D. (2018). ClusPro FMFT-SAXS: Ultra-fast Filtering Using Small-Angle X-ray Scattering Data in Protein Docking. *Journal of Molecular Biology, 430*(15), 2249-2255.

Jackson, M. R., Nilsson, T., & Peterson, P. A. (1990). Identification of a consensus motif for retention of transmembrane proteins in the endoplasmic reticulum. *EMBO Journal, 9*(10), 3153-3162.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., . . . Critical Assessment of, P. I. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins, 52*(1), 2-9.

Jaulin-Bastard, F., Saito, H., Le Bivic, A., Ollendorff, V., Marchetto, S., Birnbaum, D., & Borg, J. P. (2001). The ERBB2/HER2 receptor differentially interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 domain proteins. *Journal of Biological Chemistry, 276*(18), 15256-15263.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between

proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America, 89*(6), 2195-2199.

Kepler, T. B., & Wiehe, K. (2017). Genetic and structural analyses of affinity maturation in the humoral response to HIV-1. *Immunologcial Reviews, 275*(1), 129-144.

Kobayashi, M., Itoh, K., Suzuki, T., Osanai, H., Nishikawa, K., Katoh, Y., . . . Yamamoto, M. (2002). Identification of the interactive interface and phylogenic conservation of the Nrf2-Keap1 system. *Genes to Cells, 7*(8), 807-820.

Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R., & Vajda, S. (2013). How good is automated protein docking? *Proteins, 81*(12), 2159-2166.

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins, 65*(2), 392-406.

Kozakov, D., Clodfelter, K. H., Vajda, S., & Camacho, C. J. (2005). Optimal clustering for detecting near-native conformations in protein docking. *Biophysical Journal, 89*(2), 867-875.

Kozakov, D., Hall, D. R., Beglov, D., Brenke, R., Comeau, S. R., Shen, Y., . . . Vajda, S. (2010). Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins, 78*(15), 3124-3130.

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., . . . Vajda, S. (2017). The ClusPro web server for protein-protein docking. *Nature Protocols, 12*(2), 255-278.

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins, 77*(4), 778-795.

Kundrotas, P. J., Zhu, Z., Janin, J., & Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences of the United States of America, 109*(24), 9438-9441.

Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., & Kmiecik, S. (2015). CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research, 43*(W1), W419-424.

Lavi, A., Ngan, C. H., Movshovitz-Attias, D., Bohnuud, T., Yueh, C., Beglov, D., . . . Kozakov, D. (2013). Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. *Proteins, 81*(12), 2096-2105.

Lee, H., Heo, L., Lee, M. S., & Seok, C. (2015). GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Research, 43*(W1), W431-435.

Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., & Wodak, S. J. (2018). The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins, 86 Suppl 1*, 257-273.

Lensink, M. F., Velankar, S., & Wodak, S. J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins, 85*(3), 359-377.

Lensink, M. F., & Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins, 81*(12), 2082-2095.

London, N., Raveh, B., Cohen, E., Fathi, G., & Schueler-Furman, O. (2011). Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Research, 39*(Web Server issue), W249-253.

MacCallum, R. M., Martin, A. C., & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *Journal of Molecular Biology, 262*(5), 732-745.

Morrone, J. A., Perez, A., MacCallum, J., & Dill, K. A. (2017). Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. *Journal of Chemical Theory and Computation, 13*(2), 870-876.

Nussinov, R., Papin, J. A., & Vakser, I. (2017). Computing the Dynamic Supramolecular Structural Proteome. *PLoS Computational Biology, 13*(1), e1005290.

Ohashi, E., Hanafusa, T., Kamei, K., Song, I., Tomida, J., Hashimoto, H., . . . Ohmori, H. (2009). Identification of a novel REV1-interacting motif necessary for DNA polymerase kappa function. *Genes to Cells, 14*(2), 101-111.

Olesen, L. E., Ford, M. G., Schmid, E. M., Vallis, Y., Babu, M. M., Li, P. H., . . . Praefcke, G. J. (2008). Solitary and repetitive binding motifs for the AP2 complex alpha-appendage in amphiphysin and other accessory proteins. *Journal of Biological Chemistry, 283*(8), 5099-5109.

Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., . . . Kozakov, D. (2016). Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proceedings of the National Academy of Sciences of the United States of America, 113*(30), E4286-4293.

Porter, K. A., Desta, I., Kozakov, D., & Vajda, S. (2019). What method to use for protein-protein docking? *Current Opinion in Structural Biology, 55*, 1-7.

Poy, F., Yaffe, M. B., Sayos, J., Saxena, K., Morra, M., Sumegi, J., . . . Eck, M. J. (1999). Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Molecular Cell, 4*(4), 555-561.

Regep, C., Georges, G., Shi, J., Popovic, B., & Deane, C. M. (2017). The H3 loop of antibodies shows unique structural characteristics. *Proteins, 85*(7), 1311-1318.

Ritchie, D. (2008). Recent Progress and Future Directions in Protein-Protein Docking. *Current Protein & Peptide Science, 9*(1), 1-15.

Rose, P. W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., . . . Burley, S. K. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research, 45*(D1), D271-D281.

Rudiger, S., Germeroth, L., Schneider-Mergener, J., & Bukau, B. (1997). Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO Journal, 16*(7), 1501-1507.

Schindler, C. E. M., de Vries, S. J., & Zacharias, M. (2015). Fully Blind Peptide-Protein Docking with pepATTRACT. *Structure, 23*(8), 1507-1515.

Schmidt, A. G., Xu, H., Khan, A. R., O'Donnell, T., Khurana, S., King, L. R., . . . Harrison, S. C. (2013). Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proceedings of the National Academy of Sciences of the United States of America, 110*(1), 264-269.

Schuetz, A., Allali-Hassani, A., Martin, F., Loppnau, P., Vedadi, M., Bochkarev, A., . . . Min, J. (2006). Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO Journal, 25*(18), 4245-4252.

Sheng, Y., Saridakis, V., Sarkari, F., Duan, S., Wu, T., Arrowsmith, C. H., & Frappier, L. (2006). Molecular recognition of p53 and MDM2 by USP7/HAUSP. *Nature Structural and Molecular Biology, 13*(3), 285-291.

Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology, 268*(1), 209-225.

Smith, D. M., Chang, S. C., Park, S., Finley, D., Cheng, Y., & Goldberg, A. L. (2007). Docking of the proteasomal ATPases' carboxyl termini in the 20S proteasome's alpha ring opens the gate for substrate entry. *Molecular Cell, 27*(5), 731-744.

Smith, G. R., & Sternberg, M. J. E. (2002). Prediction of protein–protein interactions by docking methods. *Current Opinion in Structural Biology, 12*(1), 28-35.

Soding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research, 33*(Web Server issue), W244-248.

Srinivasula, S. M., Hegde, R., Saleh, A., Datta, P., Shiozaki, E., Chai, J., . . . Alnemri, E. S. (2001). A conserved XIAP-interaction motif in caspase-9 and Smac/DIABLO regulates caspase activity and apoptosis. *Nature, 410*(6824), 112-116.

Stein, A., & Aloy, P. (2008). Contextual specificity in peptide-mediated protein interactions. *PLoS One, 3*(7), e2524.

Sternberg, M. J. E. G., H. A.; Jackson, R. M.; Moont, G. (2000). *Protein-Protein Docking: Generation and Filtering of Complexes*. In Vol. 143. W. D. M. (Ed.), *Protein Structure Prediction. Methods in Molecular Biology* (pp. 399-415).

Tharakaraman, K., Robinson, L. N., Hatas, A., Chen, Y. L., Siyue, L., Raguram, S., . . . Sasisekharan, R. (2013). Redesign of a cross-reactive antibody to dengue virus with broad-spectrum activity and increased in vivo potency. *Proceedings of the National Academy of Sciences of the United States of America, 110*(17), E1555-1564.

Tiller, K. E., & Tessier, P. M. (2015). Advances in Antibody Design. *Annual Review of Biomedical Engineering, 17*, 191-216.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M. I., & Dunbrack, R. L., Jr. (2010). Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Computational Biology, 6*(4), e1000763.

Vajda, S., Yueh, C., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., . . . Kozakov, D. (2017). New additions to the ClusPro server motivated by CAPRI. *Proteins, 85*(3), 435-444.

Vakser, I. A. (1996). Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers, 39*(3), 455-464.

Vakser, I. A. (2014). Protein-protein docking: from interaction to interactome. *Biophysical Journal, 107*(8), 1785-1793.

Wang, Q., Zhang, P., Hoffman, L., Tripathi, S., Homouz, D., Liu, Y., . . . Cheung, M. S. (2013). Protein recognition and selection through conformational and mutually

induced fit. *Proceedings of the National Academy of Sciences of the United States of America, 110*(51), 20545-20550.

Warbrick, E. (1998). PCNA binding through a conserved motif. *Bioessays, 20*(3), 195-199.

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics, 54*, 5 6 1-5 6 37.

Wodak, S. J., & Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of Molecular Biology, 124*(2), 323-342.

Wu, T. T., & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *Journal of Experimental Medicine, 132*(2), 211-250.

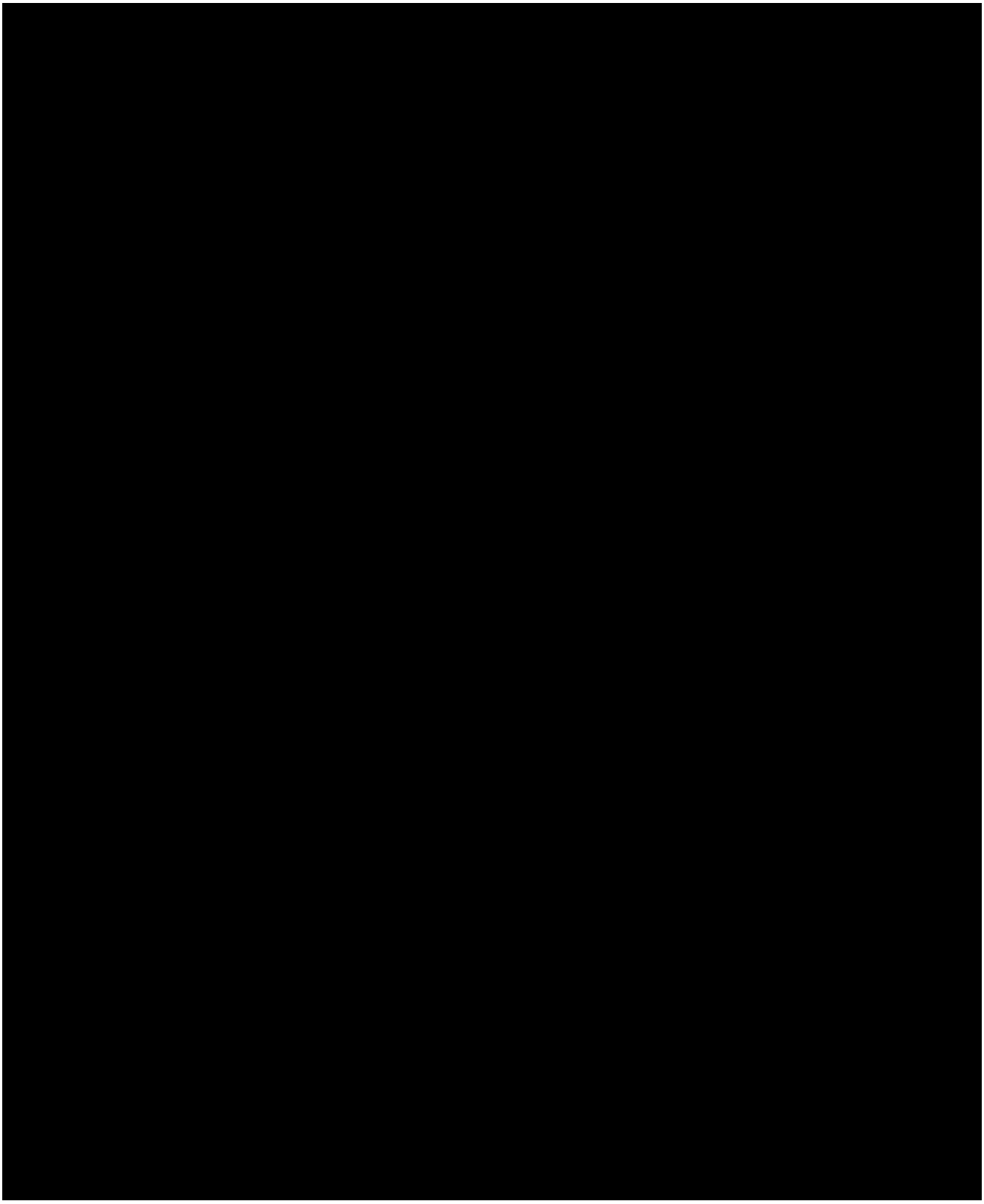
Xia, B., Mamonov, A., Leysen, S., Allen, K. N., Strelkov, S. V., Paschalidis, I., . . . Kozakov, D. (2015). Accounting for observed small angle X-ray scattering profile in the protein-protein docking server ClusPro. *Journal of Computational Chemistry, 36*(20), 1568-1572.

Xia, B., Vajda, S., & Kozakov, D. (2016). Accounting for pairwise distance restraints in FFT-based protein-protein docking. *Bioinformatics, 32*(21), 3342-3344.

Yueh, C., Hall, D. R., Xia, B., Padhorny, D., Kozakov, D., & Vajda, S. (2017). ClusPro-DC: Dimer Classification by the Cluspro Server for Protein-Protein Docking. *Journal of Molecular Biology, 429*(3), 372-381.

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kubler, J., Lozajic, M., . . . Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology, 430*(15), 2237-2243.

CURRICULUM VITAE



- Presented analyzed experimental results for critical review in weekly meetings with supervisors.

Univ. of Maine; Dept. of Chemical and Biological Eng. May 2009 - May 2013
Student Research Assistant V, PI: Dr. Darrell Donahue
- Participated in research which incorporated NIR in the analysis of wood extracts.
- Analyzed results using data modeling software and PLS regressions.

Univ. of Maine; School of Biology and Ecology Summer 2007 & Summer 2008
Student Research Assistant II, PI: Dr. Andrei Alyokhin
- Maintained insect colonies for pesticide resistance experiments.
- Participated in potato production research.

DESIGN PROJECTS

Ventricular Catheter Guidance System September 2013 – September 2014
- Designed a neurosurgical stereotactic system to improve ventricular catheter placement in the ICU setting.
- Developed design and performance specifications based on field observations, clinical interviews, and competitive research.
- Performed rapid-prototyping, risk-assessment, and design of killer experiments.

Miniature Fluorescence Microscope Design January 2013 – March 2013
- Produced a preliminary design for an integrated fluorescence microscope capable of *in-vivo* imaging of neuronal circuit changes in fluorescently-tagged mice.
- Utilized optical modeling and ray tracing for component placement.
- Selected components based on desired performance characteristics and project specifications.

SKILLS

Computer Skills and Frameworks
- Linux systems and shell scripting
- Strong programming skills in Python, MATLAB, SQL
- Website development using PHP, HTML, CSS
- Programming on high performance computing clusters

Bioinformatics Tools
- Sequence alignment and homology modeling including BLAST, HHSuite, Modeller
- Molecular viewing and analysis tools including PyMol, Maestro, Desmond, VMD

PUBLICATIONS

- **Porter KA,** Padhorny D, Desta I, Ignatov M, Beglov D, Kotelnikov S. et al. "Template-Based Modeling by ClusPro in CASP13 and the Potential for Using Co-evolutionary Information in Docking." *Proteins. Accepted, In Revision.*
- **Porter KA**, Desta I, Kozakov D, Vajda S. "What Method to Use for Protein-Protein Docking?" *Current Opinion in Structural Biology.* 2019, 55: 1-7.

- Alam N, Goldstein O, Xia B, **Porter KA**, Kozakov D, Schueler-Furman O. "High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock." *PLoS Comput Biol.* 2017, 13 (12): e1005905.
- **Porter KA**, Xia B, Beglov D, Bohnuud T, Alam N, Schueler-Furman O, Kozakov D. "ClusPro PeptiDock: Efficient global docking of peptide recognition motifs using FFT." *Bioinformatics.* 2017, 33 (20): 3299-3301.
- Kozakov D, Hall DR, Xia B, **Porter KA**, Padhorny D, Yueh C, Beglov D, Vajda S. "The ClusPro web server for protein-protein docking." *Nature Protocols.* 2017, 12 (2): 255-278.
- Padhorny D, Kazennov A, Zerbe BS, **Porter KA**, Xia B, Mottarella SE, et al. "Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds." *Proc Natl Acad Sci U S A*. 2016:113(30): E4286-93.

PRESENTATIONS

Oral Presentations

- **Porter KA,** Xia B, Beglov D, Bohnuud T, Alam N, Schueler-Furman O, Kozakov D. *ClusPro PeptiDock: Efficient global docking of peptide recognition motifs.* 12th Critical Assessment of protein Structure Prediction (CASP), Gaeta, Italy 2016.
- **Porter, KA**. *Prediction of peptide-protein interactions using motif-derived fragments*. Emerging Technologies in Computational Chemistry, 250th American Chemical Society National Meeting & Exposition, Boston, MA 2015.
- **Porter KA**, Bohnuud T, Kozakov D, Vajda S. *Evidence of Conformational Selection Driving the Formation of Ligand Binding Sites in Protein-Protein Interfaces*. Discovery On Target, Boston, MA 2014

Poster Presentations

- **Porter KA,** Sun Z, Ataca S, Kozakov D, Beglov D, Kepler T, Kolossvary I, Vajda S. *Docking provides insight into impact of mutation on antibody maturation. 13th Critical Assessment of protein Structure Prediction (CASP 13)*, Riviera Maya, Mexico 2018.
- **Porter KA,** Xia B, Beglov D, Bohnuud T, Alam N, Schueler-Furman O, Kozakov D. *ClusPro PeptiDock: Efficient global docking of peptide recognition motifs.* Conference on Modeling of Protein Interactions (MPI), Lawrence, KS 2018.
- **Porter KA,** Xia B, Beglov D, Bohnuud T, Alam N, Schueler-Furman O, Kozakov D. *ClusPro PeptiDock: Efficient global docking of peptide recognition motifs. 12th Critical Assessment of protein Structure Prediction (CASP 12)*, Gaeta, Italy 2016.
- **Porter KA,** Xia B, Beglov D, Bohnuud T, Alam N, Schueler-Furman O, Kozakov D. *Prediction of peptide-protein interactions using motif-derived fragments.* Conference on Modeling of Protein Interactions (MPI), Lawrence, KS 2016.
- **Porter KA,** Xia B, Beglov D, Alam N, Schueler-Furman O, Kozakov D. *Prediction of peptide-protein interactions using motif-derived fragments.* Gordon Research Conference on Computer Aided Drug Design, West Dover, VT 2015.
- **Porter KA**, Bohnuud T, Kozakov D, Vajda S. *Evidence of Conformational Selection Driving the Formation of Ligand Binding Sites in Protein-Protein Interfaces*. Discovery On Target, Boston, MA 2014