2020

Computational modeling of protein conformational change and molecular interactions

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

COMPUTATIONAL MODELING OF PROTEIN CONFORMATIONAL CHANGE AND MOLECULAR INTERACTIONS

by

ZHUYEZI SUN

B.A., Macalester College, 2014
M.S., Boston University, 2017

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

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

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

Fourth Reader

Karen N. Allen, Ph.D.
Professor of Chemistry
Professor of Materials Science and Engineering

Fifth Reader

Adrian Whitty, Ph.D.
Associate Professor of Chemistry

ACKNOWLEDGMENTS

I would like to sincerely thank Professor Sandor Vajda for being a supportive research advisor, and the rest of my committee for guidance and close collaboration. In addition, special thanks to Dr. Istvan Kolossvary and Professor Dima Kozakov for research advice and helpful discussions.

I would also like to extend my gratitude to current and a few former members at the Vajda group. Big thanks to long-time office buddy Katie Porter for lots of fun and intellectual conversations about research, coding, cats, crossword puzzles, etc. Also special thanks to Christine Yueh for tasty baked goods. I would also like to thank Bing Xia and David Hall for their tremendous help and detailed knowledge about research and programming.

Finally, I would like to express appreciation to my family including my parents, my brother and my grandparents. Thank you all for your encouragement, often via Skype calls in a different time zone. Also, a very special thank you to my husband Chen, thank you for your continuous unwavering support. I am very lucky to have you.

COMPUTATIONAL MODELING OF PROTEIN CONFORMATIONAL CHANGE AND MOLECULAR INTERACTIONS

ZHUYEZI SUN

Boston University College of Engineering, 2020

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

ABSTRACT

Protein conformational change and interactions with other molecules serve as the basis for many biological processes such as metabolic control and cell signaling that are important for drug design and development of diagnostics. Structural details of these changes and interactions can be experimentally determined by X-ray crystallography or nuclear magnetic resonance (NMR) methods, whereas binding affinity can be measured by a variety of tools such as surface plasmon resonance (SPR). Under many circumstances, direct experimental analysis is either not possible or too time-consuming and expensive. Therefore, it is important to develop computational methods capable of characterizing protein conformational change and modeling interacting complexes. This thesis details four projects which involve computational modeling of binding pocket dynamics, protein-small molecule interactions and the formation of protein-protein complexes.

The first project provides structure-based analysis of cryptic site opening. Cryptic sites are pockets formed in ligand-bound proteins but not observed in unbound protein structures. Through analysis of crystal structures supplemented by molecular dynamics (MD) with enhanced sampling techniques, it was shown that cryptic sites can be grouped

into three types: 1) “genuine” cryptic sites, which do not form without ligand binding, 2) spontaneously forming cryptic sites, and 3) cryptic sites impacted by mutations or off-site ligand binding. The second project details the effort to improve the accuracy of the solvent mapping server FTMap, which finds small molecule binding hot spots on proteins. More specifically, a statistical pairwise potential, which adopts the Decoy As Reference State (DARS) framework, was developed and added to the scoring function of FTMap. The third and fourth projects both involve using docking to predict the quality of protein-protein interactions in the form of binding affinity. The structural impact of a frequent mutation in the human pancreatic secretory trypsin inhibitor (PSTI) was explored using a hybrid approach of MD simulations and molecular docking. The fourth project aims at establishing a relationship between docked near-native hits and experimentally measured binding free energies of antibody-antigen interactions.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

ABSTRACT v

TABLE OF CONTENTS vii

LIST OF TABLES xi

LIST OF FIGURES xiii

LIST OF ABBREVIATIONS xvi

CHAPTER 1 Introduction 1

1.1 Motivation 1

1.2 Protein-Protein Docking with ClusPro 2

1.2.1 FFT-Based Docking Program PIPER 3

1.2.2 Scoring Function 4

1.2.3 The Automated Protein Docking Server ClusPro 6

1.3 Computational Solvent Mapping with FTMap 6

1.4 Applications of Molecular Dynamics 10

1.5 Contributions 12

CHAPTER 2 Structure-Based Analysis of Cryptic Site Opening 13

2.1 Introduction 13

2.2 Methods 18

2.2.1 The Extended CryptoSite Data Set 18

2.2.2 Adiabatic Biased Molecular Dynamics 19
2.2.3 Calculating Druggability Scores 20
2.3 Results 21
2.3.1 Proteins in the CryptoSite Set 21
2.3.2 Group 1: Proteins that Require Ligand Binding for Forming the Cryptic Site 22
2.3.3 Group 2: Proteins with Spontaneously Forming Cryptic Pockets 29
2.3.4 Group 3: Cryptic Site Opening Impacted by Mutations or Off-Site Binding .. 33
2.4 Discussions and Conclusions 40
CHAPTER 3 Pairwise Statistical Potentials for Protein-Small Molecule Interactions 43
3.1 Introduction 43
3.2 Methods 48
3.2.1 Collection of Structural Data for Developing the DARS Potential 48
3.2.2 Construction of the Protein-Small Molecule DARS Potential 52
3.2.3 Incorporating DARS into FTMap 54
3.2.4 Testing Solvent Mapping with the New DARS 57
3.2.5 Evaluation of Mapping Results 58
3.3 Results and Discussions 59
3.3.1 The New Small Molecular Atom Types 59
3.3.2 The Small Molecule DARS Interaction Energies 62
3.3.3 Fractional Overlap with Crystal Ligands 65
3.3.4 Case Studies 70
3.4 Conclusions and Future Directions 76

CHAPTER 4 Modeling the Interaction between Trypsin and the Pancreatic Secretory Trypsin Inhibitor (PSTI) Encoded by *SPINK1* 78
4.1 Introduction 78
4.2 Methods 82
4.2.1 Preparation of Structures 82
4.2.2 Molecular Dynamics and Clustering Protocol 84
4.2.3 Ensemble Docking 86
4.3 Results and Discussions 87
4.3.1 Molecular Dynamics Results 87
4.3.2 Ensemble Docking Results 89
4.4 Conclusions 92
CHAPTER 5 Modeling of Antibody-Antigen Interactions 94
5.1 Introduction 94
5.2 Methods 98
5.2.1 Data Selection and Preparation 98
5.2.2 Side Chain Packing 100
5.2.3 Rigid Body Docking Using a Box 101
5.2.4 Connecting Docking and Binding Free Energy 102
5.2.5 MD Ensemble Docking 105
5.3 Results and Discussions 107
5.3.1 Rigid Body Docking Case Studies 107
5.3.2 Docking of Lysozyme to Antibody MD Ensemble 115

5.4 Conclusions 117

Appendix A: Supplemental Information for Chapter 2 119

Appendix B: Supplemental Information for Chapter 5 121

BIBLIOGRAPHY 128

CURRICULUM VITAE 145

LIST OF TABLES

Table 3.1: The set of nine Tripos MOL2 atom types present in the 16 FTMap small molecule probes and their original definitions. 59

Table 3.2: The set of 11 ACP atom types present in the 16 FTMap small molecule probes and their original definitions. 60

Table 3.3: Pairwise contact energies $\varepsilon_{i,j}$ of small molecule DARS in kcal/mol. 63

Table 3.4: Top ten eigenvalues and eigenvectors for the new DARS potentials, rank-ordered by the magnitude of eigenvalues.. 65

Table 3.5: Overlap between crystal ligands and hot spots (or CSs). 66

Table 5.1: Summary of antibody-antigen cases in the data sets by Purisima and colleagues. 99

Table A.1: List of 32 proteins used for analysis in Chapter 2. 119

Table A.2: TEM β-lactamase druggability scores (DS) and mutations. 120

Table B.1: The ten sets of docking parameters. 121

Table B.2: Detailed docking results of antibody HyHEL-63 and antigen HEWL (wildtype PDB ID: 1dqj). 122

Table B.3: Detailed docking results of antibody Herceptin and antigen HER2 (wildtype PDB ID: 1n8z). 123

Table B.4: Detailed docking results of antibody bH1 and antigen HER2 (wildtype PDB ID: 3be1). 124

Table B.5: Detailed docking results of antibody D1.3 Fv and antigen hen egg white lysozyme (wildtype PDB ID: 1vfb). 126

Table B.6: Detailed MD ensemble docking results of antibody D1.3 Fv and antigen hen egg white lysozyme (parent PDB ID: 1vfb).....127

LIST OF FIGURES

Figure 2.1: Forming the pocket at the site of high affinity phosphotyrosine binding in PTP1B. 24

Figure 2.2: Conformational change and a snapshot from the ABMD simulation of protein tyrosine phosphatase 1B (PTP1B). 25

Figure 2.3: Druggability scores (DSs) of unliganded structures of proteins with DS distributions skewed toward the unbound state. 28

Figure 2.4: Forming the cryptic ligand binding site in beta-secretase 1 (BACE-1). 30

Figure 2.5: Conformational change and a snapshot from the ABMD simulation of BACE-1. 31

Figure 2.6: Druggability scores (DSs) of unliganded structures of proteins with a cryptic site that is frequently well formed. 32

Figure 2.7: Opening the cryptic allosteric site in TEM-1 β-lactamase. 36

Figure 2.8: Conformational change and a snapshot from the ABMD simulation of TEM-1 β-lactamase. 37

Figure 2.9: Druggability scores (DSs) of unliganded structures of proteins with cryptic sites impacted by mutations or binding at distant sites. 39

Figure 3.1: The 16 molecular probes currently used in FTMap. 44

Figure 3.2: Collection of structural data from the training set. 51

Figure 3.3: Conversion from the asymmetric protein-small molecule pairwise potential matrix to the equivalent symmetric matrix. 55

Figure 3.4: Number of atom types assigned to elements carbon (C), nitrogen (N) and oxygen (O) in the 16 FTMap probes, using the Tripos MOL2 system (new DARS) and the ACP atom types (current DARS). 62

Figure 3.5: Performance statistics of the new DARS compared to the current FTMap.... 70

Figure 3.6: Mapping results of *Staphylococcal* nuclease (apo structure 1stn, grey cartoon). 71

Figure 3.7: Mapping results of beta lactamase precursor (apo structure 1djb, wheat surface)..... 72

Figure 3.8: Mapping results of carboxypeptidase A (apo structure 5cpa, wheat surface). 73

Figure 3.9: Mapping results of human thrombin (apo structure 1hxf, transparent surface). 74

Figure 3.10: Mapping results of liver alcohol dehydrogenase (apo structure 8adh, grey cartoon). 75

Figure 4.1: Key interactions at the PSTI-chymotrypsinogen interface. 80

Figure 4.2: Modeled interface of trypsin-human PSTI. 83

Figure 4.3: The ensemble docking protocol. 87

Figure 4.4: Distributions of RMSD values in the MD trajectories of the PSTIs, using all C-alphas in the bound structure (PDB ID 1cgi) as the reference. 88

Figure 4.5: Weighted average near-native hits from docking of PSTI MD snapshots to trypsin. 89

Figure 4.6: Weighted average near-native hits from docking of PSTI MD snapshots to trypsin, where the mutations introduced in crystallization process were reversed. ..91

Figure 5.1: A typical IgG antibody structure. 95

Figure 5.2: Focused docking of antigen to antibody using a box. 102

Figure 5.3: Interface of antibody HyHEL-63 and antigen HEWL (wildtype PDB ID: 1dqj). 108

Figure 5.4: Docking-predicted -ΔG plotted against experimentally measured -ΔG for HyHEL-63 and HEWL (wildtype PDB ID: 1dqj). 109

Figure 5.5: Interface of antibody Herceptin and antigen HER2 (wildtype PDB ID: 1n8z). 110

Figure 5.6: Docking-predicted -ΔG plotted against experimentally measured -ΔG for Herceptin-HER2 (wildtype PDB ID: 1n8z). 111

Figure 5.7: Interface of Fab bH1 and antigen HER2 (wildtype PDB ID: 3be1). 112

Figure 5.8: Docking-predicted -ΔG plotted against experimentally measured -ΔG for bH1-HER2 Fab (wildtype PDB ID: 3be1). 113

Figure 5.9: Interface of D1.3 Fv and antigen lysozyme (wildtype PDB ID: 1vfb)..... 114

Figure 5.10: Docking predicted -ΔG plotted against experimentally measured -ΔG for D1.3 Fv-lysozyme binding (wildtype PDB ID: 1vfb). 115

Figure 5.11: MD ensemble docking-predicted -ΔG plotted against experimentally measured -ΔG for D1.3 Fv binding to lysozyme (wildtype PDB ID: 1vfb)..... 116

LIST OF ABBREVIATIONS

ABMD ..... adiabatic biased molecular dynamics

ACP ..... Atomic Contact Potential

ADAPT ..... Assisted Design of Antibody and Protein Therapeutics

ADARS ..... asymmetric DARS

AMBER ..... Assisted Model Building with Energy Refinement

BACE-1 ..... beta-secretase 1

CAPRI ..... Critical Assessment of Prediction of Interactions

CDR ..... complementarity-determining region

CS ..... consensus site

DARS ..... Decoys As the Reference State

DS ..... druggability score

FBDD ..... fragment-based drug discovery

FEP ..... Free Energy Perturbation

FFT ..... Fast Fourier Transform

GAFF ..... general AMBER force field

GPCR ..... G protein-coupled receptors

GPU ..... graphics processing unit

GUI ..... graphical user interface

iRMSD ..... interface root-mean-square deviation

MD ..... molecular dynamics

MM/GBSA ..... molecular mechanics generalized Born surface area

MM/PBSA .....molecular mechanics Poisson-Boltzmann surface area

MSM.....Markov state model

NMR.....nuclear magnetic resonance

PDB ..... Protein Data Bank

PSTI.....pancreatic secretory trypsin inhibitor

PTP1B..... protein tyrosine phosphatase 1B

RMSD..... root-mean-square deviation

SiPMAB.....Single-Point Mutant Antibody Binding

SPINK1.....serine peptidase inhibitor Kazal type 1

SPR..... surface plasmon resonance

TMD ..... targeted molecular dynamics

CHAPTER 1 Introduction

1.1 Motivation

Proteins play a central role in the functions and organizations of biological systems. Protein-protein interactions can be studied in genome-wide proteomics such as yeast two-hybrid assays, and the structural information can also be determined through experimental analysis such as X-ray crystallography or nuclear magnetic resonance (NMR). However, for many protein complexes and conformational change, experimental elucidation is still impossible. Insightful alternatives are computational docking methods and simulations that can provide details for those cases at atomic level (Smith & Sternberg, 2002; Halperin, Ma, Wolfson, & Nussinov, 2002; Ritchie, 2008) (Hollingsworth & Dror, 2018), which have advanced our understanding of processes such as metabolic control, signal transduction, and gene regulation, etc. On the other hand, structural details such as binding sites of many small molecule-protein interactions are not easily captured by experimental analysis due to the weak binding of those small compounds and also their sensitivity to conformational change. Instead, small regions on protein surfaces that are major free energy contributors to binding of ligands, known as hot spots, are well-studied in drug discovery community (DeLano, 2002; DeLano, Ultsch, de Vos, & Wells, 2000; Thanos, DeLano, & Wells, 2006). Experimental techniques for determining those binding hot spots are time-consuming and costly, and thus the computational analogue of such screening method has been implemented (Hall & Enyedy, 2015). Also known as computational protein mapping, this method reproduces the available experimental solvent mapping results and provide wide applications such as

to fragment-based drug discovery (FBDD). For those reasons, developing methodologies for modeling protein conformational change, predicting protein-protein binding and also improving protein mapping methods are crucial tasks in the field of structure-based drug design.

1.2 Protein-Protein Docking with ClusPro

For transient protein-protein interactions not amenable to experimental analysis, computational docking approaches are of crucial importance to understanding the molecular details of those interacting proteins. Docking describes a computational scheme to find the best "fit" between a receptor and a ligand. More specifically, the goal of molecular docking is to predict the association of two proteins given their atomic coordinates (Halperin et al., 2002), as provided on Protein Data Bank (PDB). The first docking approaches emerged in late 1970s when Wodak and Janin published about their computational generations of several models for the bovine pancreatic trypsin inhibitor-trypsin complex (Wodak & Janin, 1978), and when Greer and Bush described a method for calculating the macromolecular surface (Greer & Bush, 1978). Those early approaches largely implemented the global search for docking poses assuming rigid body motion (Vakser, 2014). Since then, the number of experimentally acquired structural data has increased tremendously and the computational power has also fast improved. Sampling techniques such as Fast Fourier Transform (FFT) enabled the more rapid docking process with high numerical efficiency. The FFT algorithm developed by Katchalski-Katzir and colleagues essentially exploits the Fourier transform and

correlation techniques; the molecular surface of the protein undergoes a 3D digitalization and the geometric information is preserved in grids (Katchalski-Katzir et al., 1992). Henceforward several groups have incorporated this Fourier correlation and grid-based technique into their docking methods (Sternberg, Gabb, Jackson, & Moont, 2000). The Vajda lab developed the docking program PIPER (Kozakov, Brenke, Comeau, & Vajda, 2006) and the automated docking server ClusPro implementing the algorithm (Kozakov et al., 2013). ClusPro is heavily used (by October 2019 it had over 13,000 users and 330,000 docking jobs) and was the best performer among servers for many rounds of Critical Assessment of Prediction of Interactions (CAPRI) (Vajda et al., 2017).

1.2.1 FFT-Based Docking Program PIPER

PIPER performs a global and systematic sampling in discretized 6D space of two interacting proteins (Kozakov et al., 2006). The larger protein is considered the receptor and its center of mass is fixed at the origin of the coordinate system, while the smaller protein is considered the ligand and its rotational and transitional poses are evaluated. The energy-like scoring function is defined on a grid, which is then expressed as the sum of P correlation functions for all possible translations α, β, γ of the ligand with respect to the receptor. In the following equation, $R_p(i,j,k)$ and $L_p(i,j,k)$ indicate the components for the receptor and the ligand, respectively.

$$E(\alpha,\beta,\gamma) = \sum_p \sum_{i,j,k} R_p(i,j,k) L_p(i+\alpha, j+\beta, k+\gamma) \qquad [1.1]$$

Using P forward and one inverse FFT, namely FT and IFT, the above expression could be calculated efficiently. In the following formulation, $i = \sqrt{-1}$; N_1, N_2 and N_3 are the dimensions of the grid along the three coordinates.

$$E(\alpha, \beta, \gamma) = IFT\left\{\sum_p^p FT^*\{R_p\}FT\{L_p\}\right\}(\alpha, \beta, \gamma) \quad [1.2]$$

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

$$IFT\{f\}(i, j, k) = \frac{1}{N_1 N_2 N_3}\sum_{l,m,n} f(l, m, n)e^{2\pi i(\frac{li}{N_1}+\frac{mj}{N_2}+\frac{nk}{N_3})} \quad [1.4]$$

Assuming $N_1 = N_2 = N_3$, the transformation allows the efficiency of the calculation to become $O(N^3 \log(N^3))$ instead of $O(N^6)$. Note that PIPER also assumes rigid body. As a result, PIPER can exhaustively sample billions of conformations of protein-protein complexes.

1.2.2 Scoring Function

The energy function of PIPER (Kozakov et al., 2013; Kozakov et al., 2006) is

$$E = E_{attr} + \omega_1 E_{rep} + \omega_2 E_{elec} + \omega_3 E_{pair} \quad [1.5]$$

The coefficients ω_1, ω_2 and ω_3 specify the weights of each energy terms, and can be customized for specific docking problems (Kozakov et al., 2017). The first two terms, E_{attr} and E_{rep} denote the attractive and repulsive van der Waals interactions, also representative of shape complementary. Together these two terms are defined on a grid. The third term E_{elec} accounts for the electrostatic energy contribution with the following details:

$$E_{elec} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i q_j}{\sqrt{r_{ij}^2 + D^2 e^{(\frac{-r_{ij}^2}{4D^2})}}} \quad [1.6]$$

This electrostatic term is given by a simplified generalized Born-type expression. The last term E_{pair} represents the pairwise contact potential, where it can be written as $E_{pair} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{i,j}$. The N_R and N_L in those equations specify the numbers of atoms in the receptor and the ligand, respectively. The term $\varepsilon_{i,j}$ is a structure-based statistical potential, and here it describes the energy contribution by a pair of atoms that are within a certain cutoff distance from each other (Kozakov et al., 2013; Kozakov et al., 2006). The calculation of $\varepsilon_{i,j}$ was approached with the Decoys As the Reference State (DARS) formulation, which is a simple and natural approach to the construction of knowledge-based intermolecular potentials. By definition,

$$\varepsilon_{i,j} = -\mathrm{RT} ln(\frac{P_{i,j}^{nat}}{P_{i,j}^{ref}}) \quad [1.7]$$

where $P_{i,j}$ is the probability of interacting pairs within a defined sphere (Sippl, 1990). Using decoys as the reference state for $P_{i,j}^{ref}$ eliminates the error of treating protein surface residues as a uniform distribution, as those decoys were generated with only van der Waals terms in docking and are thus essentially random complexes with good shape complementary (Chuang, Kozakov, Brenke, Comeau, & Vajda, 2008; Kozakov et al., 2006). The DARS potential was obtained from a nonredundant database of native protein-protein complexes. Therefore E_{pair} was parameterized with real structural data and it has empirical significance (Chuang et al., 2008). This pairwise statistical potential has substantially improved the docking results of PIPER.

1.2.3 The Automated Protein Docking Server ClusPro

The ClusPro web server is a fully automated tool for protein-protein docking (Kozakov et al., 2013; Kozakov et al., 2017). The server is built on the PIPER rigid body docking algorithm as described in Section 1.2.1. After billions of conformations have been sampled using PIPER, the 1000 poses with the lowest energies are retained and then subject to root-mean-square deviation (RMSD)-based clustering. This clustering step outputs a list of clusters. The cluster center from the most-populated cluster represents the most likely model of the predicted protein complexes. The last step in ClusPro docking is refinement, which is CHARMM (Brooks et al., 1983) minimization for the purpose of removing steric clashes at the interface. Since its debut, the ClusPro server has been actively participating in CAPRI and is one of the best performing servers. There have also been multiple additions (Vajda et al., 2017) to the ClusPro server motivated by either CAPRI or demanding interest in the scientific community, including docking protocols specifically designed for antibody-antigen (Brenke et al., 2012), heparin (Mottarella et al., 2014), peptides (K. A. Porter et al., 2017), classification of biological vs crystal dimers (Yueh et al., 2017), restraint docking (Xia, Vajda, & Kozakov, 2016) based on prior knowledge of the complexes, etc.

1.3 Computational Solvent Mapping with FTMap

Methods such as experimental solvent mapping and computational protein mapping find hot spots, which are smaller regions included in binding sites of macromolecules. Hot spots are major contributors to binding free energy, and are hence

crucial to any ligand binding at that site (DeLano, 2002; DeLano et al., 2000; Thanos et al., 2006). Studying hotspots has much more advantage than the actual binding sites of small ligands on proteins. That is because hot spots are less sensitive to conformational changes than binding sites, and can be determined even on an unbound protein (Kuttner & Engel, 2012; Dennis, Kortvelyesi, & Vajda; Landon et al., 2009; Silberstein et al., 2003). Experimentally, one can screen libraries of fragment-sized organic molecules by soaking the target proteins in series of probes, and then NMR and X-ray crystallography are used to determine the structural details of the binding of those probes to the target protein (Allen et al., 1996; Ciulli, Williams, Smith, Blundell, & Abell, 2006; Hajduk, Huth, & Fesik, 2005; Mattos & Ringe, 1996). The knowledge of hot spots has multiple direct applications. For instance, the probes cluster at hot spots and the hit rates at these sites indicate the importance of the sites such as druggability (Hajduk et al., 2005; Mattos & Ringe, 1996). However, those experimental procedures are time-consuming and not economically favorable, and thus a virtual mapping method is highly desirable. The Vajda lab developed the FTMap server to implement the computational protein soaking experiments to globally sample the protein surface for binding hot spots (Brenke et al., 2009; Kozakov, Grove, et al., 2015). Just like ClusPro, FTMap is based on FFT correlation method that allows the global grid sampling to be computationally feasible as described in the previous sections. The probe set consists of 16 small organic molecules in varying sizes, shapes, and polarity. The program includes three main steps: 1) first, the probes were docked to the protein with FFT sampling; 2) then, energy-minimization was done to calculate the most favorable position of probes and individual type of probes

were clustered and ranked based on their Boltzmann averaged energies.; 3) the probe clusters themselves are clustered to generate the consensus sites (CSs), i.e., regions where the probe clusters overlap to form clusters. The CSs are ranked based on how many probe clusters they contain. The largest CS is defined as the most important hot spot, while the smaller CSs are considered as secondary hot spots that also contribute to ligand binding on the target protein (Brenke et al., 2009) (Kozakov, Grove, et al., 2015). Using FTMap, we were able to demonstrate in a few systems that ligand moieties interacting with the hot spots are in fact free energy anchors in binding (Kozakov, Hall, Jehle, et al., 2015). Results from ligand deconstruction studies show that a top-ranked hot spot dominates the binding free energy on protein surface while surrounding satellite hot spots, which are weaker, confer intrinsic selectivity and improved binding affinities (Kozakov, Hall, Jehle, et al., 2015). From this perspective, FTMap can provide great insight for fragment-based drug discovery (FBDD), a widely used tool in medicinal chemistry (Hajduk & Greer, 2007; Murray, Verdonk, & Rees, 2012; Erlanson, McDowell, & O'Brien, 2004), since FBDD relies on the premise that fragment binding mode will be conserved when expanded to a larger ligand. It has also be shown that FTMap (Kozakov, Grove, et al., 2015) surpasses classical mapping methods including GRID (Goodford, 1985) and MCSS (Miranker & Karplus, 1991) in terms of accuracy, and FTMap is also much faster than mixed molecular dynamics-based protein mapping (Bakan, Nevins, Lakdawala, & Bahar, 2012; Lexa & Carlson, 2013; Raman, Yu, Lakkaraju, & MacKerell, 2013). As of October 2019, the server FTMap has over 3000 registered users and has run more than 75000 mapping jobs.

Despite success and popularity, the current version of FTMap still has room for improvement in terms of its energy function. FTMap has the energy expression of:

$$E = E_{attr} + \omega_1 E_{rep} + \omega_2 E_{elec} + \omega_3 E_{cavity} + \omega_4 E_{pair} \quad [1.8]$$

which is similar to the energy expression implemented in protein docking shown in equation [1.5]. There is an additional term E_{cavity} that rewards hydrophobic enclosures. The pairwise statistical potential E_{pair} comes from the DARS potential originally developed for protein-protein docking and was extended to describe protein-small molecular probe interactions (Brenke et al., 2009). In other words, the DARS potential used in FTMap was developed *ad hoc* instead of systematically like the protein-protein DARS potential for ClusPro. The simplicity of the FTMap energy function is very robust for sampling with FFT in mapping alone, but to further establish FTMap's application in FBDD, the refinement of the energy function is of immediate interest. For instance, in FBDD, ideally when comparing binding modes, the mapping probes should have chemical properties most representative of the target fragments. The diversification and customization of the probes require a more discriminative energy function. In addition, it is desirable to generate conformers from mapping that are predicative of the fragment binding pose. Both of the above considerations point to the significance of E_{pair} for small molecules. The effort to systematically derive a protein-small molecule DARS potential will be described in further details in Chapter 3.

1.4 Applications of Molecular Dynamics

Molecular dynamics (MD) simulations are not new technologies; the first MD simulations of a protein was published in 1977 (Hollingsworth & Dror, 2018; McCammon, Gelin, & Karplus, 1977). The underlying basic principles are Newton's laws of motion. To start the simulation, one is given the positions of all atoms in a biomolecular system and then the forces exerted by all other atoms are calculated. The output is the predicted spatial position of each atom as a function of time. The end result is the MD trajectory, which is basically a movie showing the physical movements of all atoms in the biomolecular system. In recent years, MD simulations have become more popular and routinely used by both computational and experimental scientists due to several reasons. First of all, with the breakthroughs in structural biology techniques such as cryo-EM, the number of solved experimental structures have increased tremendously, including historically difficult classes such as ion channels, G protein-coupled receptors (GPCRs), etc. Since the success of molecular dynamics depends on the availability of the initial structure at atomic level of details, the increase in the amount of deposited structural data plays an important role in promoting applications of MD simulations (Hollingsworth & Dror, 2018). Second, both computer hardware and MD software have become much more powerful and accessible. For example, the new technology graphics processing units (GPUs) now enables the completion of a microsecond simulation in a couple of days. Much effort has also been put into lowering the learning curve for conducting MD simulations; many MD software packages also support graphical interface with simplified system preparation protocols, improving user experience

(Hollingsworth & Dror, 2018). Accordingly, the number of publications featuring MD simulations in the top 250 journals (ranked by impact factor) has increased from ~400 to ~1000 from the year 2007 to 2017 (Hollingsworth & Dror, 2018).

MD simulations can provide a wide range of information. For example, one can observe the physical transformation of a protein by viewing a MD trajectory, which can then reveal the dynamic behavior of that protein and supply answers to biologically-relevant questions (Hollingsworth & Dror, 2018). Chapter 2 in this thesis will discuss the usage of MD to form cryptic pockets, and the conditions of such MD simulations provide qualitative explanations for the energetics of pocket formation. Another application of MD is to determine how some perturbation such as mutation in amino acid residues can change the behavior of a biomolecular system (Hollingsworth & Dror, 2018). This approach, in combination with our in-house rigid body docking program PIPER, can help answer questions that involve structural changes at the protein-protein interface. Chapters 4 and 5 will discuss the application of ensemble docking strategies supplemented by MD simulations.

It is important to note that many biologically relevant processes need relatively long timescales. Such experiments can become too computationally expensive in unguided MD simulations, even with today's technologies. Fortunately, many enhanced sampling techniques are available for capturing long-timescale processes (Hollingsworth & Dror, 2018). The application of a strategy of biasing one protein conformation to another, known as adiabatic biased molecular dynamics (ABMD) (Harvey & Gabb, 1993; Marchi & Ballone, 1999; Paci & Karplus, 1999), will be discussed in detail in Chapter 2.

1.5 Contributions

Amanda Wakefield provided the druggability scores of X-ray structures in Chapter 2. Istvan Kolossvary designed MD experiments (Chapters 2 and 4) and performed simulations of antibodies in Chapter 5. Dmitri Beglov curated the extended CryptoSite set (Chapter 2) and collaborated with me on Chapter 3. The work in Chapter 5 was also a collaboration between myself and Kathryn Porter.

CHAPTER 2 Structure-Based Analysis of Cryptic Site Opening

2.1 Introduction

The binding of proteins to small molecules is central to various biological functions, including enzyme catalysis, receptor activation, and drug action, and thus detection, comparison and analyses of binding pockets are pivotal to structure-based drug design (Perot, Sperandio, Miteva, Camproux, & Villoutreix, 2010). In many proteins significant differences in protein conformation exist between the unbound and bound states, and in some cases the binding site is not even detectable in ligand-free structures. These so-called cryptic sites can be important for drug discovery because they can provide previously undescribed pockets and thus enable targeting of proteins that would otherwise be considered undruggable. For example, it was predicted that considering cryptic sites of the structurally characterized proteins increases the size of the potentially "druggable" disease-associated human proteome from ~40% to ~78% (Cimermancic et al., 2016). Thus, targeting of cryptic binding sites represents an attractive but somewhat under-explored approach to modulating protein function with small molecules (Acker et al., 2017) (Cimermancic et al., 2016). An important related question is whether the pockets are already present in some of the unliganded structures, since this information affects the choice of methods used for the identification of such sites.

The search for novel cryptic sites has been intensified with the improving performance of molecular dynamics (MD) simulation methods that have a history of successful applications (Durrant, Keranen, Wilson, & McCammon, 2010; Durrant & McCammon, 2011; Grant et al., 2011; Wagner et al., 2016; Wassman et al., 2013). More

recently, the development of Markov state models (MSMs) provided an even more powerful tool and stronger motivation for the discovery of cryptic sites (Bowman, Bolin, Hart, Maguire, & Marqusee, 2015; Bowman & Geissler, 2012; Hart et al., 2017; Knoverek, Amarasinghe, & Bowman, 2019; J. R. Porter et al., 2019). MSMs are built from extensive MD simulations to describe a protein's intrinsic dynamics, and provide a reduced view of the ensemble of spontaneous fluctuations the molecule undergoes at equilibrium, thereby identifying transient pockets and their probabilities (Bowman & Geissler, 2012). Recent MSM simulations revealed that the forming of ligand binding pockets at cryptic sites requires large cooperative changes to the surface of the protein, and that this property helps to identify such sites (J. R. Porter et al., 2019).

The goal of this project is to consider a set of proteins with validated cryptic sites, and to study whether the sites remain always cryptic without ligand binding, or pockets already form in some of the structures. In order to answer this question with some generality we want to study a substantial number of proteins rather than only a few. In spite of advances in methodology and computer speed, MD or MSM simulations are computationally still too demanding for a large-scale study, and hence we primarily investigate X-ray structures from the Protein Data Bank (PDB). However, for three proteins the results of the empirical analysis are supported by performing adiabatic biased molecular dynamics (ABMD) simulations (Harvey & Gabb, 1993; Marchi & Ballone, 1999; Paci & Karplus, 1999).

The starting point of our analysis is the CryptoSite set of protein pairs developed for benchmarking cryptic site detection algorithms (Cimermancic et al., 2016). Each of

the of 93 bound-unbound pairs in this set included an unbound structure without a well-formed pocket and another structure co-crystallized with a biologically relevant ligand bound at the same location. In our previous work we extended the set by adding all structures in the Protein Data Bank (PDB) having at least 95% sequence identity and no ligand bound within the 5 Å neighborhood of the cryptic site (Beglov et al., 2018). All structures in this extended set were mapped using the FTMap program (Kozakov, Grove, et al., 2015), and it was shown that the vicinity included a strong binding hot spot in some of the unbound structures for over 90% of the 93 proteins (Beglov et al., 2018). Since binding hot spots disproportionately contribute to the binding free energy of any ligand (DeLano, 2002; Hall, Kozakov, Whitty, & Vajda, 2015), and some attractive forces are clearly required for ligand binding, this result was not unexpected. However, binding hot spots can be located both in relatively flat surface regions and in crevices that are too tight to accommodate drug-sized ligands, and we did not investigate whether appropriate pockets were actually formed in any of the unbound structures. In fact, FTMap is not even suitable for such analysis, since its results are relatively invariant to conformational changes (Kozakov, Grove, et al., 2015; Kozakov et al., 2011).

To examine the statistics of pockets before any ligand binds, we considered the proteins in the CryptoSite set that had at least 10 apo structures in the Protein Data Bank (PDB). To characterize the pockets in the structures we calculated a druggability score (DS) at the cryptic site using the Fpocket program (Le Guilloux, Schmidtke, & Tuffery, 2009; Schmidtke, Le Guilloux, Maupetit, & Tuffery, 2010). Fpocket is more sensitive to conformational changes than FTMap. The Fpocket DS values vary between zero (no

pocket) and 1.0 (“perfect” for binding druglike small molecules). The number of structures for each protein is generally much higher than 10, and having multiple ligand-free X-ray structures enabled us to generate histograms of Fpocket druggability scores (Schmidtke & Barril, 2010).

As will be shown, the 32 protein can be grouped into the three different types. The first group includes eight proteins with cryptic sites that, based on the available X-ray structures, can be considered “genuine” since the pocket at the site does not form without ligand binding. In contrast, the apo structures of six proteins in the second group exhibit binding pockets that seem to spontaneously form in a substantial fraction of structures. Finally, in the largest group of 18 proteins forming of a pocket is impacted by off-site mutations or ligand binding, thus emphasizing the role of allosteric communication in the opening of the cryptic site. We assume that the X-ray structure of a protein correspond to the free energy minimum of the crystal under the condition of crystallization. However, the protein has an ensemble of slightly higher energy conformations (Hilser, Wrabl, & Motlagh, 2012; Motlagh, Wrabl, Li, & Hilser, 2014; Wrabl et al., 2011), and changes in the conditions of crystallization, introducing site directed mutations, or mutating some residues all perturb the free energy landscape and thereby can alter the X-ray structure. While analyzing the unliganded structures in the PDB provides some chance for capturing alternative structures, some possibly with better formed pockets, we readily admit that this approach is far from systematic. However, results still show the substantial information available in the PDB on the opening of cryptic sites.

To further explore how cryptic sites are formed, we selected one typical protein from each of the three groups and applied adiabatic biased molecular dynamics (ABMD) simulations (Harvey & Gabb, 1993; Marchi & Ballone, 1999; Paci & Karplus, 1999). The simulations use a biasing force to guide the proteins from their ligand-free structures to ligand-bound conformations. ABMD is similar to targeted molecular dynamics (Schlitter, Engels, & Kruger, 1994), but it is more gentle because the biasing force is only applied when the system is diverging from its path towards the target structure. Guiding the structures toward well-formed pockets enables rigorous sampling the transitions between the two states, generating a distribution of druggability scores of the pocket located at the cryptic site. By varying the value of a force constant, we can assess the extent of how energetically demanding such conformational transitions are. As mentioned, the three proteins studied by ABMD represent different pocket opening mechanisms. From the first group we consider the higher affinity phosphotyrosine (pTyr) binding pocket of protein tyrosine phosphatase 1B (PTP1B), which does not seem to form without binding a charged ligand. Accordingly, considerable force is needed in the simulation to guide the structure toward the ligand-bound conformation. In contrast, the active site of beta-secretase 1 (BACE1) is defined by a loop that can open and close essentially on its own, and hence not much force is needed to move it between the two states. The third protein we study is TEM-1 β-lactamase, in which the cryptic allosteric site is formed by moving apart two helices. As will be shown, the results of these simulations confirm the trends of pocket formation observed in the X-ray structures.

2.2 Methods

2.2.1 The Extended CryptoSite Data Set

The starting point of this work is a representative set of X-ray structures of proteins with validated cryptic binding sites. This set was originally selected for training and testing the CryptoSite cryptic site prediction protocol (Cimermancic et al., 2016), and hence is referred here as the CryptoSite set. This dataset consists of 93 bound-unbound pairs in which each unbound structure had a site considered cryptic due to its low pocket score, and each bound structure had a biologically relevant ligand bound at the site. While the original CryptoSite set included only one unbound structure in each pair, in order to study the information provided by different unbound structures of a given protein, for each bound structure in the set we added all unbound structures with at least 95% sequence identity that were available in the Protein Data Bank (Beglov et al., 2018). Structures with lower than 3.5 Å resolution were excluded. The structures were superimposed on the ligand-bound structure and structures with any ligand within 5 Å of the cryptic site ligand were also excluded. We also omitted any protein if none of its unliganded structure satisfied the FTMap druggability conditions according to our previous publication (Kozakov, Hall, Napoleon, et al., 2015). Finally, we removed all proteins that had less than 10 structures satisfying the above criteria. These selection criteria reduced the number of proteins considered in this study to 32 (Table A.1). The number of such unbound structures varied from 10 to 249 per protein.

2.2.2 Adiabatic Biased Molecular Dynamics

We applied adiabatic biased molecular dynamics (ABMD) simulations to three proteins, PTP1B, beta-secretase 1, and TEM-1 β-lactamase, all with well-validated cryptic sites. The simulations were performed using the GPU version of Desmond (Bowers et al., 2006) running on Nvidia GTX 1080 graphics cards on a 4-GPU desktop computer. We used the OPLSAA_2005 force field (Banks et al., 2005) and SPC water in our simulations. Every simulation started with an equilibration protocol including the following steps: (1) Brownian dynamics NVT, T = 10 K, Δt = 1 fs, restraints on solute heavy atoms, t = 100 ps, (2) NVT, T = 10 K, Δt = 1 fs, restraints on solute heavy atoms, t = 12 ps, (3) NPT, T = 10 K, restraints on solute heavy atoms, t = 12ps, (4) NPT, T = 310 K, Δt = 2.5 fs, restraints on solute heavy atoms, t = 12ps, and (5) NPT, T = 310 K, Δt = 2.5 fs, no restraints, t = 24 ps. The production runs were configured NPT using Nose-Hoover chain with a 1 ps relaxation time for thermostat (single temperature group), and Martyna-Tobias-Klein barostat with 2 ps relaxation time and isotropic coupling. We utilized a RESPA integrator with Δt = 2.5 fs for bonded and near nonbonded interactions and Δt = 7.5 fs for far nonbonded interactions. The particle-mesh Ewald algorithm was used with periodic boundary conditions to compute long-range electrostatic interactions with the real space cutoff set to 9 Å for both electrostatic and van der Waals interactions. Water molecules were constrained with SHAKE.

The ABMD simulations were used to guide a protein molecule from apo to holo structure (Harvey & Gabb, 1993; Marchi & Ballone, 1999; Paci & Karplus, 1999). ABMD is similar to targeted molecular dynamics (TMD) (Schlitter et al., 1994), but it is

gentler because the biasing force is only applied when the system is diverging from its path towards the target structure. The "distance" from the target ligand-bound conformation is measured by RMSD and when the system moves toward the target autonomously, no force is applied. The time-dependent ABMD/RMSD biasing potential, U is a function of the conformation of the protein, R, and at a time, t, is given by:

$$U(R,t) = \frac{1}{2}kH(\chi(R,t))\chi(R,t)^2$$

where H is a Heaviside function (H (χ) = 1 if χ > 0 and H(χ) = 0 otherwise), k is a force constant and χ(R, t) is:

$$\chi(R,t) = d(R(t),R_T) - \ min_{t'<t}d(R(t'),R_T)$$

$d(R_1,R_2)$ denotes the RMSD between conformations and R_1 and R_2, R_T is the target structure. By varying the value of the force constant k we were able to assess, qualitatively, the extent of how energetically challenged different conformational transitions were. For each system we ran three independent, short ABMD simulations (20 ns each, seeded with different initial random velocities). Values were recorded at 40 ps intervals, resulting in 502 frames for each trajectory. Frames from all 3 trajectories were combined for analysis.

2.2.3 Calculating Druggability Scores

We applied the Fpocket program (Le Guilloux et al., 2009; Schmidtke et al., 2010) with default parameters to calculate a druggability score (DS) at each of the 1506 frames collected for each MD trajectory of the three proteins. Same calculation was

applied for the each of X-ray structures. Since Fpocket captures all small transitional pockets in this process, DSs were calculated for all pockets within 5 Å of the ligand superimposed from the bound structure, and the maximum DS value was reported. The Fpocket DS values depend on shape, size, and polarity of the pocket and vary between zero (no pocket) and 1.0 (pocket ideal for binding druglike small molecules). We considered DS = 0.5 as the lower threshold for a well-formed pocket, and disregarded any protein if the cryptic pocket had DS < 0.5 in the ligand-bound structure.

2.3 Results

2.3.1 Proteins in the CryptoSite Set

The increasing number of X-ray structures determined under different conditions for the same proteins enabled us to study conformational variations, including the potential opening of cryptic pockets, in a large set of proteins, and thus arrive at conclusions that may have some level of generality. As already mentioned, the starting point of our study is the CryptoSite set of X-ray structures of proteins with validated cryptic binding sites (Cimermancic et al., 2016). As described in Section 2.2.1, we extended the original CryptoSite set with certain criteria. For this work, we restricted the analysis to 32 proteins that had at least 10 unbound structures satisfying the above criteria (Table A.1). Since we studied 32 proteins, detailed discussion had to be limited. Further details such as extended comments, supplemental figures of DS histograms and the complete list of selected ligand-free structures and their calculated DS values for all proteins are included in the supplemental material of the published work (Sun,

Wakefield, Kolossvary, Beglov, & Vajda, 2019).

2.3.2 Group 1: Proteins that Require Ligand Binding for Forming the Cryptic Site

A binding site can be considered genuinely cryptic if the binding pocket never forms without a bound ligand, and thus the DS distribution is strongly skewed toward small values, i.e., DS < 0.5 in all structures. Based on the X-ray structures of the 32 proteins considered, it appears that such proteins are relatively rare. In addition, even for these proteins there generally exist some exceptions. As will be discussed, proteins that have no detectable pockets in most unbound structures still may have such pockets due to either a mutation or ligand binding at a distant site that lead to opening the cryptic site without a bound ligand. Therefore, we indicate if the protein is a mutant or if it is a complex with a ligand or protein binding at a distant (non-cryptic) site. It is generally helpful that the structures in the PDB are supplemented by publications that provide information on the origins of cryptic site properties and help to explain why the exception may occur.

To demonstrate that some proteins are unlikely to form a pocket at the cryptic site without ligand binding we selected protein tyrosine phosphatase 1B (PTP1B), an extremely well-studied protein, in which the most important subsite of the active site is cryptic. This pocket is known as the site of the high affinity phosphotyrosine binding (Puius et al., 1997). In the CryptoSite set this site is represented by the unbound structure 2cm2 and by the structure 2h4k, co-crystallized with a small inhibitor. In Figure 2.1A we copied the inhibitor from the bound structure into the unbound one to show that in the

latter the binding site is broad and open rather than a drug-sized cavity. Such cavity forms in the inhibitor-bound structure (Figure 2.1B). Without ligand the binding site is open because loop 179 – 188 turns away from the site (Figure 2.2A). The loop moves closer to the site and forms a tight pocket in all bound structures, with the side chain of F182 acting as the lid (Figures 2.1B and 2.2A). The monocyclic thiophene inhibitor in 2H4K has low affinity (K_i = 1300- 3200 nM), but the same pocket binds an inhibitor with K_i = 4 nM in the PDB structure 2qbp. Although the pocket is very important for binding active site inhibitors and the phosphotyrosine moiety of substrates, it has a DS > 0.1 only in two unbound structures. The first is the C215D mutant (PDB ID 1pa1 DS = 0.468) and the second is the low-resolution structure 2hnp (DS = 0.338). We note that some of the 19 "unliganded" PTP1B structures in the PDB are mutants or have inhibitors binding far from the active site, but the pocket remains too open in all such structures.

Since DS > 0.1 was observed only in two of the 19 structures, the conformational transition may require overcoming some energy barriers. To test this hypothesis, we used adiabatic biased molecular dynamics (ABMD) simulations to guide the protein from its unbound to its ligand-bound state (see Section 2.2.2). The biasing force was proportional to the distance from the target structure, and was only applied when the system was diverging from its path towards this target structure. Figures 2.1D, 2.1E, and 2.1F show the distributions of DS values from the simulations with the force constants k=1.0 kcal/mol/Å^2, k=10.0 kcal/mol/Å^2, and k=60.0 kcal/mol/Å^2, respectively (see Section 2.2.2). At both k=1.0 kcal/mol/Å^2 and k=10.0 kcal/mol/Å^2 the distributions are heavily skewed toward low DS values, and the pocket is getting formed in a small fraction of

snapshots only when the much larger force, k=60.0 kcal/mol/Å^2, is applied. Figure 2.2B shows a snapshot at 12 ns from the latter simulation, attesting that loop moves toward its position in the ligand-bound structure.

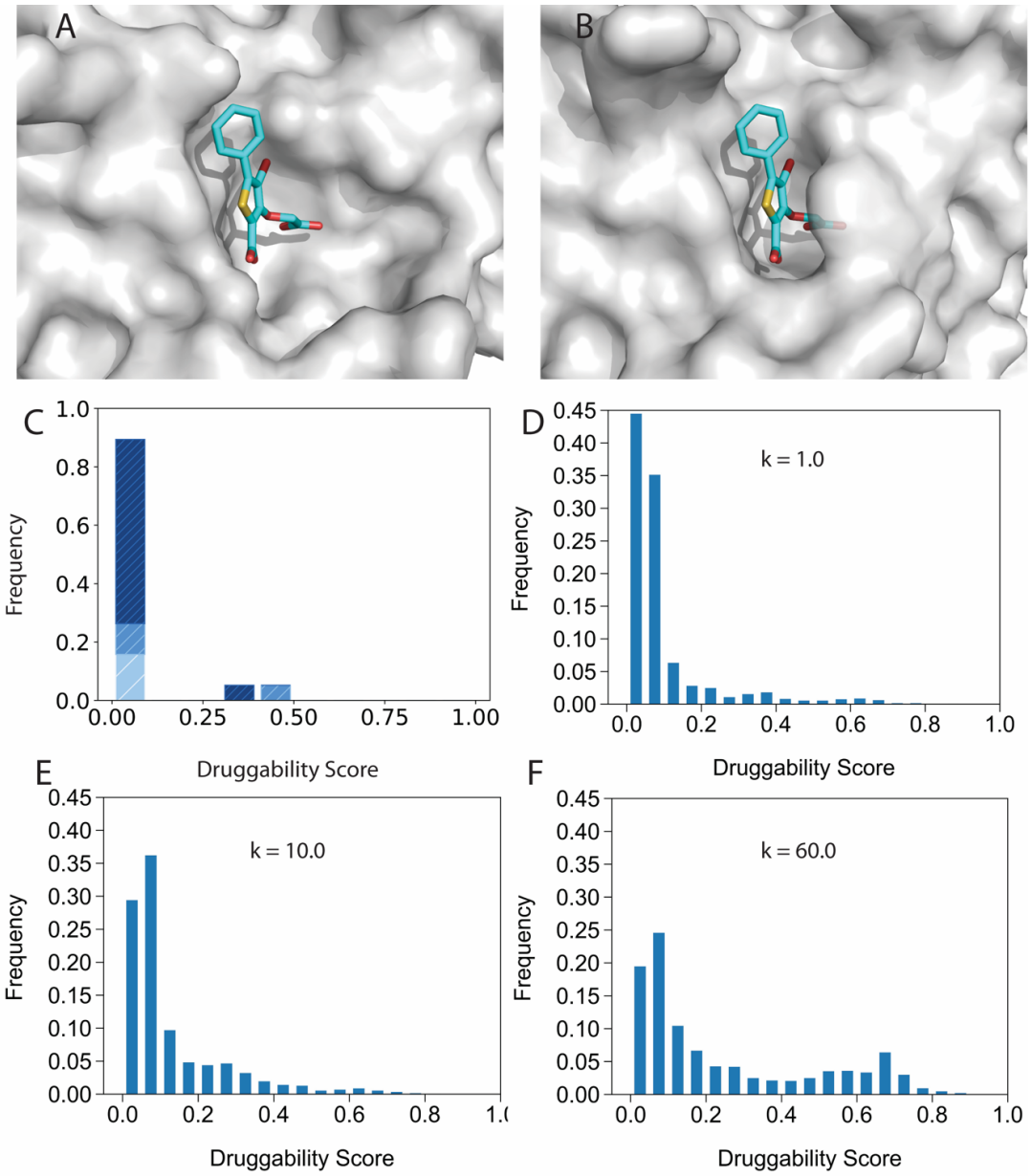

Figure 2.1: Forming the pocket at the site of high affinity phosphotyrosine binding in PTP1B. **(A)** Unbound PTP1B structure 2cm2 shown as grey surface. The inhibitor 509 from the ligand-bound structure 2h4k (cyan sticks) is shown for refence, demonstrating that the site is too open. **(B)** In the ligand-bound structure 2h4k the pocket binding the inhibitor 509 is well formed. The protein is shown as partially transparent surface for improved visibility. **(C)** Druggability scores (DSs) of unliganded

PTP1B structures in the PDB. The distribution of DS values is shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. Here "complex" means a protein or ligand binding at a distant site. **(D)** Distribution of druggability score (DS) values obtained by adiabatic biased molecular dynamics (ABMD) simulations of unliganded PTP1B at k=1.0 (kcal/mol)/Å^2. **(E)** Distribution of DS values obtained by ABMD simulations of PTP1B at k=10.0 (kcal/mol)/Å^2. **(F)** Distribution of DS values obtained by ABMD simulations of PTP1B at k=60.0 (kcal/mol)/Å^2.

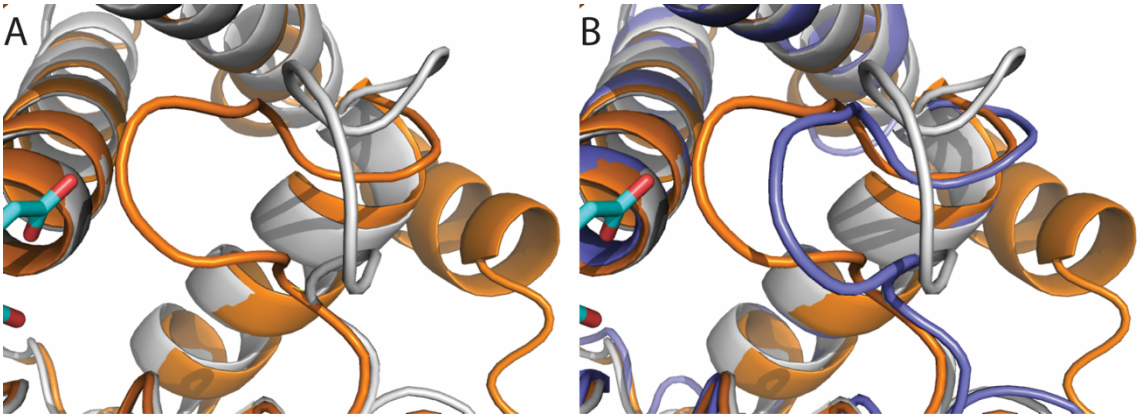

Figure 2.2: Conformational change and a snapshot from the ABMD simulation of protein tyrosine phosphatase 1B (PTP1B). All structures are shown in cartoon representation. **(A)** Loop 179 – 188 in the unbound structure 2cm2 (grey) and in the inhibitor-bound structure 2h4k (orange). **(B)** Loop 179 – 188 from a snapshot at t = 12 ns of the ABMD simulations with k = 60.0 kcal/mol/Å^2 (blue).

We show DS distributions for six more proteins with cryptic sites that almost never form without bound ligands (Figure 2.3). For clarity, the following discussions will be numbered.

(1) Pyruvate kinase from Leishmania Mexicana functions as a homotetramer, each subunit with substantial hinge motion between two domains. The active site of the enzyme has DS < 0.5 in all known ligand-free structures (Figure 2.3A). Similarly to PTP1B, the site is too open in these structures, and becomes well defined only upon binding to ATP and a substrate that cause the closing of a lid-like domain onto the site. We note that pyruvate kinase also has an allosteric site, which binds FDP (fructose 2,6 bisphosphate), almost 40 Å away from the active site, and some of the structures

considered in Figure 2.3A have FDP bound at the allosteric site. Figure 2.3A reveals that binding at the allosteric site does not affect the DS at the active site. In agreement with this result, pyruvate kinase is a known example of allostery without conformational change (Morgan et al., 2010).

(2) The active site of ricin (Figure 2.3B) is closed in most unbound structures because the side chain of Y80 protrudes into the site, stabilized with H-bond to the backbone of G121. Ligands such as pteroic acid displace Y80 and bind in the adenine pocket making specific hydrogen bonds to some active site residues (e.g., PDB ID 1br6). Without ligands the pocket is closed, but it may be affected by antibody binding at a distant site (PDB ID 4kuc), leading to DS > 0.5 in a few structures (Figure 2.3B).

(3) In ribonuclease A the cryptic site binds NADPH (PDB ID 2w5k), but in most apo structures the side chain of H119 protrudes into the site. The side chain has two alternative conformers, one of them turns out of the pocket as in the bound structure, and hence in a few structures the pocket is formed without ligand binding (Figure 2.3C). However, the H119 side chain conformation that protrudes into the pocket is stabilized by a strong H-bond with D121, and hence it is energetically preferable. Other structures with DS > 0.5 are due to a different crystallization condition or mutations.

(4) The CryptoSite set includes three cryptic allosteric sites of the hepatitis C virus polymerase NS5B. The first site is occupied by a small alpha-helix in the unbound structure 3cj0 at the tip of the N-terminal loop that connects the fingers and thumb domains. This loop becomes partially disordered in the inhibitor-bound structure 2brl, with no electron density present for residues 22–35. Another consequence of the

displacement of the helix and the resulting weaker interaction between the thumb and finger domains is a slight opening of the polymerase. Inhibitors binding at the site prevent intramolecular contacts between the two domains and consequently preclude their coordinated movements during RNA synthesis. Such conformational change does not occur in unliganded structures or, with the exception of a single complex (PDB ID 3bsc), in structures with inhibitors bound at the other two allosteric sites (Figure 2.3D). In contrast, it appears that inhibitor binding at this first site affects the pockets at the other two allosteric sites, and hence those will be discussed in the third group of proteins.

In addition to the cases discussed above, we place three more proteins into Group 1 with genuine cryptic sites: the C-terminal allosteric cryptic pocket in PTP1B (Figure 2.3E), fructose-1,6-bisphosphate aldolase (Figure 2.3F) and the Rho ADP-ribosylating Clostridium botulinum C3 exoenzyme with details given in the supplemental material in our publication (Sun et al., 2019).

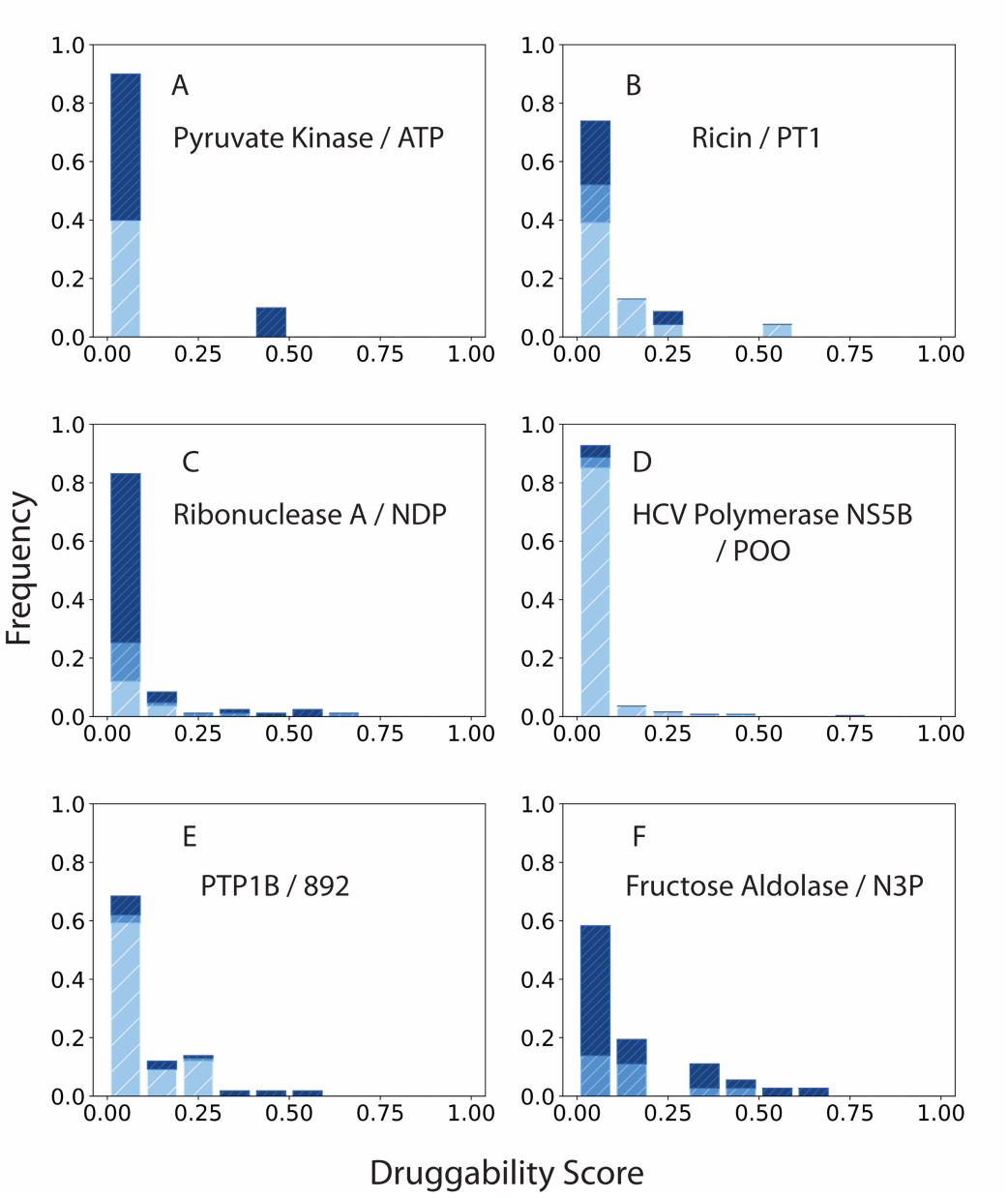

Figure 2.3: Druggability scores (DSs) of unliganded structures of proteins with DS distributions skewed toward the unbound state. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. The label shows the 3-letter code of the ligand bound at the cryptic site, and the name of the ligand is shown in parenthesis here. **(A)** Pyruvate kinase (ATP plus oxalate). **(B)** Ricin (pteroic acid). **(C)** Ribonuclease A (NADPH). **(D)** Hepatitis C virus RNA polymerase NS5B (indole-based allosteric inhibitor binding at the thumb domain). **(E)** Protein tyrosine phosphatase 1B (allosteric inhibitor binding at the C-end). **(F)** Fructose-1,6-bisphosphate aldolase enzyme from rabbit muscle (naphthol AS-E phosphate, a competitive inhibitor).

2.3.3 Group 2: Proteins with Spontaneously Forming Cryptic Pockets

As the other extreme we were looking for proteins with sites that were considered cryptic in CryptoSite, but have pockets that seem to spontaneously form in some of the ligand-free structures. Such behavior is seen in beta-secretase 1 (BACE1), represented by unbound and bound structures 1w50 and 3ixj in the CryptoSite set. In the unbound structures the loop comprising residues 71-74 is turned away from the site, making the pocket too open to score as druggable (Figures 2.4A and 2.5A). The loop is closing down on the inhibitor in the bound structure 3ixj (Figures 2.4B and 2.5A), resulting in a well-formed pocket that binds the isophthalamide ligand with high affinity (Bjorklund et al., 2010). The analysis of unbound BACE1 structures shows a broad distribution of druggability scores between conformations resembling the unbound and bound forms (Figure 2.4C), with 39% of structures with DS > 0.5. Apart from a single complex with an antibody bound far from the active site, all BACE1 structures in the PDB are of the wildtype human protein, and the various X-ray structures differ only in the crystal form and the conditions of crystallization. The overall root-mean-square deviation (RMSD) of many unbound structures with DS > 0.5 is less than 0.5 Å from the bound structure 3ixj. Thus, the variation in DS values seems to be the consequence of the variation in loop conformation, indicating significant conformational selection as part of the pocket opening. This hypothesis is supported by the results of biased molecular dynamics simulations. Indeed, simulation at k=1.0 kcal/mol/Å^2 and started from the apo state shows that the distribution of DS values is already somewhat skewed to the right, i.e., toward a well-formed binding pocket (Figure 2.4D), and loop 71-74 is getting close to its position

in the ligand-bound state as shown by a snapshot at t = 12 ns (Figure 2.5B).

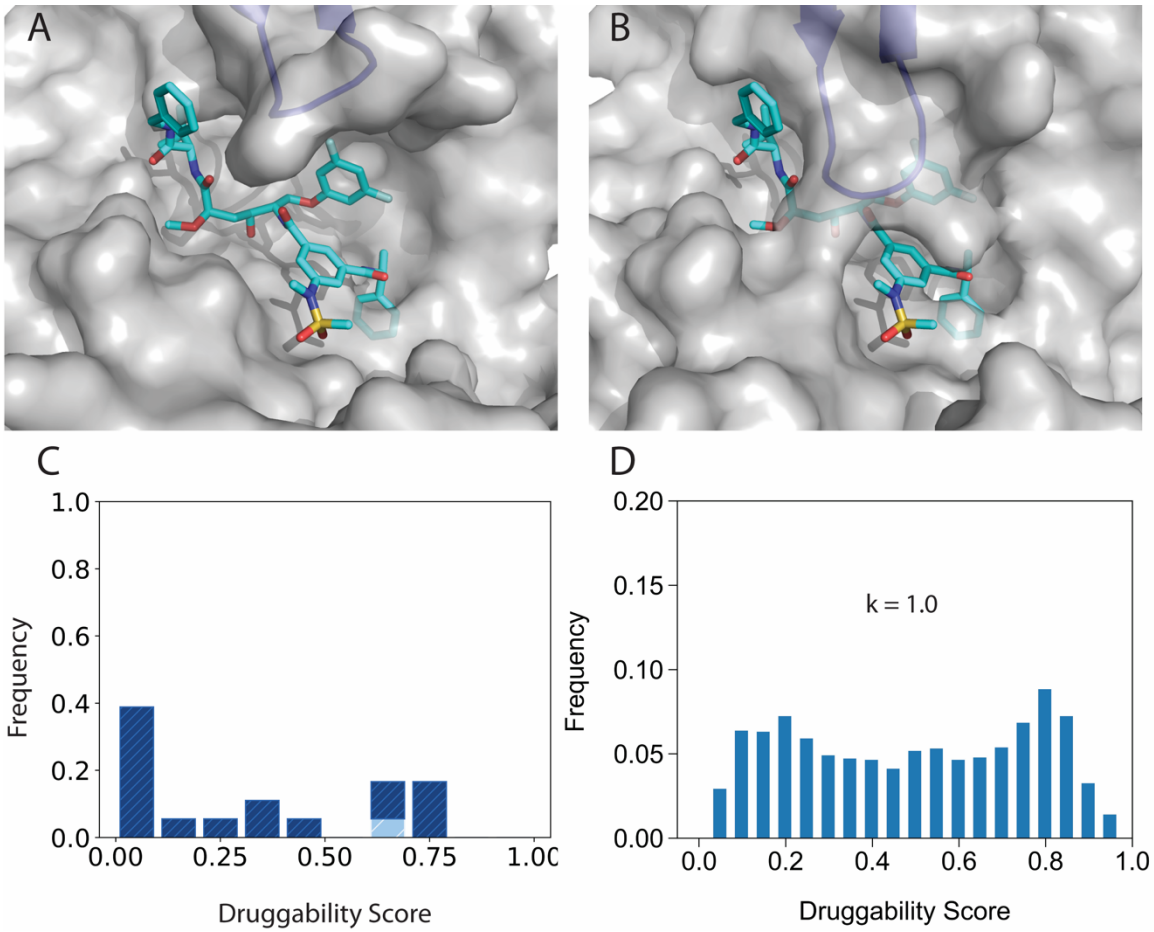

Figure 2.4: Forming the cryptic ligand binding site in beta-secretase 1 (BACE-1). **(A)** Unbound structure 1w50 (partially transparent grey surface). The inhibitor 586 from the ligand-bound structure 3ixj of BACE-1 is shown for reference (cyan sticks). The flexible loop 71-74 is shown as blue cartoon. **(B)** Structure 3ixj of BACE-1 (grey surface), co-crystallized with the inhibitor (cyan sticks). The flexible loop 71-74 is shown as blue cartoon. Based on the surface representation, the loop provides the lid of the inhibitor-binding pocket. **(C)** Druggability scores (DSs) of unliganded BACE-1 structures in the PDB. The distributions of DS values are shown in dark and light blue, respectively, for unbound structures and complexes. All structures are of the wildtype protein, and apart from a single structure with an exosite-binding antibody have no ligand bound. **(D)** Distribution of druggability score (DS) values obtained by adiabatic biased molecular dynamics (ABMD) simulations of BACE-1 at k=1.0 (kcal/mol)/Å^2.

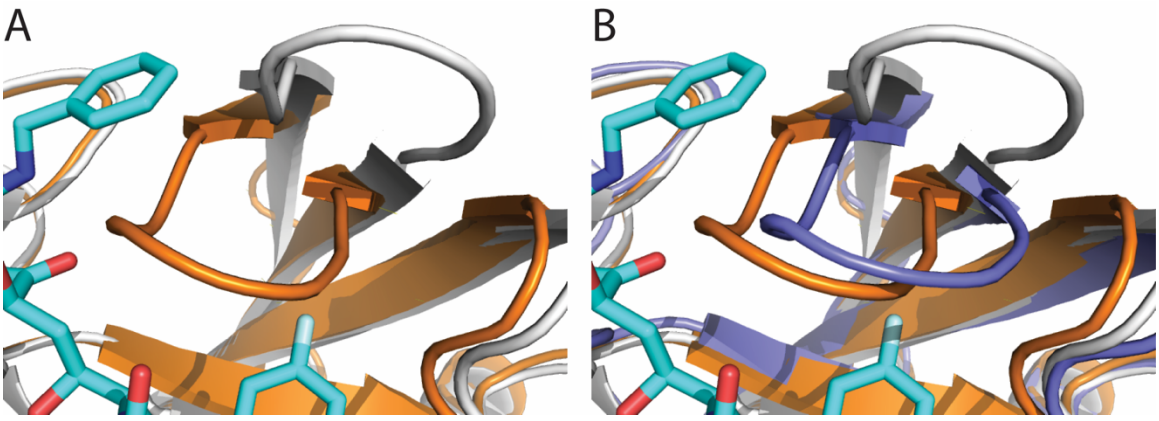

Figure 2.5: Conformational change and a snapshot from the ABMD simulation of BACE-1. All structures are shown in cartoon representation. **(A)** Loop 71-74 in the unbound structure 1w50 (grey) and in the inhibitor-bound structure 3ixj (orange). **(B)** Loop 71-74 from a snapshot at t = 12 ns of the ABMD simulations of BACE-1 with k = 1.0 kcal/mol/Å^2 (blue).

We have found only five other proteins with similar properties among the 32 studied. The first is bovine beta-lactoglobulin, which binds retinol in the middle of a β-barrel (PDB ID 1gx8). In a few unbound structures loop 84 to 90 acts as a lid that prevents access to the large and well-formed binding site. However, in most structures the flexible loop is open enough to provide access to the site (Figure 2.6A). Similarly, in a number of apo structures of human thrombin such as chain E of 1hag (which is actually a prothrombin), the active site is too open, but becomes well formed in many apo structures. Although there are mutant thrombins within the 95% sequence identity as well as complexes with ligands binding at distant sites, their impacts do not change the conclusion that the active site of thrombin can form spontaneously before any ligands bind (Figure 2.6B). The fourth protein in the CryptoSite set that does not seem to have a genuine cryptic site is the ligand binding domain of the alpha-L integrin lymphocyte function-associated antigen-1 (LFA-1). The cryptic site of LFA-1 binds an allosteric inhibitor (PDB ID 3bqm). In some structures without this inhibitor the disordered

carboxyl end protrudes into the site, but in others the binding pocket is well formed (Figure 2.6C). The fifth and sixth proteins in this group are the glutamate receptor 2 protein (Figure 2.6D) and the complex formed by the transforming protein RhoA and RhoGAP with details given in the supplemental material in our published work (Sun et al., 2019).

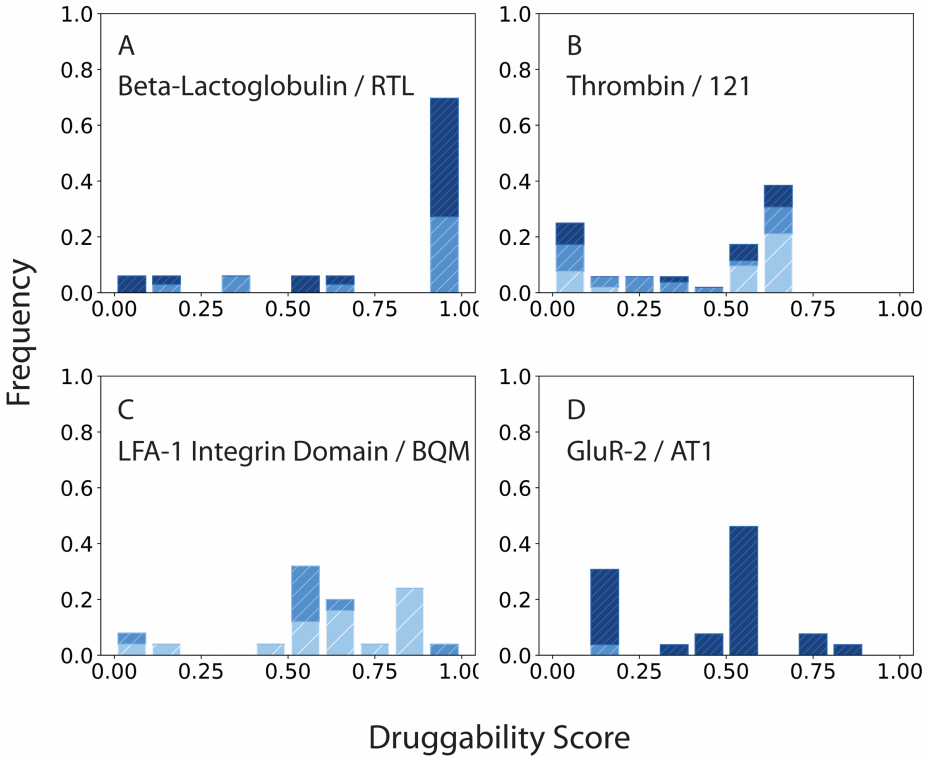

Figure 2.6: Druggability scores (DSs) of unliganded structures of proteins with a cryptic site that is frequently well formed. The ligand bound at the cryptic site is shown in the label and is listed in parenthesis here. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. **(A)** Bovine beta-lactoglobulin (retinol). **(B)** Thrombin (active site inhibitor 121). **(C)** Integrin lymphocyte function-associated antigen-1 (LFA-1) ligand binding (I) domain (inhibitor BQM). **(D)** Glutamate receptor 2 (competitive antagonist ATPO binding to the core of the receptor).

2.3.4 Group 3: Cryptic Site Opening Impacted by Mutations or Off-Site Binding

In the remaining 18 proteins the druggability score (DS) at the cryptic site substantially depends on mutations and/or on the binding of ligands or proteins at distant sites. Before discussing the other proteins, we focus on the impact of mutations on the opening of the cryptic site in TEM-1 β-lactamase, which is a textbook case of cryptic allosteric sites (Horn & Shoichet, 2004). The active site of TEM-1, with the catalytic residues S70, K73, and K234, is located between the two domains of the protein. In the unbound structures such as 1jwp helices H11 (residues 218–230) and H12 (residues 271–289), located above the active site on different domains, are close to each other (Figure 2.7A). The X-ray structure 1pzo showed two small inhibitors bound to this region by forcing apart the two helixes (Figures 2.7B and 2.8A). Although the center of this cryptic site is 16 Å from the center of the active site of the enzyme, one of the inhibitors has a second binding mode that partly occludes the active site near residues S235, G245, and G236 (Horn & Shoichet, 2004). However, it appears that binding to this second site would only be possible to a structure formed by inhibitor binding to the first, core site. This "opening" of the secondary structure results in major backbone and side-chain rearrangement that exposes mainly hydrophobic surface to the compound.

As shown in Figure 2.7C, the pocket at the cryptic site is deemed druggable (DS > 0.5) by Fpocket in over 50% of the unliganded lactamase structures. However, 19 of the 21 apo structures have some mutated amino acid residues. Introducing mutations represents the main mechanism by which opportunistic and pathogenic bacteria become resistant to β-lactam antibiotics, and hence many mutants have been generated. A

substantial number of studies examined how these mutations affect antibiotic resistance and stability (Abriata, Salverda, & Tomatis, 2012) (Brown, Pennington, Huang, Ayvaz, & Palzkill, 2010) (Dellus-Gur, Toth-Petroczy, Elias, & Tawfik, 2013) (Kather, Jakob, Dobbek, & Schmid, 2008) (Marciano et al., 2008) (Modi & Ozkan, 2018) (Orencia, Yoon, Ness, Stemmer, & Stevens, 2001) (Speck et al., 2012) (Stec, Holtz, Wojciechowski, & Kantrowitz, 2005) (Thomas et al., 2005) (X. Wang, Minasov, & Shoichet, 2002a, 2002b) (Knies, Cai, & Weinreich, 2017) (Latallo, Cortina, Faham, Nakamoto, & Kasson, 2017) (Zimmerman et al., 2017). Here we consider a new question and study how the mutations affect the druggability of the allosteric site. In general, we have observed that the stabilized mutants of β-lactamase (with higher melting temperatures T_m) have DS <0.2. On contrary, it appears that the mutations that reduce stability generally also yield a more open allosteric pocket. Since the allosteric site is located between the two domains of the protein, and the interactions between the domains affect both the stability of the protein and the volume of the cryptic site, this observation is not difficult to explain. More detailed discussions are featured in our publication (Sun et al., 2019). List of specific mutations and their DSs are summarized in Table A.2.

Since only two structures are available for the unliganded wild type TEM-1 β-lactamase, simple inspection does not provide information on the forces needed to open the site, and performing MD simulations is particularly important. Markov state models (MSMs) built from hundreds of microseconds of MD simulations have shown that the allosteric pocket was at least partially open for 53% of the simulation time (Bowman & Geissler, 2012). The cryptic pocket identified by the MSM simulations was also used for

the design of novel allosteric modulators (Hart et al., 2017). In contrast, MD simulations of the same protein by Gervasio and co-workers (Oleinikovas, Saladino, Cossins, & Gervasio, 2016). using parallel tempering failed to show appreciable opening of the site when starting from the apo crystal structure, possibly due to the lack of convergence. In order to reliably capture the conformational transition from the closed to open allosteric site we have applied the already discussed ABMD method to the M182T variant of the β-lactamase. Simulations at k=1.0 kcal/mol/Å^2 show that the pocket is already formed in some fraction of conformations, in good agreement with the MSM results (Bowman & Geissler, 2012). However, the pockets are only partially open, with the peak DS value around 0.6 (Figure 2.7D). It is interesting that the site has a "binary" behavior having closed and partially open states with limited intermediate conformations. This is in contrast to the site in BACE1, which has an almost flat distribution of DS values (Figure 2.4D). Increasing to biasing force to k=10.0 kcal/mol/Å^2 and then to k=30.0 kcal/mol/Å^2 increases the fraction of partially open sites (Figures 2.7E and 2.7F). However, even a high DS value does not necessarily mean that the allosteric site is fully open. For example, Figure 2.8B shows a snapshot at t = 20 ns from the simulation with k=30.0 kcal/mol/Å^2. Although this structure has a pocket with DS = 0.8 close to the site that binds the allosteric inhibitors, the pocket is created by some unfolding of the amino end of helix H11, and it can only partially accommodate one of the inhibitors.

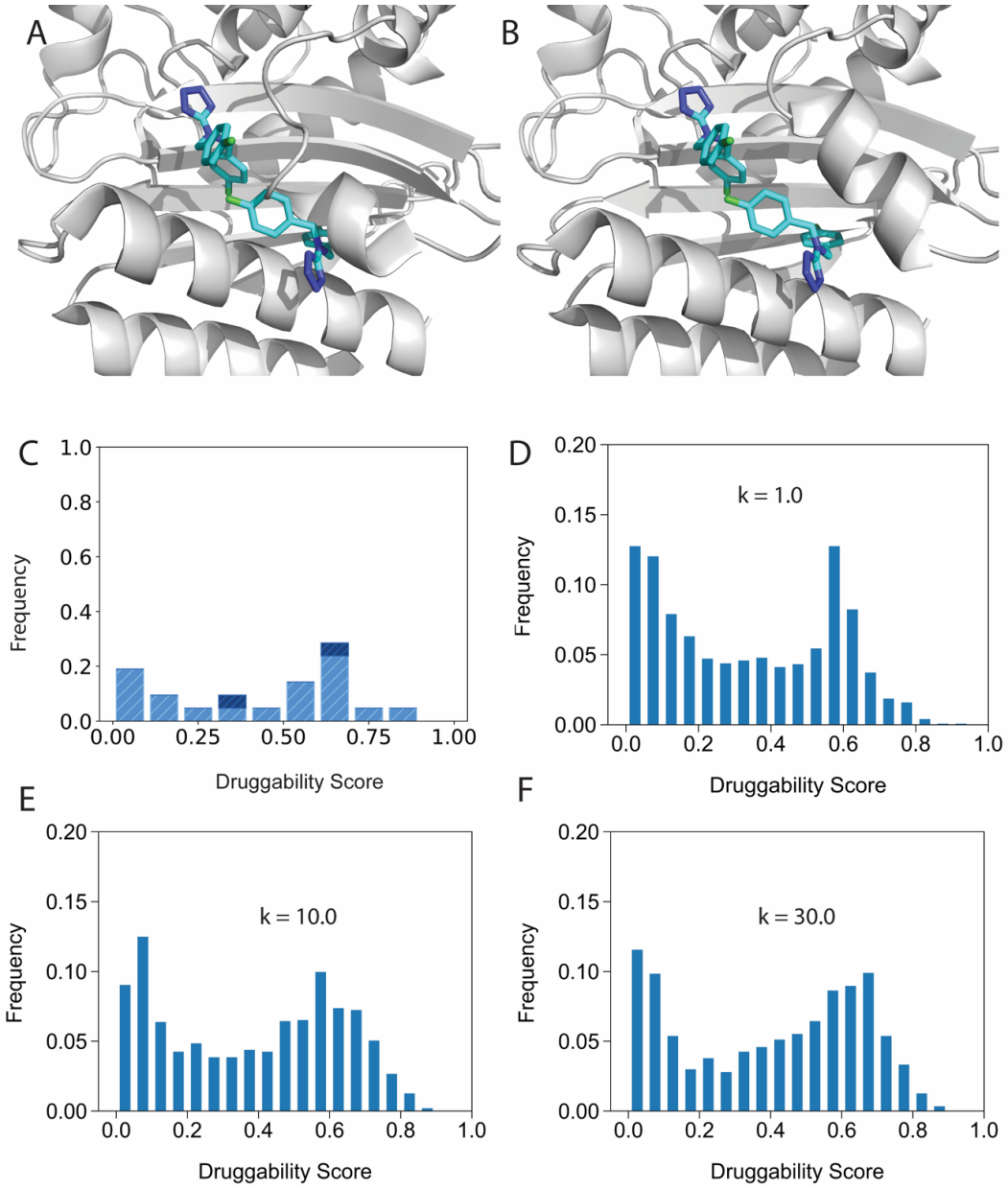

Figure 2.7: Opening the cryptic allosteric site in TEM-1 β-lactamase. **(A)** Unbound structure 1jwp of TEM-1 β-lactamase (grey cartoon). Two small allosteric inhibitors from the structure 1pzo are shown for reference (cyan sticks). **(B)** Inhibitor-bound structure 1pzo with two allosteric inhibitors, demonstrating that the two helices lining the allosteric site move apart. **(C)** Druggability scores (DSs) of unliganded TEM-1 β-lactamase structures in the PDB. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. Here "complex" means a protein or ligand binding at a distant site. **(D)** Distribution of druggability score (DS) values obtained by ABMD simulations of TEM-1 β-lactamase at k=1.0 (kcal/mol)/Å^2. **(E)** Distribution of DS values obtained by ABMD simulations of TEM-1 β-lactamase at k=10.0 (kcal/mol)/Å^2. **(F)** Distribution of DS values obtained by ABMD simulations of TEM-1 β-lactamase at k=30.0 (kcal/mol)/Å^2.

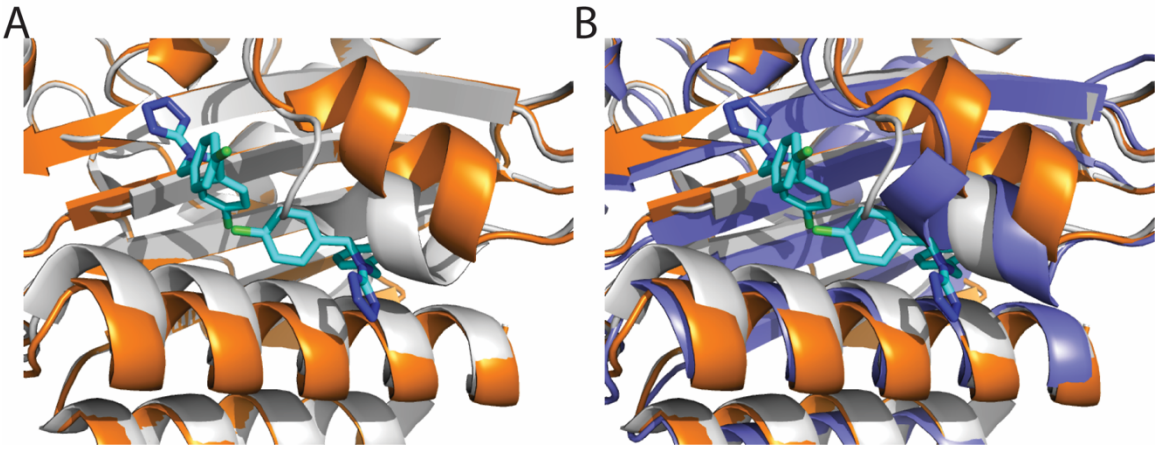

Figure 2.8: Conformational change and a snapshot from the ABMD simulation of TEM-1 β-lactamase. All structures are shown in cartoon representation. **(A)** Unbound structure 1jwp (grey), superimposed with the bound structure 1pzo (orange). The two allosteric inhibitors bound to 1pzo are shown in cyan. **(B)** Helix H11 from a snapshot at t = 20 ns of the ABMD simulations of TEM-1 β-lactamase with k = 30.0 kcal/mol/Å^2 (blue). H11 partially unfolds to open a small but druggable pocket for the binding of ligands.

Group 3 includes 17 more proteins with cryptic sites that seem to form in some structures due to mutation, binding of ligands or proteins at locations distant from the cryptic site, or simply due to changes in the conditions of crystallization. The next paragraphs will discuss why forming the cryptic site depends on such additional factors for four proteins. For clarity reasons, the following paragraphs will be numbered.

(1) The first example is AMPc beta-lactamase with a mechanism of cryptic site opening that is similar to that of TEM-1 beta-lactamase, although the two proteins exhibit limited sequence or structure similarity. In many unbound structures of the AMPc beta-lactamase residues 289-293 form a small helix protruding into the site. In the presence of fragment-sized inhibitors the same residues form a loop allowing for ligand binding (PDB ID 3gqz). Although the active site is more than 8 Å from the allosteric site, the two sites are in the same crevice, and binding of active site inhibitors seems to affect the opening of the allosteric site, which can also be impacted by mutations (Figure 2.9A).

(2) The second protein in this group is human pyruvate dehydrogenase kinase, which has a non-competitive (allosteric) inhibitor site, 33 Å from the ADP binding site (PDB ID 2bu2). Upon binding by the inhibitor TF1, the helix alpha-2 shifts by a hinge motion. The loop of residues 34-37 is found to be very flexible in all structures determined to date. This may be necessary to facilitate the hinge movement of the helix. In spite of the large distance, the opening of the cryptic site is clearly affected by binding at the ADP site, since DS<0.3 in all ADP-bound structures but DS>0.7 in all structures with bound ADP-competitive inhibitors, and thus the binding of the inhibitors helps to open the allosteric site (Figure 2.9B).

(3) At its cryptic site exodeoxyribonuclease I (ExoI) binds BCBP (PDB ID 3hl8), which inhibits its interaction with bacterial single-stranded DNA-binding proteins. In many unbound structures W245 protrudes into the weak surface site. The pocket is generally not well formed, but there are a few exceptions. Almost all structures are co-crystallized with various oligonucleotides, and such interactions affect the cryptic site, but the highest DS value occurs in a ligand-free structure (Figure 2.9D).

(4) The cryptic site in the Dengue 2 virus envelope protein is located between two domains and binds the detergent n-octyl-β-D-glucoside (PDB ID 1oke). Spontaneous variations may occur between open and closed states. The key change is the local rearrangement of the hairpin formed by residues 268–280, and the concomitant opening up of a hydrophobic pocket. The most open pockets occur in unbound structures, whereas DS is reduced by the binding of antibodies at distant sites (Figure 2.9E), motivating the placement of the protein in this category.

The other 13 proteins in this group are hepatitis C virus polymerase NS5B (Figure 2.9C), myosin (Figure 2.9F), monomeric actin, fructose 1,6-bisphosphatase, maltodextrin/maltose binding protein, the MurA dead-end complex, acid-beta-glucosidase, biotin carboxylase, glutamate receptor 2, androgen receptor, p38 map kinase, and aspartate transcarbamylase. Details for these proteins are given in the supplemental material of our publication (Sun et al., 2019).

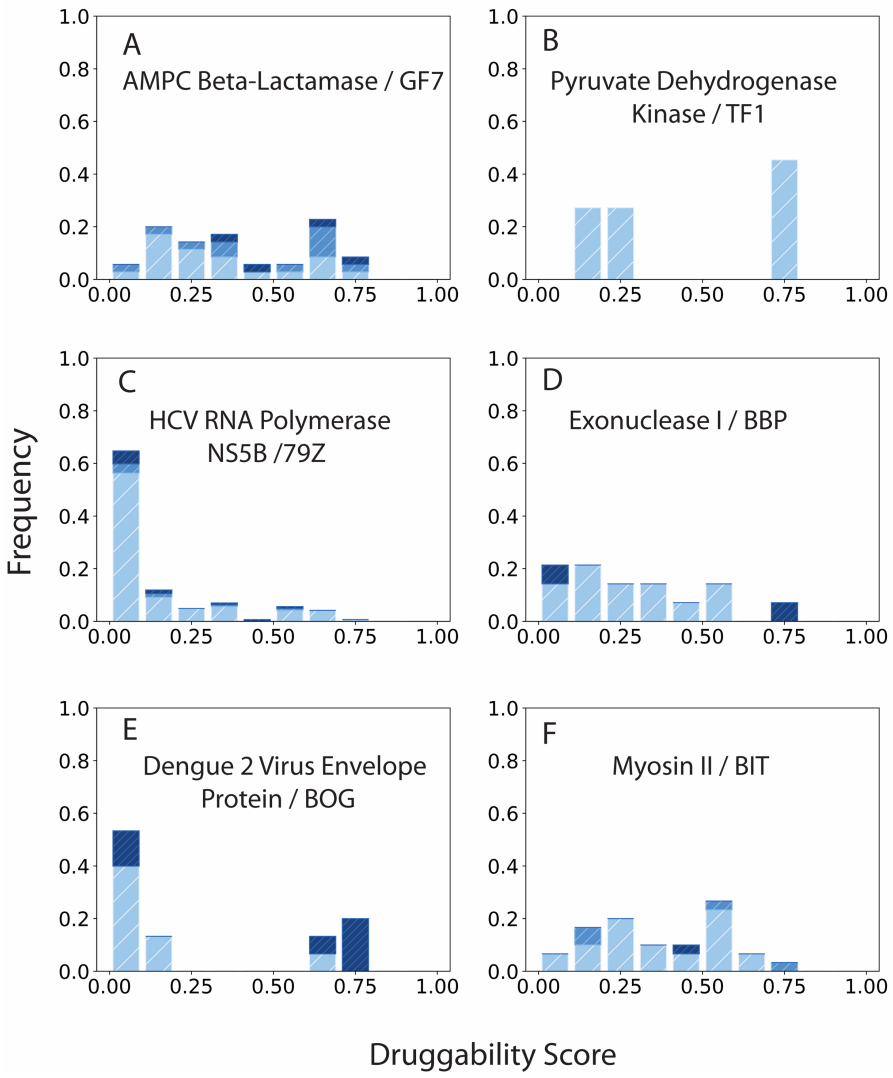

Figure 2.9: Druggability scores (DSs) of unliganded structures of proteins with cryptic sites impacted by mutations or binding at distant sites. The ligand bound at the cryptic site shown in parenthesis. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound

structures, complexes, and mutants. **(A)** AMPc beta-lactamase (Inhibitor GF7). **(B)** Human pyruvate dehydrogenase kinase (Allosteric inhibitor TF1). **(C)** Hepatitis C virus RNA polymerase NS5B (Inhibitor 79Z binding near the active site). **(D)** Exodeoxyribonuclease I (Inhibitor BCBP). **(E)** Dengue 2 virus envelope protein (Detergent n-octyl-β-D-glucoside). **(F)** Myosin II (inhibitor blebbistatin).

2.4 Discussions and Conclusions

The binding of a ligand molecule is often accompanied by conformational changes of the protein. This is definitely the case if the binding site is cryptic, thus it is not detectable in the unliganded protein. A central question is whether the ligand induces the conformational change via induced-fit, or rather selects and stabilizes a complementary conformation from a pre-existing equilibrium of ground and excited states of the protein via conformational selection (Weikl & von Deuster, 2009). Since the binding proceeds from the free energy minimum of the separate target protein to the free energy minimum of the receptor-ligand complex, the distinction is kinetic rather than thermodynamic. However, the free energy landscape of the protein determines the pathway of the association. In fact, the unbounds state is always an ensemble of conformations (Hilser, Garcia-Moreno, Oas, Kapp, & Whitten, 2006). If conformations without the pocket formed are at deep free energy minima, then the probability of pocket formation without ligand binding is small. On the other extreme, if the landscape includes minima leading to conformations with pockets formed, then the binding site is most likely cryptic only in a certain fraction of the conformational ensemble.

Molecular dynamics (MD) is increasingly considered as a valuable tool to characterize conformational ensembles of macromolecules. One of the major strengths of this approach is that it provides both thermodynamic and kinetic information (Knoverek

et al., 2019). However, as discussed for TEM-1 β-lactamase, the results of simulations depend on a multiplicity of factors (Bowman & Geissler, 2012) (Oleinikovas et al., 2016), including the force field parameters (Childers & Daggett, 2018) and on the strategy of sampling (Zimmerman, Porter, Sun, Silva, & Bowman, 2018). In addition, each timestep is on the order of a femtosecond, while many of the biological processes of interest take a millisecond or longer. Performing over 1012 iterations is computationally expensive, and limits the applicability of the method. The use of Markov state models (MSMs) enables ultra-long MD simulations (Lane, Bowman, Beauchamp, Voelz, & Pande, 2011), and helps to elucidate functional conformational changes (Chodera & Noé, 2014) (M. Huang et al., 2018). In spite of recent development, MSMs still require substantial computational resources and have been applied only to a few proteins for the analysis of cryptic site opening (Bowman & Geissler, 2012) (Knoverek et al., 2019) (Maurer et al., 2012; J. R. Porter et al., 2019).

The main goal of this chapter was to consider unliganded X-ray structures of proteins with validated cryptic sites and to study whether the sites remain always cryptic without ligand binding, or pockets already form in some of the structures. The simple approach of documenting the druggability of pockets at cryptic sites in 32 proteins enabled us to arrive at some fairly general conclusions. First, we have shown that few proteins have even approximately “genuine” cryptic pockets that are unlikely to form without ligand binding. Second, proteins on the other extreme, with spontaneously opening and closing cryptic sites, are also rare. The most well populate group includes proteins that, under some conditions, have a cryptic pocket with very low druggability,

but easily form a more druggable pocket if the conditions change. This behavior is in good agreement with the assumptions that the native state of the protein is defined by an ensemble of conformational states at free energy minima with similar energy levels (Hilser et al., 2006). Even moderate perturbations can change the free energy landscape and thereby impact the distribution of residence probabilities at the various states, also affecting the druggability of pocket at the cryptic site. The practical implication of this finding is that in order to discover cryptic allosteric site it is always advisable to investigate all homologous proteins. As shown for TEM-1 β-lactamase, it is particularly useful to study slightly destabilized versions of a protein. The conclusions from the analysis of X-ray structures were confirmed by ABMD simulations applied to one protein from each of the three groups (Harvey & Gabb, 1993; Marchi & Ballone, 1999; Paci & Karplus, 1999).

CHAPTER 3 Pairwise Statistical Potentials for Protein-Small Molecule Interactions

3.1 Introduction

The current fragment-based drug discovery (FBDD) procedure involves two main steps (Erlanson et al., 2004): 1) development of the libraries of fragments, and 2) converting fragments into hits and leads with fragment optimization, merging, and assembly. The second step especially relies on the premise that the orientation of the fragment will be conserved once converted into a drug-like full-size ligand (Kozakov, Hall, Jehle, et al.). Computational solvent mapping provided by FTMap can potentially identify the main hot spot, generate poses of the probes in the hot spot, and extrapolate such information to the binding mode of the desired ligand. The current version of FTMap, as described in the Section 1.3, provides accurate solutions of hot spot locations and strengths through simulation of crystal soaking. However, connecting the information from probe clusters to the viability of fragments, and eventually strategizing the growth of the fragments, will require an improved version of FTMap. Through analytical studies, we were able to demonstrate that ligand fragments overlapping with energetically important hot spots conserve their binding mode when the rest of the ligand is removed (Kozakov, Hall, Jehle, et al., 2015). In a scenario where the drug discovery teams have to choose among fragments, one natural approach would be to compare the predicted poses of candidate fragments and FTMap probe cluster poses. Therefore, ideally the FTMap probes should have the chemical diversity on par with the fragment library used in screening. The current set of FTMap probes consists of only 16 small organic molecules. As briefly discussed in Section 1.3, the pairwise statistical potentials for these molecules

were derived from protein DARS potentials. One of the main reasons for taking such an *ad hoc* approach was the structural similarity between the probes (Figure 3.1) and side chains of amino acid residues, for which the parameters were readily available in the original DARS potential derived for protein-protein docking (Chuang et al., 2008).

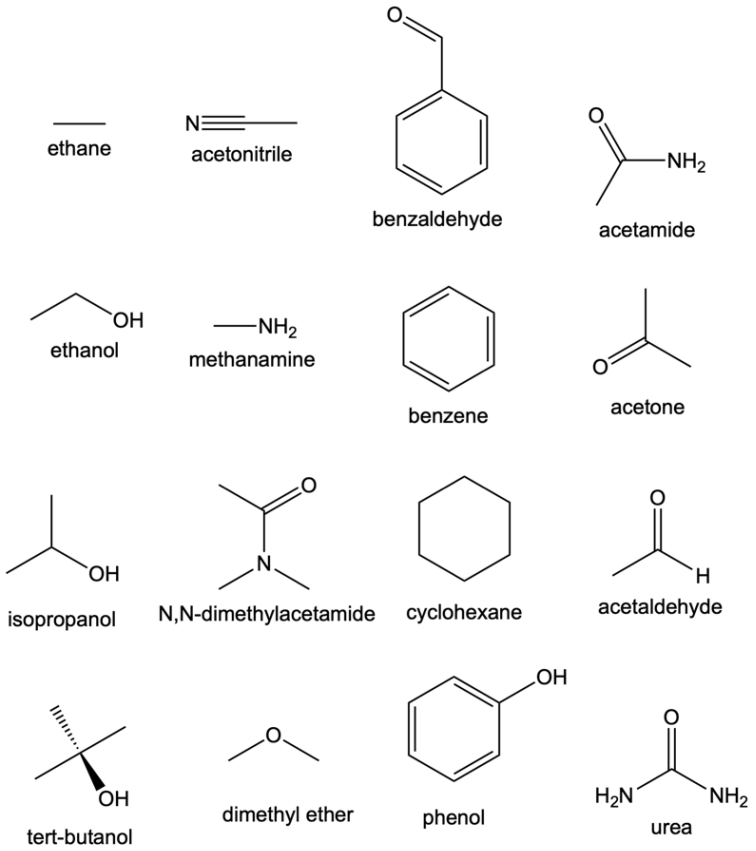

Figure 3.1: The 16 molecular probes currently used in FTMap. They bear a structural resemblance to side chains of amino acid residues (Kozakov, Grove, et al., 2015).

In addition, we have also observed that FTMap needs to improve on identifying hydrogen bonding instances. The immediate interests to 1) expand the probe set to explore more chemical diversity and 2) generate more accurate poses of the molecular probes depend on a superior energy function. More specifically, we critically need the

systematically developed pairwise potential for small molecule-protein interactions that can replace the limited *ad hoc* protein DARS assignment.

Decoys As the Reference State (DARS) is a natural framework for constructing structure and statistics-based intermolecular potentials to improve the accuracy of the energy function. The key idea is to extract interaction frequencies from a large set of docking decoys, which are essentially docked conformations considering only good shape complementary but no other criteria (Chuang et al., 2008). DARS is an innovative formalism because it already accounts for non-uniform distribution of protein surface atoms and corrects for atom type specific information in the reference set statistics. The advantages of DARS have been previously demonstrated in the improved docking quality of PIPER (Chuang et al., 2008). The application of DARS to small molecules requires careful consideration and distinctive strategies because of a few reasons. Unlike protein-protein interface, a small molecule makes limited contact at the binding site with the protein; free energy contribution per atom is high for small molecule-protein interactions. The pairwise potential should be able to reflect atom-scale considerations. On the other hand, the composition of small molecule atom types is far more diverse than amino acid atom types.

As described in Section 1.2.2, the DARS framework is essentially built on the inverse Boltzmann formula. The pairwise statistical potential between two atoms of types *i* and *j* is (Chuang et al., 2008):

$$\varepsilon_{i,j} = -\mathrm{RT} ln(\frac{P_{i,j}^{nat}}{P_{i,j}^{ref}})$$

In this equation, $P_{i,j}^{nat}$ is the probability of atoms of types *i* and *j* making contact in the native structures (i.e. crystal structures of complexes). To specify, $P_{i,j}^{nat}$ is calculated as the following (Chuang et al., 2008):

$$P_{ij}^{nat} = \frac{v_{ij}^{obs}}{\sum_{i,j} v_{ij}^{obs}}$$

where v_{ij}^{obs} is the number of observed contacts between atom types *i* and *j* within the defined pairwise interaction distance, and the denominator is the total number of atom pairs in the native structures. Likewise, $P_{i,j}^{ref}$ is the probability of atoms of types *i* and *j* to exist within the defined interaction distance in the reference set. The reference state should represent the "perfectly" random protein-ligand complexes (Chuang et al., 2008). Based on our success with the protein DARS potential, we decided to explore the same approach of using docking decoys to develop protein-small molecule DARS. Extracting valid protein-small ligand complexes from the Protein Data Bank (PDB) is not a straightforward task. Instead, we used the curated database PDBbind (Liu et al., 2015) as the training set for gathering native interaction statistics as well as generating decoys for the reference set. The PDBbind database contains structural data of chemically diverse protein-small molecule complexes as well as energetics data. The binding data in the PDBbind assured that non-specific protein-small molecule interactions are excluded from our training set. In addition, PDBbind also pre-assigns protonation states of ligand groups, in contrast to the ambiguity with valence and protonation states of small molecules from the PDB raw data. The ligands from PDBbind are also stored in Tripos MOL2 (Clark & Cramer, 1989) format, which shows all hydrogens explicitly and already

contains information for types of heavy atoms.

This chapter addresses the first step for improving FTMap application in FBDD. The primary goal of the work is to establish that systematically-derived DARS potentials from validated protein-small molecule complexes perform equally as well (or better than) the current implementation of *ad hoc* potential originated from protein DARS (i.e. *ad hoc* DARS parameters for the small molecular probes were assigned based on their similarity to atoms in amino acid side chains). We took a similar approach of collecting structural data for developing the pairwise potential, with some extra effort specific to the nature of generating docking decoys for small molecules. As will be described in the next section, only the ligand moiety overlapping with mapping hot spots are retained for collecting interaction statistics. This strategy was taken to avoid docking multiple conformations of large ligands with rotatable bonds, and was supported by our observation that ligand fragment coinciding with hot spots are the most energetically important atoms interacting with the protein (Kozakov, Hall, Jehle, et al.). Also, a large number of decoy poses (35000 lowest energy poses) had to be retained for the reference state statistics, which facilitated a diverse distribution of decoys on the protein surface. The small molecule-protein DARS potential generated using a 3.5 Å cutoff was tested on a set of 48 apo structures that were previously used for the evaluation of other binding site prediction methods (B. Huang & Schroeder, 2006). The results presented in this chapter show that the systematically-derived small molecule-protein pairwise DARS potential perform reasonably well compared to the current *ad hoc* DARS, but continued improvement of as well as exploration of robust evaluation metrics are necessary.

3.2 Methods

3.2.1 Collection of Structural Data for Developing the DARS Potential

3.2.1.1 PDBbind "Refined" Set

The "refined" set of PDBbind database was compiled to select high-quality protein-ligand complexes appropriate for evaluating and developing scoring functions (Liu et al., 2015). The selection criteria of the PDBbind refined set is fairly complicated and has evolved over the years; basically, the cases have to suffice conditions that ensure both high quality structural as well as binding data. The biological and chemical nature of the complexes were also checked (Liu et al., 2015). At the time of the data download in 2014, 3706 complexes were included in the set. The PDBbind team also prepared the data in such a way that the protein and ligand molecules were saved separately in standard formats. We saved protein structures as PDB files and ligands as MOL2 files. The protein PDB format is suitable for the next step of solvent mapping. The ligand MOL2 files contain hydrogen atoms explicitly plus the pre-assigned Tripos MOL2 atom types, which facilitates the preparation of ligands for generating of decoys as well as atom typing. The cases were also standardized by conversion to PDBQT format (Morris et al., 2009), which identified rings and substituents in the ligand structures.

3.2.1.2 Solvent Mapping of Proteins

Before generating decoys, we mapped all proteins in the refined PDBbind set (case PDB ID 4bps was removed because of the non-standard amino acid selenomethionine). The hot spots from mapping were used to compare to the ligand pose

in the crystal structure. Only the fragment part of the ligand overlapping with the hot spots were retained for docking and generating decoys. We had the following reasons to proceed with such strategy: 1) from previous studies we determined that the most important part of the ligand interacting with the protein are those atoms that coincide with the hot spots; 2) docking only fragments to generate decoys avoided too many rotatable bonds and multiple conformations of one ligand. The 3705 proteins in the "refined" PDBbind set were mapped using the commercial version of FTMap, Atlas (Brenke et al., 2009) (Kozakov, Grove, et al., 2015), provided by Acpharis. Atlas is free for academic research. For high-throughput purposes, we mapped all proteins with default settings of Atlas. Usually, we recommend mapping multi-domain proteins or structures with huge cavities using domain splitting or focused mapping to prevent probes from falling into one place. However, a large number of decoys (as will be discussed in the Section 3.2.1.4) were generated for the reference state, which should create sufficiently diverse decoy statistics.

3.2.1.3 Carving Fragments

The goal of solvent mapping was to identify the minimally flexible part of the ligands in the original PDBbind complexes. Dr. Dmitri Beglov developed a program which extracted chemically consistent fragments with minimal flexibility (the program used the information from PDBQT files to identify rings and substituents) (Morris et al., 2009). The output fragment is the portion of the ligand overlapping with the top hot spot from solvent mapping of its partner protein. During this process, we eliminated cases

where the ligand was a polymer or contained repetitions of the same atoms, including peptides. If one subunit of the polymer overlaps with the top hot spot while the other parts do not, it would be challenging to justify one subunit is more important for binding to the protein than another subunit. We also expected peptide-like ligands to exhibit a different structural pairwise potential profile than the standard small molecules. In addition, cases with no hot spots nearby the ligands were also discarded. In the end, 2907 fragments were available for generating the reference state statistics.

3.2.1.4 Generation of Decoys

We relied on the software Atlas Parameterization to process the fragments before docking. This software is built on Open Babel (O'Boyle et al., 2011) as well as Amber Tools (Case et al., 2005) to assign parameters such as charges and spring constants, using the AM1-BCC charge model (Jakalian, Jack, & Bayly, 2002) and certain parameters from the general AMBER force field (GAFF) (J. Wang, Wolf, Caldwell, Kollman, & Case, 2004). The number of cases reduced to 2522 after parameterization. Some common reasons for parameterization failures were due to flat structures, odd numbers of electrons, presence of phosphate groups, etc. To generate decoys, the fragments were docked back to their respective partner proteins using PIPER (unlike protein docking, special consideration was given to small molecules so that all of the atoms were recognized as solvent-accessible by the program). The docking was done with only van der Waals components in the scoring function, the same way the protein decoys were generated before (Chuang et al., 2008). 2499 fragments were successfully docked in the

end.

For each system, the 35000 decoy poses with lowest PIPER energies were retained for the reference state. This large number of poses was needed to generate diverse decoy statistics (including a smaller number of decoys resulted in the collection of poses in very concentrated places on the protein surface). These 35000 decoys were subject to greedy clustering using a 9 Å radius. Clustering reduced the redundancy in decoy statistics and produced a reasonable number of cluster centers for collection of interaction statistics. Figure 3.2 summarizes the steps taken so far.

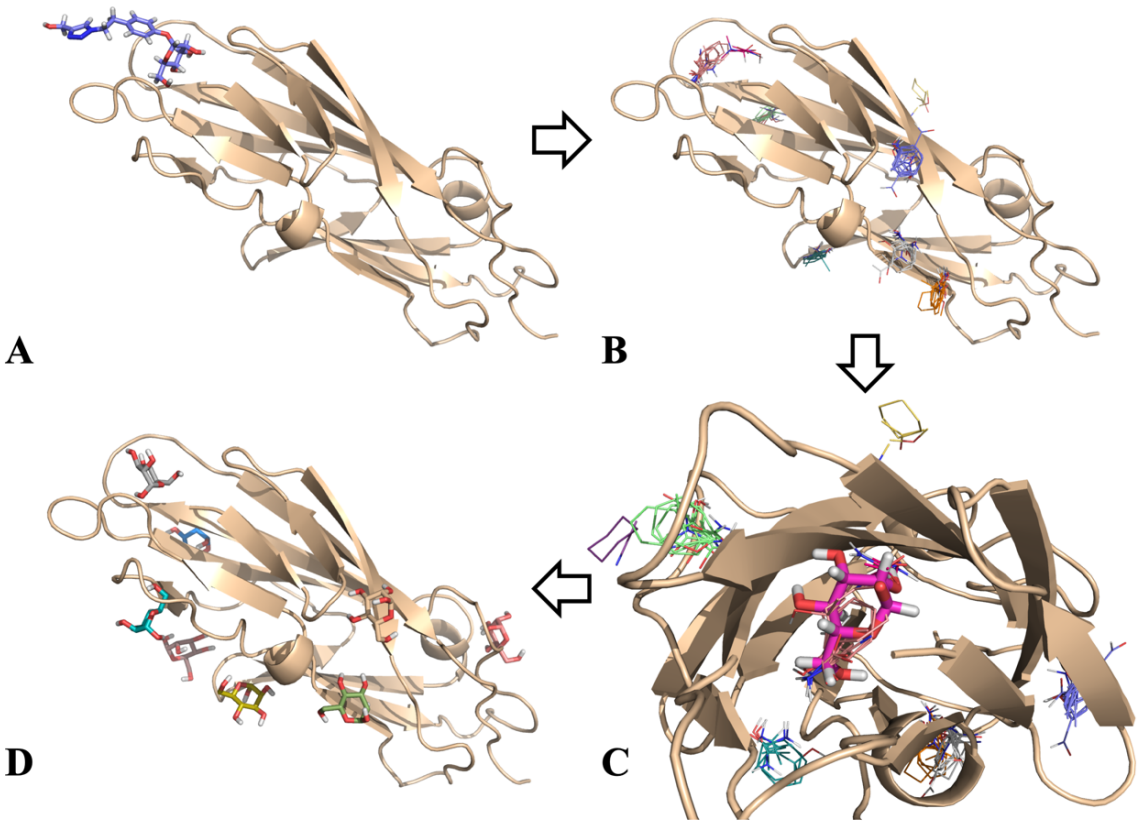

Figure 3.2: Collection of structural data from the training set. **(A)** The original PDBbind complex contains a protein and a small molecule ligand. **(B)** The unbound protein was mapped for hot spots. **(C)** The part of the ligand overlapping with the main hot spot was carved out as a fragment. **(D)** The fragment was docked back to the protein to generate decoy cluster centers.

3.2.2 Construction of the Protein-Small Molecule DARS Potential

3.2.2.1 Atom Types

Defining atom types for heavy atoms in small molecules is not a trivial task. There is a variety of classification schemes from general atom typing such as AMBER GAFF (J. Wang, Wang, Kollman, & Case, 2006), Tripos MOL2 (Clark & Cramer, 1989), and AutoDock4 (Huey, Morris, Olson, & Goodsell, 2007) atom types, to those specifically chosen for statistical potential derivations such as X-SCORE (R. Wang, Lai, & Wang, 2002) and PMF (Muegge, 2006). After preliminary experimentation we decided to try the Tripos MOL2 system. The MOL2 atom types consist of more than 40 heavy atom types, including rare elements such as aluminum. The Tripos MOL2 format should be able to accommodate the chemical diversity within the PDBbind set. On the protein side, we decided to keep the current DARS atom types used in ClusPro, which include 18 atom types originally defined for Atomic Contact Potential (ACP) (C. Zhang, Vasmatzis, Cornette, & DeLisi, 1997).

3.2.2.2 Counting Pairwise Interactions

To accommodate different types of interactions between small molecules and proteins, we initially decided to include three types of cutoff distances for evaluating interacting atoms: [0, 3.5 Å], [0, 6.5 Å] and [3.5 Å, 6.5 Å]. The hope is that the [0, 3.5 Å] cutoff will capture hydrogen bonds; [0, 6.5 Å] is the same cutoff as the protein DARS potential, which should incorporate hydrophobic interactions. The [3.5 Å, 6.5 Å] window was used to collect statistics because of our interest in developing a multi-bin DARS

potential in the future, which is beyond the scope of this chapter. The protein-fragment pairs from the training set PDBbind was used for generating v_{ij}^{obs}, the number of observed contacts between atom types *i* and *j* within the defined pairwise interaction distance in the native structures, and subsequently P_{ij}^{nat} as defined in Section 3.1. As for the decoy statistics, each protein-fragment system contained multiple decoy cluster centers. For each system, the number v_{ij}^{ref} was therefore normalized, meaning the total counts of interacting atoms of types *i* and *j* across all decoy cluster centers were divided by the number of such cluster centers. The corresponding P_{ij}^{ref} was the probability of normalized atom pair counts.

3.2.2.3 Calculation of Pairwise Potentials

Given the "step zero" for improving FTMap scoring function is to demonstrate that systematically-derived DARS performs equally as well as or better than the current *ad hoc* DARS, the work in this chapter focuses on the pairwise potential relevant for the 16 solvent probes. After close examination of atom types occurring in the 16 probes, a subset (more specifically, nine Tripos MOL2 types) of the previously mentioned small molecule atom types were retained for constructing the pairwise potential. All heavy atoms in the 16 probes were accounted for. The probe methylamine initially had the nitrogen atom assigned as sp^3 nitrogen (MOL2 type "N.3"), but was manually converted to positively charged sp^3 nitrogen (MOL2 type "N.4"). All primary amines in PDBbind were protonated and carried positive charge, and hence they were labeled as "N.4"

instead of "N.3". This is biologically relevant, as primary amine should be fully protonated at pH=7. Using the probabilities of interactions (P_{ij}^{nat} and P_{ij}^{ref}) and the inverse Boltzmann formula discussed in Section 3.1, an 18×9 matrix M_{asym} of pairwise potential $\varepsilon_{i,j}$ was generated. There were a number of atom pairs present in the reference set but not in the crystal structures. We interpreted these interactions as favorable in decoys but not favorable in native structures. Therefore, a large positive number was assigned as dummy values for $\varepsilon_{i,j}$, in order to penalize such atom pairs coming to close proximity in FFT sampling. We constructed the potentials for both [0, 3.5 Å] and [0, 6.5 Å] windows. The top priority was to evaluate if this new DARS potential could capture hydrogen bonding behavior. Consequently, the potential presented in this chapter has interaction range of 0-3.5 Å.

3.2.3 Incorporating DARS into FTMap

3.2.3.1 Eigenvalue–Eigenvector Decomposition of the Matrix

In order to evaluate the pairwise potential using FFT, the energy expression must be written as a sum of correlation functions. Based on the eigenvalue-eigenvector decomposition of the pairwise potential matrix representing $\varepsilon_{i,j}$, it can be written as:

$$\varepsilon_{i,j} = \sum_{p=1}^{K} \lambda_p u_{pi} u_{pj}$$

The terms λ_p is the *p*-th eigenvalue of the interaction matrix, and u_{pi} is the *i*-th component of the *p*-th eigenvector. The terms in the decomposition represent energy

contribution proportional to the absolute value of the eigenvalue λ_p (Brenke et al., 2009; Kozakov et al., 2006). It is not possible to perform eigen decomposition of the asymmetric 18×9 matrix M_{asym} for protein-small molecule interactions. However, we were able to create the 27×27 symmetric matrix M_{sym} equivalent to M_{asym}, a recipe we followed when developing our asymmetric antibody-antigen DARS (Brenke et al., 2012) (see Figure 3.2). Using MATLAB (version R2019a), it was straightforward to calculate the eigenvectors and eigenvalues associated with the new matrix M_{sym}. In preparation for the next steps, the eigenvectors were organized based on the magnitude of their respective eigenvalues.

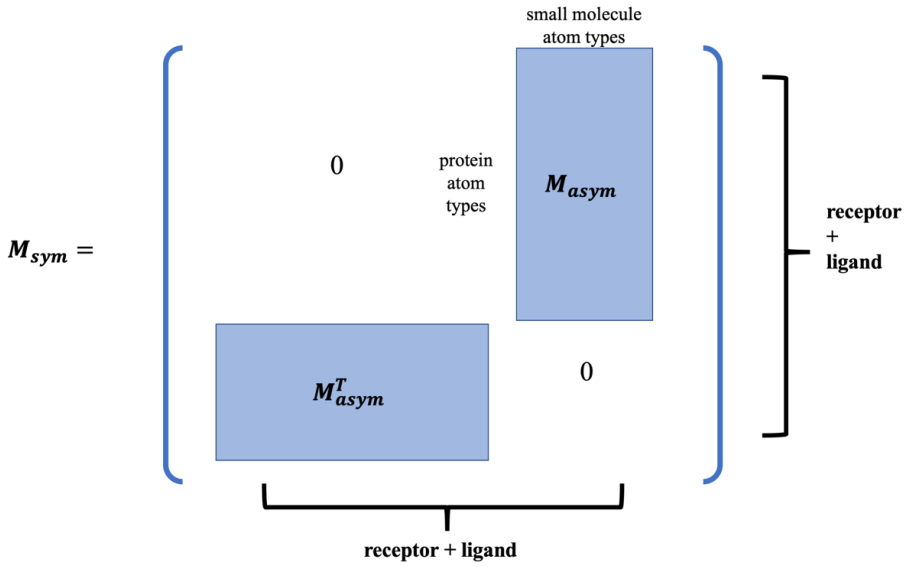

Figure 3.3: Conversion from the asymmetric protein-small molecule pairwise potential matrix to the equivalent symmetric matrix. The labels on the matrix indicate that in atom type naming for our FFT program, protein atom types took on type number 0-17, followed by small molecule atom types of 18-26.

3.2.3.2 Implementation of DARS into Scoring Function

Based on our previous experience, we only needed to use a few dominant eigenvectors to approximate the interaction matrix in FFT sampling (Kozakov et al.,

2006). With protein DARS, it was possible to yield only 10% error in energy values with only the first four eigenvectors (Kozakov et al., 2006). Since the magnitude of the eigenvalues from deconstructing M_{sym} varied widely, we chose the top ten eigenvectors all with absolute values higher than three. This number was chosen to maintain the balance between accuracy in the energy function and the computational expense (sampling could become too demanding if too many eigenvalues were requested during FFT). The neglected eigenvectors were associated with much smaller eigenvalues and we hoped it was sufficient for initial testing. It is desirable to revisit rigorous testing by systematically "zero-ing out" the smallest eigenvectors for evaluation in the future. The current FTMap was also programmed to use 6.5 Å as the radius for deriving contacts in calculating pairwise interactions. This cutoff distance was changed to 3.5 Å in consistency with the new version of DARS. The atoms in the small probes were assigned the new PIPER atom types according to Figure 3.3. All changes were implemented in a local version of FTMap running in a Docker container. The workflow of FTMap solvent mapping consists of the following steps: 1) FFT rigid body sampling docks the 16 probe molecules to the apo protein. 2) The 2000 best probe poses retained from the previous step are minimized using CHARMM potential (Brooks et al., 1983). 3) The minimized probes are clustered and clusters are ranked based on Boltzmann averaged energies. For each probe six clusters with the lowest energies are retained. 4) Hot spots, or consensus sites (CSs), are generated by clustering different probe clusters using the distance between center of mass of probe cluster centers as the distance measure. These hot spots are finally ranked based on the numbers of probe clusters they each contain, which we

also refer to as the "strength" of the hot spots (Brenke et al., 2009; Kozakov, Grove, et al., 2015).

All program settings and parameters except those discussed in the previous paragraph were left unchanged, including the PIPER energy weights for van der Waals, electrostatic and pairwise potential terms. Recall that the goal of this work is to first prove the new DARS is equivalent or superior to the *ad hoc* DARS potential. We presumed that the weights assigned to energy terms other than the pairwise potential were sufficiently optimized to push probes into the generally correct site, while the DARS potential further refined the sampling by pulling desirable atom pairs closer and simultaneously penalizing the undesirable pairs. In other words, at this point we assumed the current values of weights were acceptable for initial testing. We plan to further optimize the scoring function with benchmark studies in the future.

3.2.4 Testing Solvent Mapping with the New DARS

Previously we used the LIGSITE^{CSC} (B. Huang & Schroeder, 2006) set to evaluate binding site prediction of FTSite (Ngan et al., 2012), a program related to FTMap. The LIGSITE^{CSC} contains 48 proteins with both apo and holo structures in PDB format, including validated small molecule ligands. This data set is therefore ideal for the initial testing of DARS performance. Each apo protein from the LIGSITE^{CSC} set was mapped twice: one run with the regular FTMap and one run with the local version of FTMap where the new small molecule DARS was implemented. Prior to solvent mapping, the LIGSITE^{CSC} cases were visually inspected to check if all chains were

needed in mapping; for several cases, only certain chain(s) were mapped. But the mapping conditions were kept consistent between regular FTMap and mapping with the new DARS potential.

3.2.5 Evaluation of Mapping Results

In addition to visual inspection of the hot spots, we needed a quantitative metric to evaluate the performance of mapping. Previously we formulated a measure of the degree of spatial overlap between a molecule and hot spots called fractional overlap (Kozakov, Hall, Jehle, et al., 2015). To use such formulation, all mapping results were superimposed onto the holo structure, where the overlap between hot spots and the co-crystalized ligand was calculated. Dr. David Hall developed a Python program using Pybel (O'Boyle, Morley, & Hutchison, 2008) that loops through all atoms in the ligand and the hot spots, and marks any atoms that are within 2 Å as "nearby". One of the outputs reported by the program is how much of the ligand is covered by hot spots (i.e. percentage of the ligand atoms overlapping with the hot spots). This fractional overlap evaluates how well the mapping algorithm can push solvent probes into the location of the ligand, which is highly relevant to the research aim of this chapter. It was also important to check if the new DARS could push dominant hot spots into the ligand binding site. In previous publication we reported the hot spot strength could indicate druggability (the number of probe clusters in the hot spot higher than 16 points to a druggable site) (Kozakov et al., 2011; Kozakov, Hall, Napoleon, et al., 2015). However, this criterion was somewhat irrelevant since the analytical work was done based on the

previous scoring function with the *ad hoc* DARS. To reduce arbitrariness in defining "dominant" hot spots, top five hot spots were selected. The reported results therefore contain fractional overlap calculated using 1) all hot spots and 2) the top five hot spots ranked by strength (number of probe clusters).

3.3 Results and Discussions

3.3.1 The New Small Molecular Atom Types

As described above in Section 3.2, we kept the same 18 atom types for the protein amino acid residues. For small molecules, about 20 atom types were gathered from the training set of 2499 protein-fragment complexes. The atom types included halogens (Br, Cl, and F), carbon, nitrogen, oxygen and sulfur with various hybridizations and properties (sp^3, sp^2, *sp*, aromatic, cationic, etc.). Our current 16 solvent probes were only populated with nine Tripos MOL2 types in terms of heavy atoms (see Table 3.1). Note that the nitrogen atom in methylamine was perceived as positively charged, consistent with the protonated primary amines in the original PDBbind ligands. All hydrogen atoms were treated as "type-less", the same as in protein DARS assignment.

Table 3.1: The set of nine Tripos MOL2 (Clark & Cramer, 1989) atom types present in the 16 FTMap small molecule probes and their original definitions.

Atom Type Code	Definition
C.1	*sp* carbon
C.2	sp^2 carbon
C.3	sp^3 carbon
C.ar	aromatic carbon
N.1	*sp* nitrogen
N.4	sp^3 nitrogen positively charged
N.am	nitrogen amide

O.2	*sp*^{2} carbon
O.3	*sp3* carbon

The DARS potential used in current version of FTMap assigns atom types to the 16 probes using the 18 atom types described in the atomic desolvation energies publication (C. Zhang et al., 1997), also known as the atomic contact potential (ACP). The assignment was done manually, based on the resemblance between the probe atoms and amino acid side chains. For example, the carbon atoms in the probe phenol were assigned as "YC$^\gamma$" and "FC$^\gamma$", which were originally designated for carbons in side chains of tyrosine and phenylalanine. In total, excluding hydrogens, 11 ACP atom types are currently present in the FTMap probes.

Table 3.2: The set of 11 ACP atom types present in the 16 FTMap small molecule probes and their original definitions (C. Zhang et al., 1997).

	Definition	
Atom Type Code	**Amino Acid**	**PDB Atom Name**
N	Backbone	N
C^α	Backbone	C^α
C	Backbone	C
O	Backbone	O
C^β	Ala	C^β
	Arg	C^β
	Asn	C^β
	Asp	C^β
	Cys	C^β
	Gln	C^β
	Glu	C^β
	His	C^β
	Ile	C^β
	Leu	C^β
	Lys	C^β
	Met	C^β
	Phe	C^β
	Pro	C^β, C^γ, C^δ
	Thr	C^β

	Trp	C^{β}
	Tyr	C^{β}
	Val	C^{β}
RN^{η}	Arg	C^{ζ}, $N^{\eta 1}$, $N^{\eta 2}$
NN^{δ}	Asn	C^{γ}, $O^{\delta 1}$, $N^{\delta 2}$
	Gln	C^{δ}, $O^{\varepsilon 1}$, $N^{\varepsilon 2}$
SO^{γ}	Ser	C^{β}, O^{γ}
	Thr	$O^{\gamma 1}$
	Tyr	O^{η}
YC^{ζ}	Tyr	$C^{\varepsilon 1}$, $C^{\varepsilon 2}$, C^{ζ}
FC^{ζ}	Arg	C^{γ}
	Gln	C^{γ}
	Glu	C^{γ}
	Ile	$C^{\gamma 1}$
	Leu	C^{γ}
	Lys	C^{γ}
	Met	C^{γ}, S^{δ}
	Phe	C^{γ}, $C^{\delta 1}$, $C^{\delta 2}$, $C^{\varepsilon 1}$, $C^{\varepsilon 2}$, C^{ζ}
	Thr	$C^{\gamma 2}$
	Trp	C^{γ}, $C^{\delta 1}$, $C^{\delta 2}$, $C^{\varepsilon 2}$, $C^{\varepsilon 3}$, $C^{\zeta 2}$, $C^{\zeta 3}$, $C^{\eta 2}$
	Tyr	C^{γ}, $C^{\delta 1}$, $C^{\delta 2}$
LC^{δ}	Ile	$C^{\gamma 2}$, C^{δ}
	Leu	$C^{\delta 1}$, $C^{\delta 2}$
	Met	C^{ε}
	Val	$C^{\gamma 1}$, $C^{\gamma 2}$

At a first glance, it was hard to compare which atom type assignment gave a more diverse range of atom types for the 16 probes, especially since with ACP types one atom type could point to multiple elements (i.e., RN^{η}, NN^{δ} and SO^{γ} point to carbon, oxygen and/or nitrogen). After examining specific atom type assignment to each probe, we were able to account for "hybrid" atom types (i.e. one atom type covering multiple elements) and showed that the *ad hoc* DARS atom types actually classified the chemistry of the 16 probes to more specific groups (Figure 3.4). The Tripos MOL2 atom types are based on the chemistry of the atoms and have more built-in physical significance, but this system

was designed for a more general purpose (Tripos is a molecular mechanics force field) (Clark & Cramer, 1989). On the other hand, ACP types were constructed for estimating effective atomic contact energies and specifically desolvation (C. Zhang et al., 1997). The classification of ACP atoms is somewhat intuitive and is generally based on chemical properties. Nevertheless, the assignment of ACP atom types in FTMap was at most a rough estimate by comparing solvent atoms to amino acid side chains. In the future we may need to revisit a more diverse set of atom types if necessary.

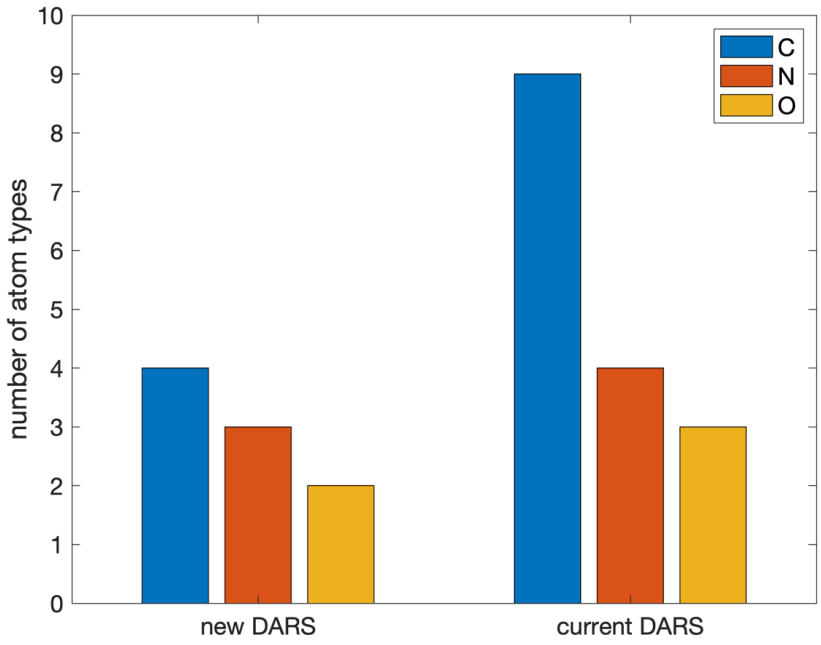

Figure 3.4: Number of atom types assigned to elements carbon (C), nitrogen (N) and oxygen (O) in the 16 FTMap probes, using the Tripos MOL2 system (new DARS) and the ACP atom types (current DARS).

3.3.2 The Small Molecule DARS Interaction Energies

In Table 3.3, the contact energies $\varepsilon_{i,j}$ between the 18 protein atom types (listed in the first column) and nine small molecule atom types (first row) are reported. The protein

atom types are defined in detail in the original ACP article (C. Zhang et al., 1997). Essentially the backbone atoms were each listed in a separate type; side chain atoms were further categorized, with the most hydrophobic side chain atoms included in categories FC^{ζ} and LC^{δ}. Detailed descriptions of nine small molecule atom types are included in Table 3.1.

Table 3.3: Pairwise contact energies $\boldsymbol{\varepsilon_{i,j}}$ of small molecule DARS in kcal/mol. The asymmetric matrix shows the interaction between 18 protein atom types (first column) and nine small molecule atom types (first row). Favorable interactions are highlighted with light orange color.

	C.1[†]	C.2	C.3	C.ar	N.1	N.4	N.am	O.2	O.3
N[*]	-0.04	0.19	1.21	0.65	-0.57	1.06	0.40	-0.79	-0.64
C^{α}	10.00[‡]	1.34	1.80	1.19	10.00	10.00	1.88	-0.48	0.03
C	-0.12	1.35	1.44	0.79	-0.70	0.57	1.03	0.27	-0.38
O	0.44	0.51	0.26	0.34	0.01	-0.51	-0.41	0.59	0.24
GC^{α}	0.05	0.67	0.97	0.70	-0.03	0.66	0.87	-0.96	-0.86
C^{β}	0.60	0.95	1.37	0.69	-0.44	1.19	0.46	-0.09	-0.15
KN^{ζ}	10.00	0.50	1.33	1.27	0.33	10.00	0.50	0.04	-0.57
KC^{δ}	10.00	10.00	1.95	1.84	10.00	10.00	10.00	0.49	0.45
DO^{δ}	10.00	0.06	-0.53	0.68	10.00	-1.61	-0.56	0.13	-1.45
RN^{η}	10.00	0.35	0.88	0.96	-0.56	10.00	1.04	-0.01	-0.50
NN^{δ}	0.91	0.62	0.39	0.64	-0.59	-0.20	0.28	-0.31	-0.66
RN^{ε}	10.00	0.80	1.59	2.01	-0.12	10.00	1.44	0.63	0.24
SO^{γ}	0.48	-0.24	0.05	0.16	-0.59	-0.77	-0.67	-0.40	-0.32
HN^{ε}	10.00	0.07	0.06	0.18	10.00	0.10	-1.47	-1.13	-1.49
YC^{ζ}	10.00	0.03	0.32	0.10	-0.32	-0.57	0.40	-0.77	-0.66
FC^{ζ}	1.02	0.17	0.40	0.31	-0.90	-0.03	0.02	-0.06	-0.30
LC^{δ}	-0.27	0.41	0.53	0.10	-0.32	0.95	0.62	-0.42	-0.13
CS^{γ}	-1.37	-0.07	0.33	0.36	10.00	-0.94	-0.35	-0.82	0.14

[*]Protein atom types are described in further details in the ACP article (C. Zhang et al., 1997). In general, a side chain atom type is a combination of the single letter amino acid code and one of the atoms of the type (e.g. $SO^{\gamma} = O^{\gamma}$ of serine).
[†]Small molecule atom types are defined in the Tripos article (Clark & Cramer, 1989). The code is composed of element and chemical property. For example, C.3 = carbon sp^3, C.ar = aromatic nitrogen, N.am = amide nitrogen, etc.
[‡]Dummy values.

It is interesting to see that both "C.1" and "N.1" seem to strongly repel many types of protein atoms. Note that a large positive number was assigned to the contact energy $\varepsilon_{i,j}$ if the interaction was only observed in the decoys but not in the native structures. It is possible that nitrile or alkyne groups simply rarely occurred in PDBbind. The protonated amine nitrogen "N.4" was also found to form strongly repulsive interaction with positively charged arginine, lysine and histidine side chains within the defined distance (< 3.5 Å). A large number of attractive interactions were observed in the columns for amide nitrogen and oxygens, pointing to the possibility of hydrogen-bonding. The columns for hydrophobic atoms such as sp^3 and aromatic carbons contained lots of positive values, indicating that within the 3.5 Å cutoff, hydrophobic interactions could be penalized in FFT sampling. One surprise is the $\varepsilon_{i,j}$ in the "N.1" column, where a lot of moderately attractive interactions were observed.

As described in Section 3.2.3.1, we performed eigen decomposition of the symmetric matrix built from Table 3.3. Table 3.4 shows the top ten eigenvalues and their corresponding eigenvectors, ordered by the magnitude of the eigenvalues. Note there is a wide range of absolute values of these eigenvalues. Based on visual inspection of the eigenvalues, the top ten dominant eigenvalues and eigenvectors were included in the FFT sampling step.

Table 3.4: Top ten eigenvalues and eigenvectors for the new DARS potentials, rank-ordered by the magnitude of eigenvalues.

Atom	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
types	-38.11	38.11	-18.65	18.65	-13.97	13.97	-8.41	8.41	-3.29	3.29
N	-0.01	-0.01	-0.04	-0.04	0.03	0.03	-0.02	-0.02	-0.34	-0.34
C^{α}	-0.32	-0.32	0.07	0.07	0.05	0.05	-0.32	-0.32	-0.02	-0.02
C	-0.01	-0.01	-0.04	-0.04	0.07	0.07	0.08	0.08	-0.28	-0.28
O	0.00	0.00	0.01	0.01	-0.01	-0.01	0.04	0.04	-0.01	-0.01
GC^{α}	-0.01	-0.01	-0.02	-0.02	0.05	0.05	0.03	0.03	-0.34	-0.34
C^{β}	-0.02	-0.02	-0.05	-0.05	0.04	0.04	0.02	0.02	-0.26	-0.26
KN^{ζ}	-0.23	-0.23	-0.23	-0.23	-0.12	-0.12	-0.14	-0.14	-0.03	-0.03
KC^{δ}	-0.37	-0.37	0.04	0.04	0.49	0.49	0.32	0.32	0.10	0.10
DO^{δ}	-0.19	-0.19	0.33	0.33	-0.23	-0.23	0.12	0.12	0.07	0.07
RN^{η}	-0.23	-0.23	-0.26	-0.26	-0.12	-0.12	-0.10	-0.10	0.07	0.07
NN^{δ}	-0.01	-0.01	-0.02	-0.02	-0.01	-0.01	0.09	0.09	-0.19	-0.19
RN^{ε}	-0.24	-0.24	-0.25	-0.25	-0.08	-0.08	-0.09	-0.09	0.01	0.01
SO^{γ}	0.01	0.01	0.00	0.00	-0.06	-0.06	0.04	0.04	-0.11	-0.11
HN^{ε}	-0.21	-0.21	0.30	0.30	-0.23	-0.23	0.00	0.00	-0.09	-0.09
YC^{ζ}	-0.13	-0.13	-0.02	-0.02	-0.30	-0.30	0.37	0.37	-0.05	-0.05
FC^{ζ}	-0.01	-0.01	-0.03	-0.03	-0.04	-0.04	0.07	0.07	-0.11	-0.11
LC^{δ}	-0.01	-0.01	-0.03	-0.03	0.04	0.04	-0.01	-0.01	-0.12	-0.12
CS^{γ}	-0.05	-0.05	0.33	0.33	0.13	0.13	-0.28	-0.28	-0.12	-0.12
C.1	0.50	-0.50	0.04	-0.04	0.41	-0.41	-0.27	0.27	-0.05	0.05
C.2	0.12	-0.12	0.00	0.00	-0.35	0.35	-0.34	0.34	0.04	-0.04
C.3	0.06	-0.06	0.05	-0.05	-0.07	0.07	0.02	-0.02	0.49	-0.49
C.ar	0.06	-0.06	0.03	-0.03	-0.03	0.03	0.01	-0.01	0.27	-0.27
N.1	0.29	-0.29	-0.58	0.58	-0.16	0.16	0.22	-0.22	-0.02	0.02
N.4	0.35	-0.35	0.39	-0.39	-0.19	0.19	0.41	-0.41	-0.06	0.06
N.am	0.12	-0.12	0.06	-0.06	-0.37	0.37	-0.30	0.30	-0.07	0.07
O.2	0.00	0.00	0.04	-0.04	-0.03	0.03	-0.02	0.02	-0.31	0.31
O.3	-0.02	0.02	0.03	-0.03	-0.08	0.08	0.04	-0.04	-0.29	0.29

3.3.3 Fractional Overlap with Crystal Ligands

To evaluate the performance of the new DARS in FTMap energy functions, fractional overlap was calculated. This value, as described previously in Section 3.2.5,

indicates how well the solvent mapping can push probes into the area occupied by the small molecule ligand in the co-crystalized structure. Table 3.5 reports the overlap results (in percentage) of 48 benchmark cases from the LIGSITE^{csc} set (B. Huang & Schroeder, 2006). Referencing previous results from testing FTMap-related binding site prediction server FTSite, visual inspection was done to the apo proteins in the set. If deemed appropriate, certain chain(s) of the proteins instead of the whole PDB assembly was mapped. The mapped apo protein was superimposed back to the reported holo protein, and subsequently fractional overlap between hot spots, or consensus sites (CSs), and the crystal ligand was calculated. Two types of fractional overlap are documented: using 1) all hot spots and 2) the top five hot spots according to cluster strength.

Table 3.5: Overlap between crystal ligands and hot spots (or CSs). Columns report unbound protein PDB code, bound protein PDB code, crystal ligand PDB code, overlap percentages calculated using 1) all hot spots (CSs) from current DARS/FTMap, 2) all hot spots from new DARS, 3) top five strongest hot spots from current DARS/FTMap and 4) top five strongest hot spots from new DARS.

Unbound PDB	Bound PDB	Ligand PDB	FTMap all CSs (%)	New DARS all CSs (%)	FTMap top 5 CSs (%)	New DARS top 5 CSs (%)
3lck	1qpe	PP2	85.71	85.71	85.71	85.71
1pdy	1pdz	PGA	100.00	100.00	100.00	100.00
1a4j_AB†	1igj_AB	DGX	48.65	48.65	48.65	48.65
1a6u	1a6w	NIP	52.94	58.82	0.00	0.00
1ahc	1mrg	ADN	100.00	100.00	100.00	90.91
1bbs_A	1rne_A	C60	88.24	92.16	72.55	78.43
1brq	1rbp	RTL	71.43	66.67	71.43	61.90
1bya	1byb	GLC	93.33	93.33	73.33	71.11
1hsi	1ida	0PO	53.85	61.54	50.00	53.85
1ifb	2ifb	PLM	100.00	100.00	100.00	100.00
1krn	2pk4	ACA	66.67	66.67	66.67	66.67
1nna	1ivd_A	ST1	100.00	100.00	100.00	94.12
1pts_A	1srf_A	MTB	81.82	81.82	81.82	81.82
1qif	1acj	THA	80.00	80.00	80.00	80.00

1stn	1snc	THP	24.00	48.00	24.00	48.00
2sil	2sim	DAN	90.00	90.00	90.00	90.00
2tga	1mtw	DX9	54.55	36.36	27.27	36.36
7rat	6rsa	UVC	95.24	95.24	90.48	90.48
6ins	3mth	MPB	0.00	0.00	0.00	0.00
3app	1apu	IVA	76.47	76.47	50.00	61.76
1psn	1pso	IVA	81.25	81.25	81.25	81.25
1chg	3gch	OAC	8.33	8.33	0.00	0.00
1ypi	2ypi	PGA	100.00	100.00	88.89	55.56
1djb	1blh	FOS	100.00	86.67	100.00	86.67
1cge	1hfc	PLH	72.00	68.00	64.00	68.00
1ime_A	1imb_A	LIP	100.00	100.00	100.00	100.00
4ca2	1okm	SAB	88.89	100.00	88.89	77.78
2cba	2h4n	AZM	100.00	100.00	100.00	92.31
3p2p_A	5p2p_A	DHG	75.00	78.57	71.43	78.57
5cpa	7cpa	FVF	58.54	80.49	58.54	80.49
1l3f	2tmn	0FA	92.31	92.31	92.31	92.31
2ctb	2ctc	HFA	100.00	100.00	100.00	100.00
1phc	1phd	PIW and HEM	94.44	94.44	94.44	85.19
1esa	1inc	ICL	100.00	100.00	79.17	79.17
1gcg	1gca	GAL	100.00	100.00	100.00	100.00
1hel	1hew	NAG	51.16	46.51	51.16	44.19
1hxf	1dwd	MID	91.89	81.08	91.89	59.46
1npc	1hyt	BZS	73.33	73.33	73.33	73.33
1swb_A	1stp	BTN	100.00	100.00	100.00	100.00
1ula	1ulb	GUN	90.91	100.00	0.00	100.00
2ctv	5cna_A	MMA	100.00	100.00	100.00	100.00
2fbp	1fbp	F6P	87.50	90.63	75.00	78.13
3phv	4phv	VAC	18.48	33.70	18.48	33.70
3ptn	3ptb	BEN	100.00	100.00	100.00	100.00
3tms	1bid	UMP	70.00	75.00	70.00	75.00
5dfr	4dfr_A	MTX	54.55	54.55	54.55	54.55
8adh	1cdo_A	NAD	84.09	84.09	47.73	75.00
8rat	1rob	C2P	90.48	100.00	90.48	90.48

[†]Underscore followed by capital letters indicate specific chain(s) mapped.

In summary, when considering fractional overlap between all hot spots and crystal ligands, the new small molecule DARS performance was generally satisfactory. For 30 out of 48 cases in the LIGSITE^{csc} set, the small molecule DARS performed just as well as the present DARS in FTMap. In fact, for 12 cases the new DARS increased the fractional overlap. The new DARS performance only worsened for 6 cases, which is 13% of the total LIGSITE^{csc} set. In total, 88% of the cases in the LIGSITE^{csc} benchmark showed that the systematically-derived DARS was on par with or outperforms the *ad hoc* DARS in terms of enriching hot spots in the validated ligand binding sites (Figure 3.5). However, if solely factoring in the five strongest hot spots, only 77% of the benchmark cases demonstrated that systematically-generated DARS performed well (50% cases had the same overlap percentage, 27% cases did better). The number of under-performing cases increased to 23% (Figure 3.5). A closer look at Table 3.5 suggests that certain systems were more sensitive to the updated DARS. There was one case where the current FTMap did not have any one of the top five hot spots overlapping with the crystal ligand: PDB 1ula (bound: 1ulb), which is purine nucleoside phosphorylase. The fifth hot spot using new DARS mapping coincided with the small ligand guanine. On the other hand, it had come to the seventh hot spot using the current FTMap to overlap with guanine. Granted it was noticeable that the overlapping hot spots were relatively weak in both cases: the CSs contained eight and seven probe clusters for new and old DARS, respectively. Also, as an example, with PDB structure 1hxf (bound PDB 1dwd, the human thrombin), the binding site of its inhibitor was better predicted by the current FTMap. Using all hot spots, current FTMap showed 91.89% overlap with the inhibitor

while the new DARS reduced the number to 81.08%. Accounting for top five hot spots reduced overlap to 59.46% with the new DARS. Another representative case was PDB 1ypi (bound 2ypi). Both the current FTMap and the new DARS showed that 100% of the inhibitor of this triosephosphate isomerase protein was covered by all hot spots. Once switching to calculating overlap with only top five hot spots, the percentage overlap dropped to only 55.56% for the new DARS, as opposed to 88.89% overlap using the present version of FTMap. In the case of thrombin (PDB 1hxf), it seemed that the new DARS predicted additional hot spots and the top CSs became weaker. By visually inspecting triosephosphate isomerase (PDB 1ypi) mapping, the overlap of hot spots to the ligand actually seemed quite similar. The reason why overlap showed a 33.33% difference was probably the small size of the ligand 2-phosphoglycolic acid (its molecular formula is $C_2H_5O_6P$). These results are thought-provoking in multiple ways. First, the seemingly system-dependent performance indicates that we most likely need a more diverse set of benchmark cases, which will not only return more informative performance statistics but also give us better ideas why the DARS performance is system-dependent. Ideally there should be analysis on which proteins map better with which DARS potential. In addition, we will also need alternative evaluation metrics other than ligand fractional overlap, as smaller ligands may be excessively sensitive to positions of probes and lead to computational artifact in the form of exaggerated results. Detailed analyses of cases paired with visual representation will be included in the next section.

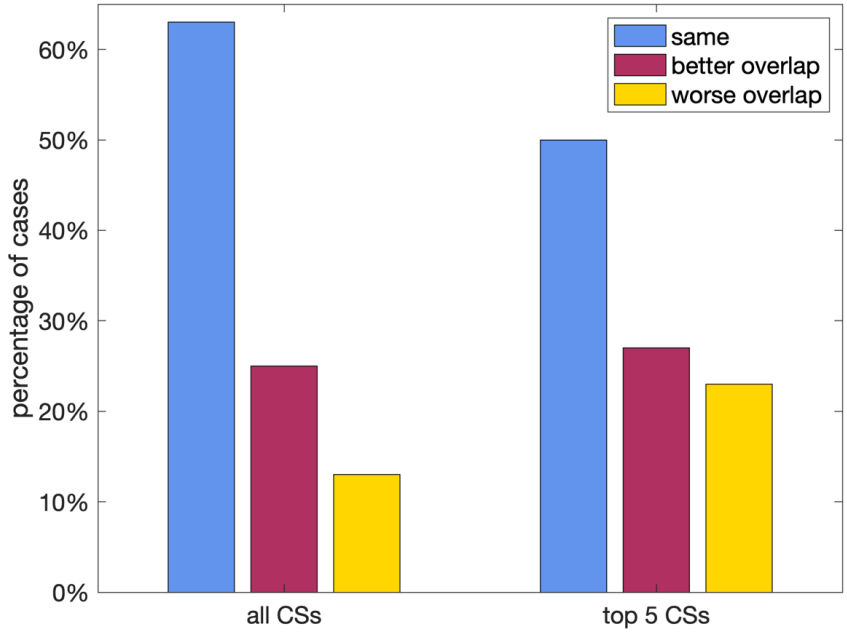

Figure 3.5: Performance statistics of the new DARS compared to the current FTMap. There are three categories: "same", "better overlap" and "worse overlap". Respectively, these categories mean the overlap stayed the same, increased or declined with the updated DARS for the 48 benchmark cases. The y-axis shows the percentage of the cases out of the whole set corresponding to the categories. There are two groups of results: overlap calculated using all CSs or only the top five CSs.

3.3.4 Case Studies

Five cases with the most pronounced differences in crystal ligand fractional overlap will be discussed in this section. For easier navigation, each case was summarized in numbered paragraphs. These cases demonstrate that the small molecule DARS performance is system-dependent. It will be desirable to evaluate performance on a more diverse set of proteins, which will help us comprehend the nature of the small molecule pairwise potential and provide ideas for improvement.

(1) *Staphylococcal* nuclease (unbound PDB 1stn, bound PDB 1snc, ligand PDB THP) (Hynes & Fox, 1991; Loll & Lattman, 1989). Fractional overlap increased with updated DARS using all and top five hot spots. As shown in Figure 3.6, the spatial overlap between hot spots and the ligand seemed better with the new DARS (pink line representation); probes were placed closer to the ligand at both the di-phosphate and the thymine ends. The overlapping hot spots also contained more probe clusters with the new pairwise potential (46 probe clusters in total) than the current version of FTMap (42 probe clusters).

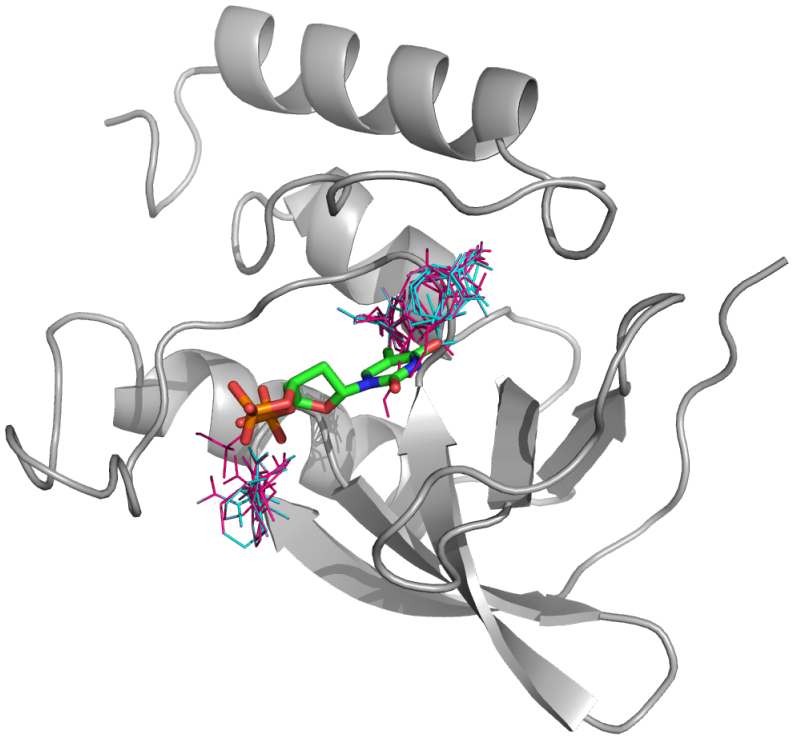

Figure 3.6: Mapping results of *Staphylococcal* nuclease (apo structure 1stn, grey cartoon). Green sticks represent the ligand THP, thymidine-3',5'-diphosphate. Pink lines are the overlapping hot spots from mapping with the new DARS. Cyan lines are overlapping hot spots from mapping with the regular FTMap.

(2) Beta-lactamase precursor (unbound PDB 1djb, bound PDB 1blh, ligand PDB FOS) (Chen, Rahil, Pratt, & Herzberg, 1993; Chen et al., 1996). Fractional overlap dropped with updated DARS using all and top five hot spots. The ligand is fairly small. Using the new small molecule DARS, two out of the three top hot spots directly coincided with the ligand; the third hot spot was approximately 2.3 Å away from the phosphate group (Figure 3.6). On the other hand, all top three regular FTMap hot spots directly overlapped with the ligand, and there was no gap between hot spots (cyan lines in Figure 3.7).

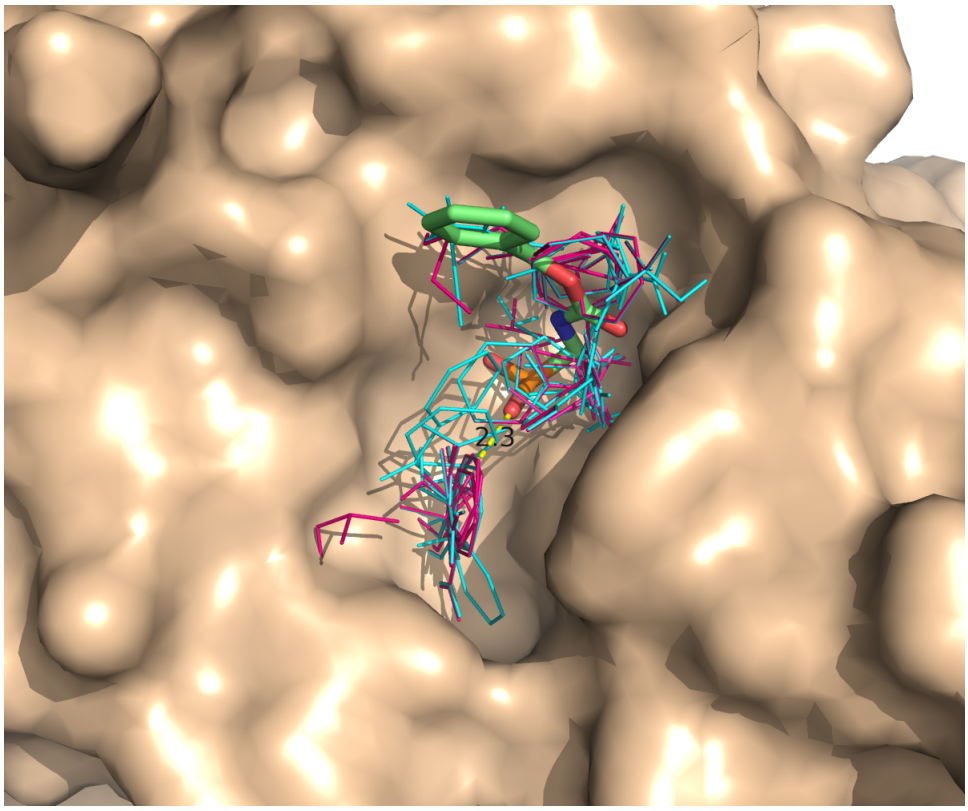

Figure 3.7: Mapping results of beta lactamase precursor (apo structure 1djb, wheat surface). Green sticks show the ligand FOS. Pink lines are hot spots from mapping with the new DARS. Cyan lines are hot spots from mapping with the regular FTMap. The yellow dotted line is the measured distance between the new DARS hot spot and the closest atoms of the ligand, which is 2.3 Å.

(3) Carboxypeptidase A (unbound PDB 5cpa, bound PDB 7cpa, ligand PDB FVF) (Kim & Lipscomb, 1991; Rees, Lewis, & Lipscomb, 1983). Fractional overlap increased with updated DARS using all and top five hot spots. The new pairwise potentials not only detected hot spots deep within the cavity, covering most part of the ligand FVF, but also identified the patch at the entrance of the cavity. This spot was missed by the current FTMap, even though the crystal ligand has an aryl group interacting with the patch (see Figure 3.8).

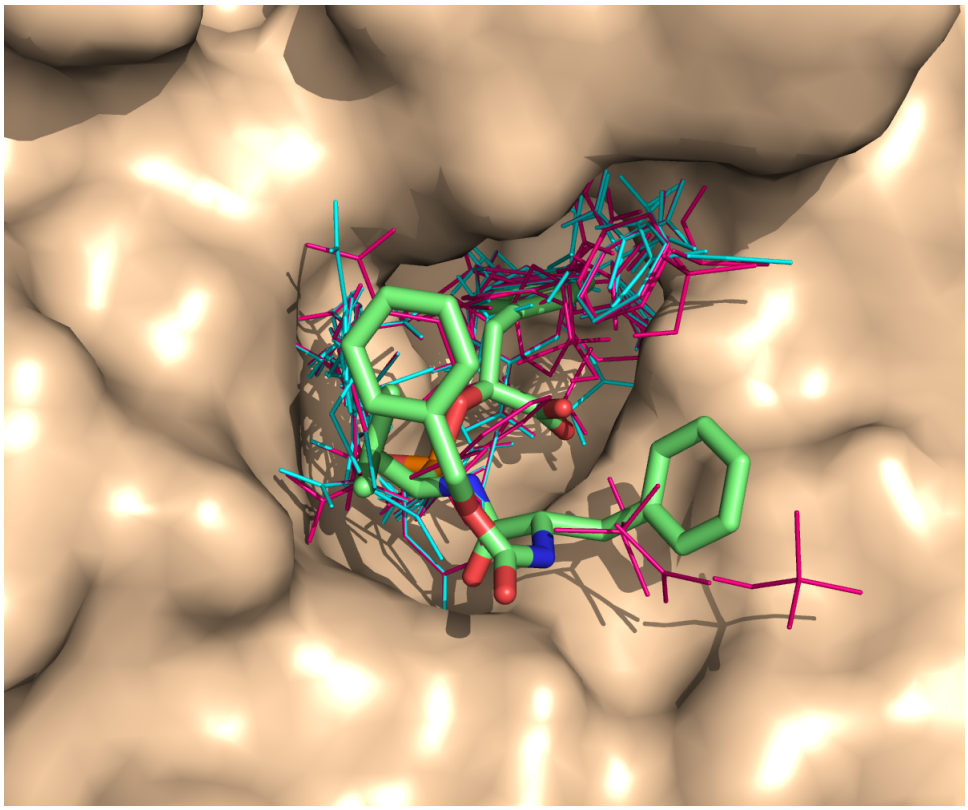

Figure 3.8: Mapping results of carboxypeptidase A (apo structure 5cpa, wheat surface). Green sticks show the ligand FVF. Pink lines are hot spots from mapping with the new DARS. Cyan lines are hot spots from mapping with the regular FTMap.

(4) Human thrombin (unbound PDB 1hxf, bound PDB 1dwd, ligand PDB MID (Banner & Hadvary, 1991; E. Zhang & Tulinsky, 1997). Fractional overlap dropped with

updated DARS using both all and top five hot spots. It seemed the new pairwise potential detected new spots on the protein surface and diluted the strengths of all hot spots (the third largest hot spot with the new DARS only included eight probe clusters; the third hot spot had 15 probe clusters in the current FTMap results). This could be a scenario where the new DARS was finding more false positive results. There were ten weak hot spots with strengths < 10 in addition to the top two strongest hot spots with strengths >15. On the contrary, the current FTMap results contained four dominant hot spots with at least 15 probe clusters and only four additional weak hot spots. It also seemed that the new DARS missed to place enough probes to at the binding site of the naphthalene ring (see Figure 3.9), showing a weakened capacity to detect hydrophobic interactions.

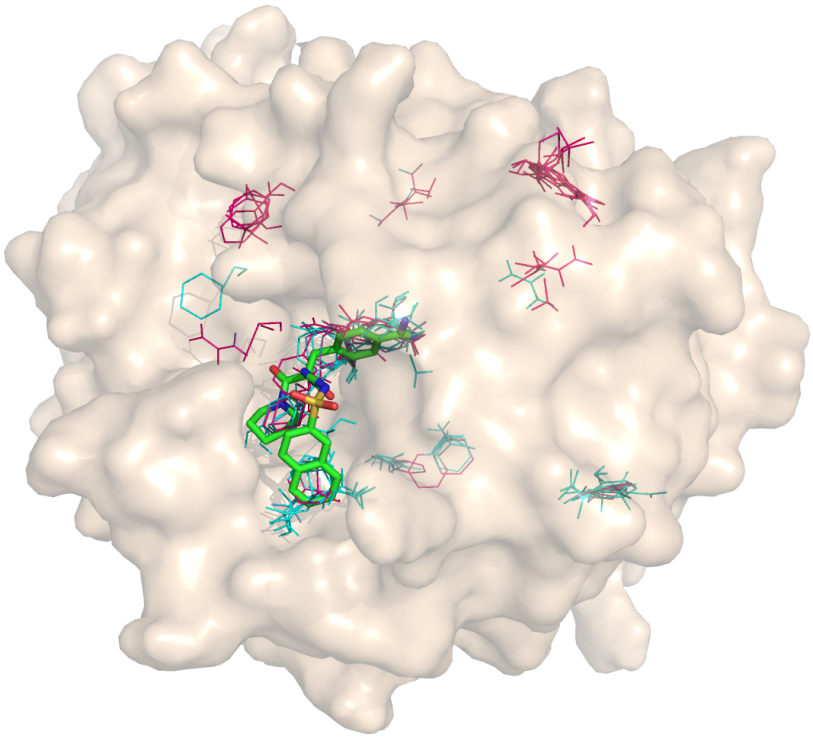

Figure 3.9: Mapping results of human thrombin (apo structure 1hxf, transparent surface). Green sticks show the inhibitor MID. Pink lines are hot spots from mapping with the new DARS. Cyan lines are hot spots from mapping with the regular FTMap.

(5) Liver alcohol dehydrogenase (unbound PDB 8adh, bound PDB 1cdo chain A, ligand PDB NAD) (Colonna-Cesari et al., 1986; Ramaswamy, el Ahmad, Danielsson, Jornvall, & Eklund, 1996). Fractional overlap was the same with updated DARS using all hot spots. But the top five hot spots with the new DARS increased overlap by 27.3%. Both FTMap and the new DARS predicted the hot spots along the belt-like ligand binding site well. But the new DARS hot spots showed a tighter overlap with di-nucleotide molecule (Figure 3.10).

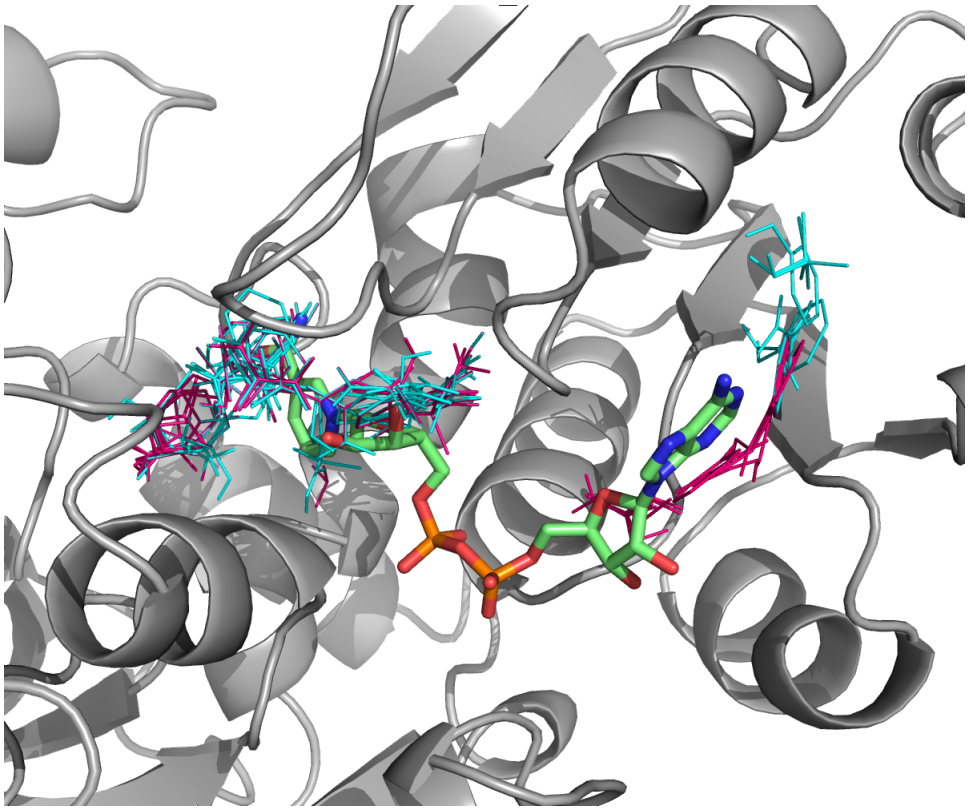

Figure 3.10: Mapping results of liver alcohol dehydrogenase (apo structure 8adh, grey cartoon). Green sticks show the long nicotinamide-adenine-dinucleotide. Pink lines are the top five hot spots from mapping with the new DARS. Cyan lines are top five hot spots from mapping with the regular FTMap.

3.4 Conclusions and Future Directions

There has been an increased need to improve the current energy function in FTMap. The solvent mapping server can robustly identify hot spots on protein surface and provide guidance in terms of druggability (Kozakov, Hall, Napoleon, et al., 2015). However, for better prediction of hydrogen bonds and placement of the probes, a superior energy function is of immediate necessity. We are also seeking to expand the solvent probes beyond the current 16 types to include fragments and building blocks more relevant to medicinal chemistry and fragment-based drug discovery (FBDD). This chapter described the effort to replace the *ad hoc* DARS potential implemented in the present version of FTMap. This chapter detailed that native structural data was collected from the training set of 2499 PDBbind protein-fragment complexes, as well as the reference decoy statistics. Initial round of testing new DARS has concluded, using a cutoff distance of 3.5 Å, which is most applicable to forming hydrogen bonds. Based on the results of LIGSITE^{csc} benchmark cases, the new DARS showed prospects of improving the performance of FTMap but needed continued enhancement.

In terms of possible future research strategies, first of all, it is important to note that there were only nine atom types present in the solvent probes, which may not be diverse enough to discriminate amongst key chemical interactions. After all, the protonation states of small molecules could not be assessed fully based on such a discreet atom typing system. In reality, this restrictive atom typing scheme would not work due to the continuity nature in atom protonation states. We may need to seek alternatives in order to increase the diversity of small molecule atom types. Second, many sampling

parameters should be re-calibrated with the new DARS. Ideally more effort should go into rigorously choosing the number of eigenvectors to include in FFT, and re-adjusting the weights of energy terms in the scoring function. Furthermore, the evaluation metric reported in this chapter is not the most rigorous or fair one. The fractional overlap is a high-throughput strategy for evaluating solvent mapping performance, but there is arbitrariness in the calculation. For instance, smaller ligands are overly sensitive to the metric and could lead to skewed results; using only the top five hot spots could also over-represent the weak but still highly-ranked CSs. In the future it will be more informative to examine the hydrogen bonds and contacts formed between probes and the protein, using measurements such as mapping and ligand fingerprints (Bohnuud, Kozakov, & Vajda, 2014), as well as examining false negatives and false positives of hot spots. Finally, it will be interesting to try multi-bin DARS application. This new implementation in PIPER could be the next breakthrough in FTMap solvent mapping, as we will be able to apply distance-based interaction energies. In other words, the DARS potential constructed from [0, 3.5 Å] cutoff will be applied to emphasize hydrogen bonding between probes and the protein; simultaneously the [0, 6.5 Å] pairwise potential can recover hydrophobic interactions removed by the [0, 3.5 Å] DARS potential. The structural statistics for both cutoff values have been collected, which will facilitate the development of multi-bin DARS in the future.

CHAPTER 4 Modeling the Interaction between Trypsin and the Pancreatic Secretory Trypsin Inhibitor (PSTI) Encoded by *SPINK1*

4.1 Introduction

The *SPINK1* gene encodes the serine peptidase inhibitor Kazal type 1, which is also known as the pancreatic secretory trypsin inhibitor (PSTI). This protein is secreted from pancreatic acinar cells and released into pancreatic juice (https://www.ncbi.nlm.nih.gov/gene/6690). A common mutation that has been linked to elevated risk for chronic pancreatitis is N34S in the *SPINK1*-encoded trypsin inhibitor (Aoun et al., 2008; Whitcomb, 2013). However, the exact molecular mechanism behind N34S as the chronic pancreatitis-predisposing mutation is still unclear (Boulling, Chen, Callebaut, & Férec, 2012) (Hegyi & Sahin-Toth, 2017). There is a trypsin-dependent pathological pathway in chronic pancreatitis. Accumulation of trypsin in pancreas can lead to autodigestion of the cells and harm pancreatic tissues. The 6.2-kDa protein PSTI can potentially inhibit trypsin, acting as a protective mechanism. It is hypothesized that the N34S mutant has reduced inhibition of trypsin compared to the wildtype, which increases the risk of chronic pancreatitis (Hegyi & Sahin-Toth, 2017). A few research groups have reported their findings in terms of the trypsin inhibitory activity differences between the wildtype and the N34S mutant. The Kuwata group expressed the proteins in *Saccharomyces cerevisiae* and measured residual trypsin activities under a range of alkaline and acidic pH conditions. They found only minor differences in residual trypsin activity of wildtype and N34S mutants, and it was established that their results were inconclusive (Kuwata et al., 2002). Another paper published by Kiraly and colleagues

confirms that biochemical defect caused by N34S is unrelated to trypsin inhibitory activity using proteins expressed by human embryonic kidney 293T cells, supporting the previous findings by the Kuwata group (Kiraly, Wartmann, & Sahin-Toth, 2007). It is recognized that the experimental evidence for diminished trypsin inhibition of N34S is insufficient at this point (Hegyi & Sahin-Toth, 2017). Nonetheless, the structural impact of N34S on trypsin binding was supported by a molecular modeling study, plus secondary structure predictions and analysis of the PSTI crystal structures (Boulling et al., 2012; Kuwata, Hirota, Nishimori, Otsuki, & Ogawa, 2003; Pfutzer et al., 2000). Consequently, there is still interest to explore the structural impact of N34S on trypsin inhibition.

At the time this research project was formulated, there were no crystal structures available for the human trypsin-PSTI complex. By examining the crystal structures of human PSTI in un-complexed and complexed forms (the latter co-crystalized with bovine chymotrypsinogen), Boulling and colleagues proposed that the N34S mutation sits in a critical loop that might indirectly impact the binding to trypsin (see Figure 4.1) (Boulling et al., 2012). The bovine chymotrypsinogen and the human trypsin share overall sequence similarity of 54% according to the EMBOSS Needle alignment program (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) (Needleman & Wunsch, 1970); the amino acids lining the trypsin inhibitor binding sites differ by only a few residues and the backbone are essentially identical (as illustrated in Figure 4.2). Based on this similarity, the chymotrypsinogen-PSTI complex sets a reasonable basis for understanding the interaction between trypsin and PSTI. In the complex (PDB ID 1cgi) (Hecht,

Szardenings, Collins, & Schomburg, 1991), the side chain hydroxy group of Y33 forms a hydrogen bond with T40 (see Figure 4.1), an interaction not present in the un-complexed form (PDB ID 1hpt) (Hecht, Szardenings, Collins, & Schomburg, 1992). Since Y33 is the immediate neighbor of N34 and the Y33-T40 hydrogen bond facilitates the correct conformation of the loop at the inhibitor-enzyme complex interface, it seems reasonable to speculate that N34 plays an important role in the conformational dynamics of the interface. Boulling's analysis points to a possible structure-based explanation for N34S as the disease-causing mutation. Pfutzer and colleagues came to a similar conclusion through molecular modeling using the porcine PSTI (Pfutzer et al., 2000).

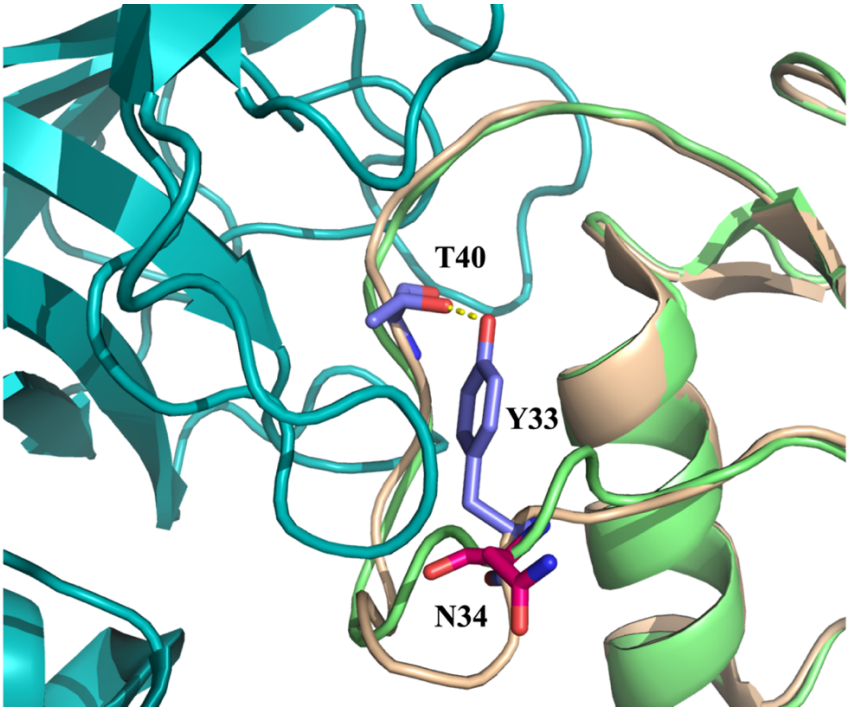

Figure 4.1: Key interactions at the PSTI-chymotrypsinogen interface. Important interactions form at the loop where N34 resides. Superposition of unbound human PSTI (green cartoon, PDB ID 1hpt) with the PSTI-chymotrypsinogen complex (wheat and cyan cartoons, respectively; PDB ID 1cgi) shows the bound loop conformation depends on the hydrogen bond between Y33 and T40 at the interface (purple residues). The mutation N34S (magenta) happens in a region neighboring the interface.

This chapter details our effort to further elucidate the structural origin of the N34S mutation-associated pancreatic disease. We used molecular dynamics (MD) to generate conformational ensembles for both wildtype and the N34S mutant of the human PSTI, in order to examine the change in loop dynamics introduced by the mutation. We then subjected the conformational ensembles of the two PSTI variants to rigid body docking using PIPER (Kozakov et al., 2006). Using physics-based scoring function (described in further details in Section 1.2), PIPER docking yields a large number of protein complexes, including structures that are not near-native. Such docked poses represent encounter complexes, which are important transient states along the protein association pathway (Kozakov et al., 2014). The number of near-native poses therefore represents the productive encounters between the receptor and ligand proteins. Additionally, based on our experience, the number of docked near-native structures could be used as an indicator for the quality of protein-protein interface. For instance, the number of docked near-native hits was used to distinguish between crystal or biological dimers (Yueh et al., 2017). Based on the hypothesis that the N34S mutation impairs trypsin inhibition, we expected that wildtype PSTI MD ensemble should generate more productive conformations that can bind trypsin than the mutant ensemble, meaning that the wildtype should have a higher number of near-native hits than the docked N34S mutant. The detailed protocol of ensemble docking of MD snapshots and the calculation of ensemble near-native hits will be outlined in the next section.

4.2 Methods

4.2.1 Preparation of Structures

We used the bound conformation of the human PSTI crystal structure to generate MD trajectories. The PSTI protein was co-crystalized with chymotrypsinogen (PDB ID 1cgi) (Hecht et al., 1991). This structure was a variant of the human PSTI and it included four additional mutations: K41Y, I42E, D44R and N52D. These four residues were mutated back to match the original PSTI sequence, using Maestro GUI (Schrodinger, 2017). The N34S mutation was also introduced using Maestro. There is an un-complexed form of human PSTI (PDB ID 1hpt) (Hecht et al., 1992), but the N34 residue in the structure adopts an unusual high-energy conformation in the left-handed alpha-helical region of the Ramachandran plot (Boulling et al., 2012). Although such conformation may be possible for asparagine (Deane, Allen, Taylor, & Blundell, 1999), we determined it to be an artifact from crystal contact. Running simulations with crystal contact may lead to structures being trapped in a pre-fixed stable conformation, and thus it was undesirable to start MD with the PDB structure 1hpt. More importantly, it was beneficial to start the MD simulations from a conformation closest to the expected bound form, so we decided to focus on the bound conformation of human PSTI as the starting structure of our simulations.

As described above in Section 4.1, at the time this work was done no 3D structure of the PSTI-trypsin complex was available. To model the near-native interface, we superimposed the trypsin-bovine pancreatic trypsin inhibitor (BPTI) complex (PDB ID 2ra3) (Salameh, Soares, Hockla, & Radisky, 2008) onto the chymotrypsinogen-human

PSTI complex (PDB ID 1cgi). The modeled interface of trypsin-human PSTI then became the reference structure later used for evaluating docked near-native poses. Chymotrypsinogen is the inactive precursor of chymotrypsin. Chymotrypsinogen is converted into fully active enzyme through cleavage of a single peptide bond (Berg, Tymoczko, & Stryer, 2002). Chymotrypsin and trypsin are both serine proteases; they share high sequence identity and also very similar tertiary structures (Ma, Tang, & Lai, 2005). As shown in Figure 4.2, the interfaces of the two complexes overlay well. In fact, the backbone of the key interactions at the interface loops are nearly identical (Figure 4.2, lower right corner). Based on these facts, we determined it was reasonable to use this model as the reference near-native interface.

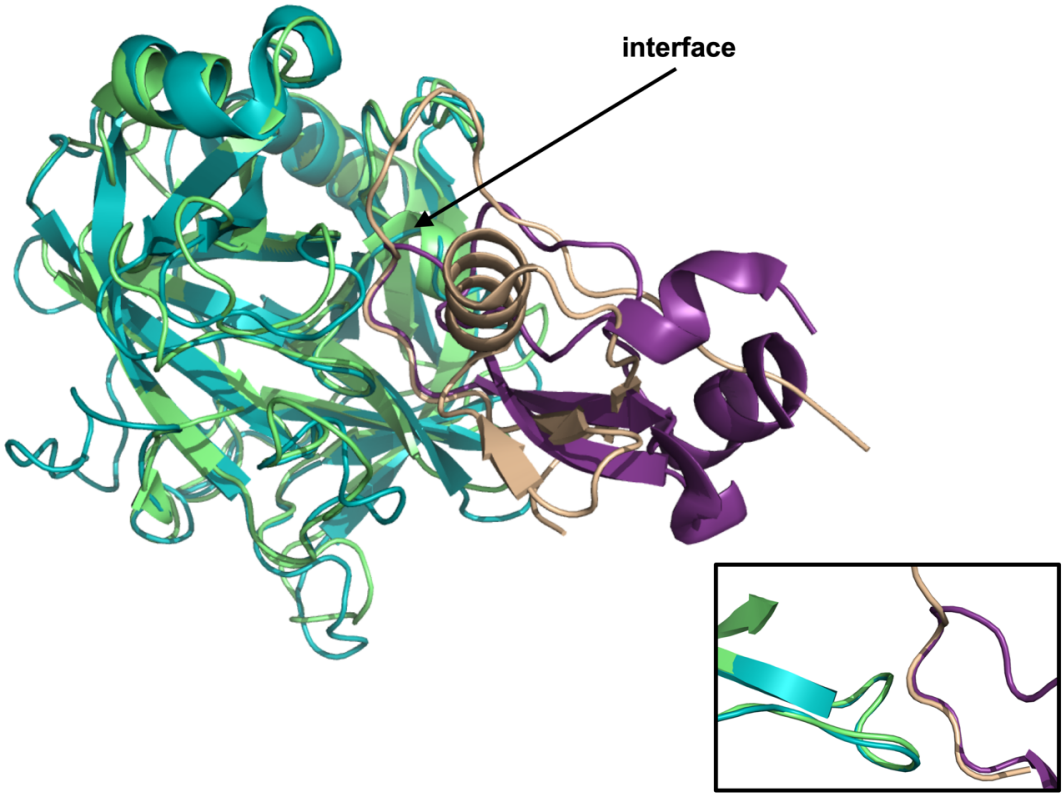

Figure 4.2: Modeled interface of trypsin-human PSTI. The trypsin-bovine pancreatic trypsin inhibitor (BPTI) complex was available (PDB ID 2ra3, green: trypsin, purple: BPTI). The complex was

superimposed onto the chymotrypsinogen-human PSTI complex (cyan and wheat, respectively, PDB ID: 1cgi), making a modeled trypsin-PSTI interface. The loops of BPTI and PSTI at the interface are nearly identical. The zoomed-in image on the lower right corner shows the key interaction sites are aligned.

4.2.2 Molecular Dynamics and Clustering Protocol

We applied MD simulation to two proteins: 1) wildtype PSTI extracted from the co-crystalized form (PDB ID 1cgi) and 2) the N34S mutant. We used AMBER99SB-ILDN (Lindorff-Larsen et al., 2010) force field and the TIP3P water model. Before setting up systems and MD parameters, all proteins were processed by Maestro GUI's Protein Preparation Wizard (Sastry, Adzhigirey, Day, Annabhimoju, & Sherman, 2013). The simulations were performed using the GPU version of Desmond (Bowers et al., 2006) on a desktop computer with four Nvidia GTX 1080 graphics cards. Before production runs, every simulation started with the built-in standard Desmond equilibration and relaxation protocol. The production runs were configured NPT using Nose–Hoover chain with a 1 ps relaxation time for thermostat (single temperature group), and Martyna–Tobias–Klein barostat with 2 ps relaxation time and isotropic coupling. The RESPA integrator was set to Δt= 2.5 fs for bonded and near nonbonded interactions and Δt = 7.5 fs for far nonbonded, a previously used standard protocol (Ignatov et al., 2019). Water molecules were constrained with SHAKE. Each simulation was set to run for 500 ns, with recording intervals of 1000 ps. For each protein, we ran ten independent simulation each starting at a random initial velocity, as an effort to average the noisiness of MD simulations. The resulting ten simulations were concatenated to generate an aggregate of 5 μs MD trajectory. For RMSD analysis of the trajectories, we used the

Desmond utility program Simulation Event Analysis implemented in Maestro (Schrodinger, 2017) (Bowers et al., 2006). The RMSD values were calculated referencing the bound conformation of PSTI in the PDB structure 1cgi based on all C-alphas.

To extract representative structures from the MD simulations, we subjected the trajectories to clustering. First, using the Desmond Trajectory Clustering program (Bowers et al., 2006) implemented in Maestro (Schrodinger, 2017), pairwise fitted interface root-mean-square deviation (RMSD) matrices for all frames were generated based on C-alphas of the proteins. A greedy clustering algorithm, which finds nearest neighbors within a certain radius, uses those RMSD matrices as the distance measure. The clustering radii were tailored to individual trajectory with respect to pairwise RMSD distributions, based on our previous experience (Ignatov et al., 2019; Kozakov, Clodfelter, Vajda, & Camacho, 2005). The general thought process was the following: first, if there was a bimodal distribution, the minimum between the two peaks was chosen as the optimal clustering radius; if there was no such distribution, the peak RMSD value of the distribution was treated as an optimal clustering radius. The clustering radii we eventually applied were 2.8 Å and 2.6 Å for the wildtype and the N34S mutant, respectively. Finally, the cluster centers, which were structures with the largest numbers of neighbors, were extracted from the trajectories using Desmond utilities (Bowers et al., 2006) (Schrodinger, 2017). Only “significant” cluster centers were saved for the next step. More specifically, if a cluster had lower than 2% of total population, meaning that the cluster consisted of fewer than 100 frames out of the total 5020 frames from the concatenated trajectories, that cluster was assessed to be “insignificant”. Based on these

criteria, 12 wildtype cluster centers and ten N34S mutant cluster centers were extracted as representative structures from the simulations.

4.2.3 Ensemble Docking

All MD snapshots were docked to the trypsin structure (chain A of PDB ID 2ra3), using global FFT sampling implemented in PIPER (Kozakov, Brenke, Comeau, & Vajda, 2006). The docking parameters were configured to mimic standard ClusPro protein dockings, including the standard weight coefficients used in the scoring scheme (Kozakov et al., 2017) (further discussions about ClusPro can be found in Section 1.2). Similar to ClusPro docking, 70000 rotations corresponding to ~5 degrees in terms of the Euler angles were explored. However, instead of clustering the 1000 lowest energy poses to generate models like in ClusPro, we counted the number of near-native hits (N_{nn}) in these 1000 poses. To be considered near-native, a docked pose needed to have interface root-mean-square deviation (iRMSD) lower than 10 Å compared to the reference structure discussed above in Section 4.2.1.

For each trypsin-PSTI variant (wildtype or N34S mutant) docking, it was further represented by the ensemble docking of the MD snapshots to the receptor trypsin. To account for the ensemble effect, a weighted average N_{nn} was calculated for each of the PSTI docking. The weights for these MD snapshots (i.e. cluster centers) were calculated by dividing the cluster sizes by the total number of frames in the merged trajectory, which was essentially the probability of that particular cluster center. And the weighted N_{nn} for each PSTI ensemble docking was therefore the sum of the products of probability

and N_{nn} for each MD snapshot. Figure 4.3 below explains the protocol in further details.

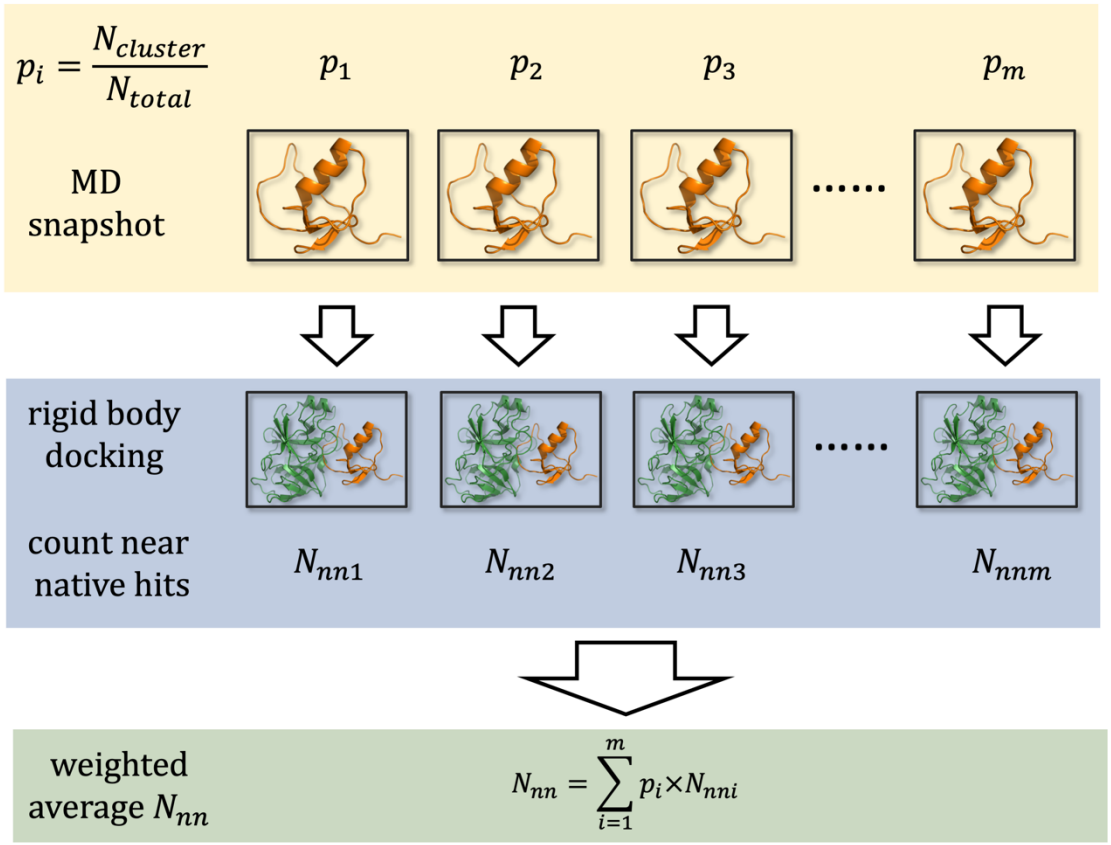

Figure 4.3: The ensemble docking protocol. The probability for each MD snapshot was calculated as the cluster size divided by the total number of frames (top yellow box). Each snapshot was docked to the receptor trypsin, and each generated a number of near-native hits (middle blue box). The weighted average near-native hits were calculated for that protein docking (bottom green box).

4.3 Results and Discussions

4.3.1 Molecular Dynamics Results

Based on the hypothesis that N34S has impact on the loop dynamics of PSTI, we expected the distributions of RMSD deviation from the bound conformation to be different in the MD trajectories of the wildtype and N34S mutant. As illustrated in Figure 4.4, the wildtype has a more even distribution of RMSD values in the trajectories,

especially in the range of 1.5 Å to 2.5 Å. On the contrary, the N34S mutant has a sharper peak near 2 Å with a tail extending beyond 5 Å. The histograms in Figure 4.4 demonstrate the change in loop conformation statistics introduced by the N34S mutation. The residence time of loop conformations at roughly 2 Å away from the bound form is clearly longer in the N34S simulations than in the wildtype simulations. Using the two-sample Kolmogorov-Smirnov test (MathWorks, 2019) implemented in MATLAB (version R2019a), it is shown that the RMSD distributions of wildtype and N34S MD trajectories are unequal at the 5% significance level.

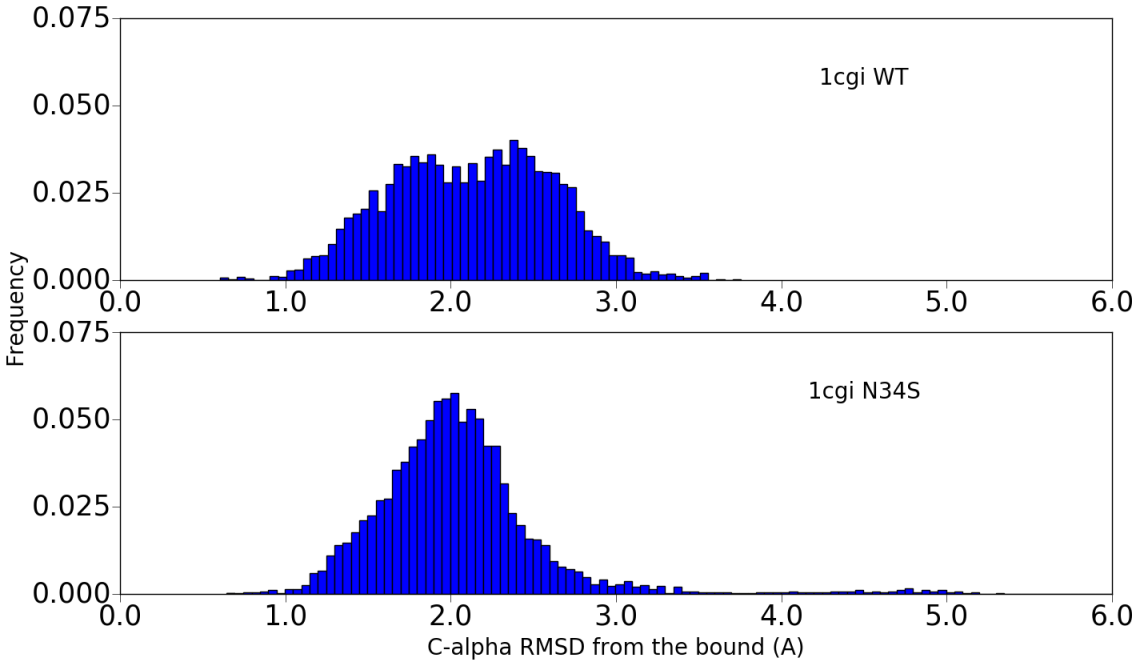

Figure 4.4: Distributions of RMSD values in the MD trajectories of the PSTIs, using all C-alphas in the bound structure (PDB ID 1cgi) as the reference. The histograms in the top and bottom panels describe results from wildtype and the N34S mutant, respectively.

4.3.2 Ensemble Docking Results

We expected the weighted average near-native hits (N_{nn}) of the wildtype PSTI docked to trypsin to be higher than that of the N34S mutant. As described, the N34S mutant was hypothesized to have diminished inhibition of trypsin, a possible underlying mechanism for the predisposition to pancreatitis (Boulling et al., 2012; Kuwata et al., 2003; Pfutzer et al., 2000). As shown in Figure 4.5, the wildtype PSTI generated weighted average N_{nn}'s higher than the N34S mutant using all three sets of standard ClusPro docking weight coefficients (Figure 4.5).

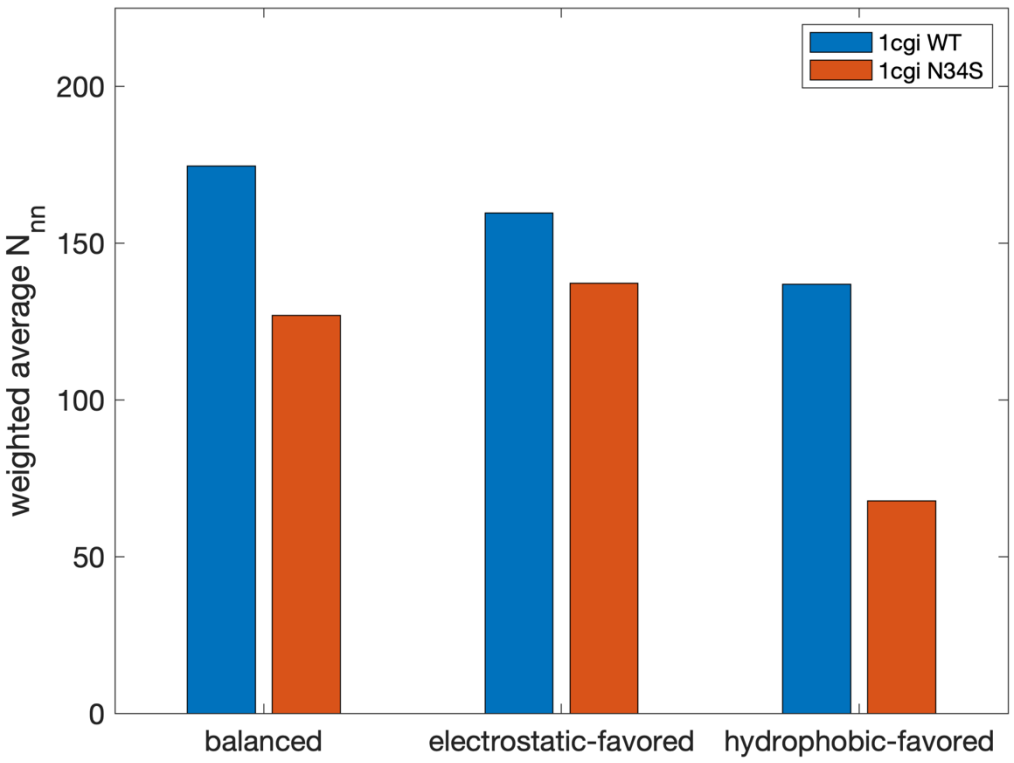

Figure 4.5: Weighted average near-native hits from docking of PSTI MD snapshots to trypsin. There are two types of MD snapshots, each generated by 1) the wildtype PSTI in the bound form (PDB ID 1cgi) or 2) the N34S mutant. The bars in the charts are grouped by the weight coefficients used in docking: balanced set, electrostatic-favored set and hydrophobic-favored set.

As we informed our users in the Nature Protocols paper (Kozakov et al., 2017), the selection of weight coefficients is not a trivial task. The coefficients we offer in ClusPro represent real biophysical driving forces that facilitate protein-protein interactions. The names of the coefficients associate preferred weights with assumed interactions. For example, the electrostatic-favored weight coefficient set is recommended if the association of the proteins is driven by electrostatic interactions. Same for the hydrophobic-favored coefficient set. The balance set is considered as the "default" or if no information about the proteins for docking is available. It is also known that the balanced set generally provide good results for docking enzyme-inhibitor complexes (Kozakov et al., 2017). The trypsin-PSTI interface is lined with residues with charged side chains such as glutamic acid, arginine and aspartic acid, indicating electrostatic interactions at play. Therefore, both balanced and electrostatic-favored weight sets should be highly applicable here. Hydrophobic-favored weight coefficient set may be the least optimal, but its results are also included for reference. As shown in Figure 4.5, all coefficient sets demonstrated that the wildtype MD snapshots generated higher weighted average near-native hits than the N34S MD snapshots, as we expected. In other words, the wildtype PSTI generated more productive conformations for binding trypsin than the N34S mutant according to the docking experiment.

We later discovered that the trypsin structure used in docking (PDB ID 2ra3) contained two mutations (Salameh et al., 2008): S195A and R117H. The first mutation was introduced to produce the catalytically inactive trypsin and was actually located at the inhibitor binding site. The second mutation was used to remove a sensitive site more

than 20 Å away from the PSTI binding interface. The combination of the two mutations allowed for improved crystallization results. For more accurate docking results, we used the same Maestro program (Schrodinger, 2017) described in Section 4.2.1 to mutate A195 and H117 back to match the original trypsin sequence, and then performed the same docking protocol. We did not expect docking to generate drastically different encounter complexes. The loop conformational change caused by the N34S mutation was anticipated to have a bigger impact on docking-predicted protein encounters. As shown in Figure 4.6, after these point mutations were applied to trypsin, the wildtype still produced more productive encounters than the mutant.

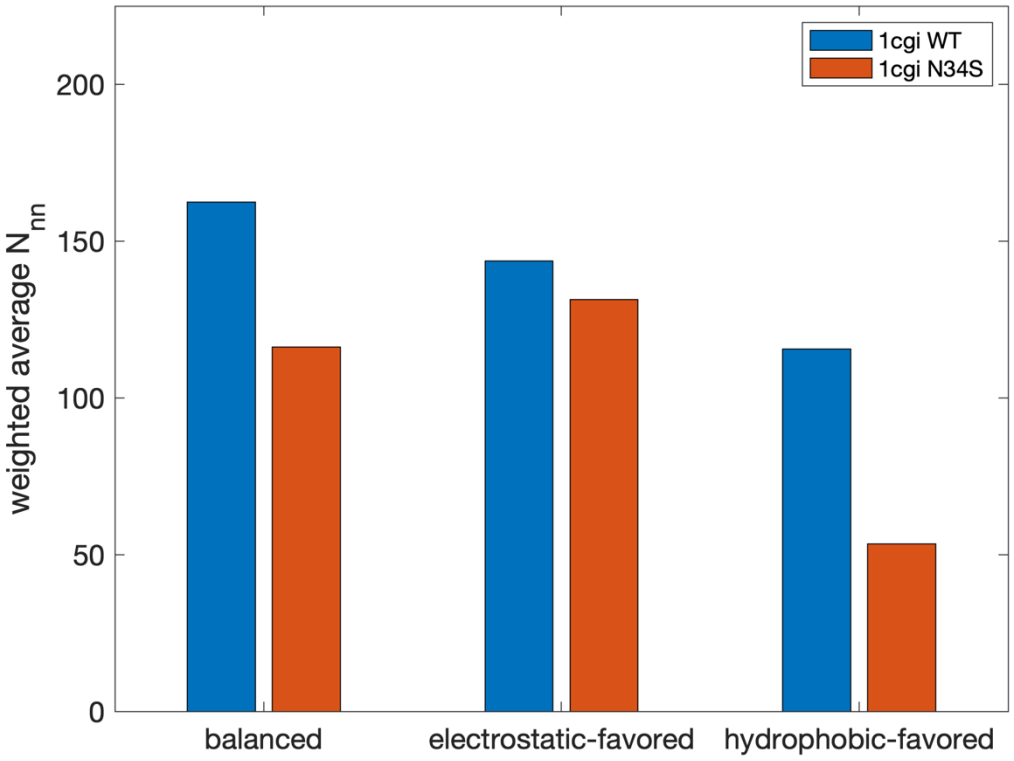

Figure 4.6: Weighted average near-native hits from docking of PSTI MD snapshots to trypsin, where the mutations introduced in crystallization process were reversed. The MD snapshots were generated starting from 1) the wildtype PSTI in the bound form (PDB ID 1cgi) or 2) the N34S mutant. The bars in the charts are grouped by the weight coefficients used in docking: balanced set, electrostatic-favored set and hydrophobic-favored set.

Despite interests and efforts to uncover the diminished trypsin inhibition of the N34S mutant, it has been difficult to demonstrate the altered inhibitory capacity of PSTI in *in vitro* studies (Hegyi & Sahin-Toth, 2017; Kiraly et al., 2007; Kuwata et al., 2002). The differences in trypsin inhibition between the wildtype *SPINK1*-encoded PSTI and the mutant were reported to be very minor, and the experimental results were recognized as inconclusive at most. Some authors propose that more sensitive assays are necessary to detect the reduced affinity of N34S for trypsin (Kuwata et al., 2002). A recent review reported that human PSTI should only inhibit as much as 13% of the potential trypsin (Hegyi & Sahin-Toth, 2017). Therefore, it is possible that one should focus on a low molar ratio of PSTI to trypsin in biochemical assays. It is also desirable to re-consider varying pH conditions in experiments in order to better mimic the physiological environment in the pancreas. Our computational modeling results demonstrated the structural impact of N34S mutation on trypsin inhibition, which is a molecular mechanism for the N34S-associated chronic pancreatitis that has not be fully validated by experimental outcomes.

4.4 Conclusions

We considered the hypothesis that the N34S mutant PSTI has a lower affinity for trypsin than the wildtype, due to the mutation-induced loop dynamics change. Based on the analysis of ensemble docking to trypsin, we observed differences in the numbers of productive conformations the wildtype PSTI and the N34S mutant generated. The differences, demonstrated by the numbers of weighted average near-native hits, were

consistent across three sets of standard ClusPro docking weight coefficients. However, so far there is not enough experimental evidence for the reduced trypsin inhibition of N34S (Hegyi & Sahin-Toth, 2017; Kiraly et al., 2007; Kuwata et al., 2002). A definitive experimental measurement of the effect of the N34S mutation on the binding affinity for trypsin is therefore required to better support the hypothesis. Of course, the functional role of the N34S mutation could lie elsewhere such as in impaired cellular folding, decreased levels of secretion, and reduced mRNA expression, none of which have been completely solved by experimental studies either (Boulling et al., 2012; Boulling et al., 2007; Kiraly et al., 2007). More recently, researchers have started to explore the uncharacterized flanking regions of the *SPINK1* gene (Hegyi & Sahin-Toth, 2017). The mechanism of action of the N34S-associated chronic pancreatitis remains a mystery. Our computational results provided a structure-based explanation, but stronger experimental evidence is essential to support the functional role of N34S in trypsin inhibition and elevated risk of chronic pancreatitis.

CHAPTER 5 Modeling of Antibody-Antigen Interactions

5.1 Introduction

Antibodies have been used as diagnostic and analytical tools for almost a century. More recently, antibodies have also been gradually accepted as therapeutic reagents for cancers and other diseases. Therefore, it has become increasingly important to understand the scientific details and properties of antibodies to improve antibody-related technologies for therapeutic uses (Maynard & Georgiou, 2000). Antibodies defend against pathogens by binding to and inactivating them, and also elicit other mechanisms of immune response. Antibodies in humans include five different isotypes, each presenting distinct structural features and mediating specific immune functions (Irani et al., 2015). The most abundant isotype in human serum is Immunoglobulin G (IgG) (Vidarsson, Dekkers, & Rispens, 2014). The basic structure of an IgG antibody molecule consists of four polypeptides: two identical light chains and two identical heavy chains. Each antigen binding site is formed by both the heavy and light chains (Irani et al., 2015). The antigen binding site is composed of six hypervariable loops (L1, L2, L3, H1, H2, and H3) (Marks & Deane, 2017), which then make up the complementarity-determining regions (CDRs) (see Figure 5.1). In a process called somatic hypermutation, CDRs of the immunoglobulin genes are mutated, forming a diverse library of antibodies. As a result, through repeated stimulation by the antigen, the affinities of antibodies for the antigen progressively increase. This process is called affinity maturation and acts as screening for the highest affinity antibody in response to the antigen (Alberts et al., 2002; Haynes, Kelsoe, Harrison, & Kepler, 2012).

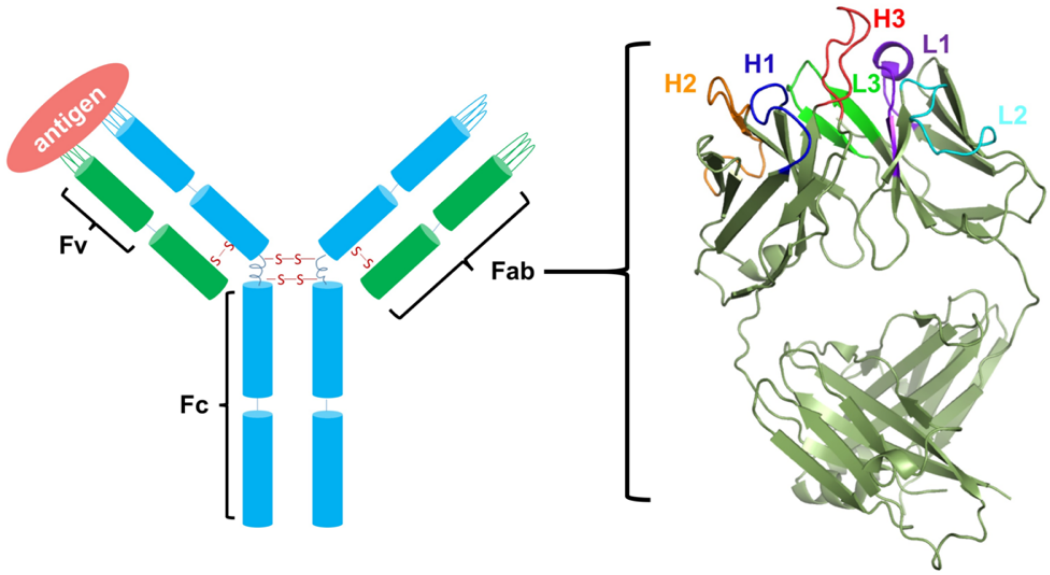

Figure 5.1: A typical IgG antibody structure. Each antigen binding fragment (Fab) consists of both a heavy (blue) and light chain (green), each with one constant and one variable domain. The variable domain is also referred to as the Fv region. The variable region consists of complementarity-determining regions (CDR) which bind to specific antigens. In addition to Fab, the antibody contains a tail region called fragment crystallizable region (Fc) that is critical to proper immune response by binding to Fc receptors. An example Fab is shown on the right (PDB ID: 3sm5).

The large structural repertoire associated with CDRs allows for tight and specific antigen binding properties and has been widely applied in therapeutics (Marks & Deane, 2017; Maynard & Georgiou, 2000; Shirai, Kidera, & Nakamura, 1999; Wilson & Stanfield, 1994). Over the years, antibodies as therapeutic agents have grown significantly. It was predicted that combined worldwide sales would reach nearly $125 billion by 2020. The number of approved monoclonal antibody drugs has steadily increased at the rate of three to five new products per year in recent years (Ecker, Jones, & Levine, 2015). Many of the antibody drugs on the market such as Humira®, Enbrel®, Avastin®, and Herceptin® have achieved sales greater than $6 billion. These antibodies target a wide range of therapeutic areas, including but not limited to oncology, cardiology, infectious disease, and immunology (Lagasse et al., 2017). It has been

predicted that the combination of aging worldwide population and efficient platform-based approaches will continue to fuel the growth of the antibody drug market (Ecker et al., 2015).

In drug development and antibody engineering, knowledge of the antibody structure is critical information. When developing a new biotherapeutic agent, a high-quality structure would allow for rational design and further optimization of the candidate, such as increasing binding affinities through guided-mutations (Marks & Deane, 2017). However, it is often time-consuming and expensive to improve affinity via experimental approaches such as combinatory mutagenesis (Maynard & Georgiou, 2000) (Sulea, Vivcharuk, Corbeil, Deprez, & Purisima, 2016). Therefore, computational alternatives or methods that can expedite experimental approaches are highly desirable. In terms of antibody structure prediction, five out of the six hypervariable CDR loops have limited structural diversity and can be modeled with canonical structures (Chothia & Lesk, 1987), the only exception is the structurally diverse H3 loop (Marks & Deane, 2017; Weitzner, Dunbrack, & Gray, 2015). The determination of the H3 loop naturally becomes the most challenging part of antibody modeling. However, even with the improvement in the accuracy of H3 loop prediction in the recent years, the computational methods still need further enhancement to support rational design effectively. Canonical modeling of the other five CDRs has reached sub-angstrom level of accuracy, while the accuracy of many H3 loop methods, including computationally expensive ones, can range from 1.5 Å to 3 Å, or even worse (Marks & Deane, 2017).

In our research, we have attempted to predict H3 loop conformational change

upon direct introduction of mutations, using molecular dynamics (MD) (Bowers et al., 2006) and low mode (LMOD) search (Kolossváry & Guida, 1996). Based on our experience, we could not predict the loop conformational change if the mutation introduced too much perturbation to the overall structure of the H3 loop. This chapter describes our effort to model antibody-antigen interactions, but is limited to antibodies where the H3 loop is generally rigidified. Moreover, we presumed that H3 loop rigidification and antibody affinity maturation go hand-in-hand (Schmidt et al., 2013). Instead of running lengthy MD simulations in order to capture accurate conformational change in H3 within the biologically relevant time scale, we focused on high-affinity antibodies where the mutations supposedly did not cause extensive perturbation in loop conformation. More specifically, we modeled the mutations with side chain packing and rotamer selection, leaving the H3 backbone essentially unchanged. Then we applied rigid body docking of the antigen to antibody models, in order to correlate the number of docked near-native poses with reported binding affinities. The formulation of this approach will be further explained in the next section. The aim for the work in this chapter is to establish that there is a numerical relationship between docking results and experimentally measured binding free energies, limited to high-affinity antibodies where the mutations are not altering the H3 loop backbone extensively. The goal is to establish that docking can predict the impact of mutations on rigidified H3 loops, in the form of a change in binding free energies.

5.2 Methods

5.2.1 Data Selection and Preparation

The data used in our docking studies originated from two publications by Purisima and colleagues (Sulea et al., 2016; Vivcharuk et al., 2017). In the first article, the researchers curated a dataset comprised of seven antibody-antigen systems called Single-Point Mutant Antibody Binding (SiPMAB) (Sulea et al., 2016). This data set was carefully assembled to contain 212 single-point mutations of antibodies in their CDRs, each labeled with binding affinity data reported in literature. These systems all came with PDB IDs of the parent (wildtype) antibody-antigen complexes. In some cases, multiple types of experiments were performed to measure binding affinities in multiple publications. When we computed correlations, we included only single-source data, meaning that data reported using different experiments or in different publications were considered as different data series. Purisima and colleagues also deposited their own experimental data in another paper published in 2017. They provided three sets of surface plasmon resonance (SPR) data from testing their Assisted Design of Antibody and Protein Therapeutics (ADAPT) platform, which was developed to aid the selection of mutations that can improve antibody binding affinities (Vivcharuk et al., 2017). We considered these two data sets as good starting points because wildtype antibody-antigen structures and high-quality binding affinities were available. The original ADAPT data set contained single-, double- and triple-point mutants, but we only included single-point mutants in our study, in consistency with the SiPMAB set and the rigid loop assumption. A summary of the cases in the SiPMAB and ADAPT data sets are presented in Table 5.1.

Table 5.1: Summary of antibody-antigen cases in the data sets by Purisima and colleagues. These cases were reported in two publications which aimed to computationally improve antibody binding affinities (Sulea et al., 2016; Vivcharuk et al., 2017). Original names for the datasets are listed below. The system name, PDB IDs, resolution, and experimental assays that measured the binding are reported. There were also different types of mutations. Some data series included mutations to alanine only.

antibody	antigen	PDB ID	resolution (Å)	binding assay	mutations
SiPMAB (Sulea et al., 2016)					
A4.6.1	VEGF	1bj1	2.4	SPR, ELISA	mostly to alanine
D1.3	HEWL	1vfb	1.8	SPR, ITC, ELISA	various
HyHEL-10	HEWL	1c08	2.0	ITC, spectro-photometric	various
Herceptin	HER2	1n8z	2.5	ITC	mostly to alanine
A6	INFγR	1jrh	2.8	SPR	to alanine only
D1.3	E5.2	1dvf	1.9	SPR	to alanine only
HyHEL-63	HEWL	1dqj	2.0	SPR	to alanine only
ADAPT test cases (Vivcharuk et al., 2017)					
bH1	VEGF	3bdy	2.6	SPR	various
bH1	HER2	3be1	2.9	SPR	various
Herceptin	HER2	1n8z	2.5	SPR	various

Because we wanted to focus on high-affinity antibody mutants that have relatively rigidified H3, we further narrowed down the mutants that should be included in docking experiments. Following guidelines from a previously published protein binding benchmark (Kastritis et al., 2011), we specifically eliminated low affinity mutants with measured $K_D > 1\ \mu M$. We also visually inspected the locations of the mutations, and observed that almost all are within 10 Å from the antigen. We used the provided K_D data to calculate the experimental binding free energies with the formula $\Delta G = -RTlnK_D$,

assuming the gas constant $R = 1.987 \times 10^{-3}\ kcal \cdot K^{-1} \cdot mol^{-1}$ and the temperature $T = 298\ K$. If a ratio of K_D^{mutant}/K_D^{WT} was provided instead of absolute values of K_D, the K_D^{mutant} was converted based on the absolute values of K_D^{WT} provided by the original publications.

We conducted very minimal structure preparation of the wildtype antibody. Before docking, we manually removed constant domains of the antibodies. If the structure of the antigen was deemed too big and thus inappropriate for focused docking, which will be described later in Section 5.2.3, the part far from the interface was also removed.

5.2.2 Side Chain Packing

For all systems, we started with the structures of the wildtype antibody-antigen complexes, downloaded from Protein Data Bank (PDB). To generate mutant structures, we used the side chain packing program SCWRL4 developed by the Dunbrack group at Fox Chase Cancer Center (available at http://dunbrack.fccc.edu/scwrl4/). SCWRL4 is a fast and accurate protein side chain conformation prediction program, which uses a new backbone-dependent rotamer library. The program also considers averaging over samples of conformations, hydrogen bonding and soft van der Waals atom-atom interactions (Krivov, Shapovalov, & Dunbrack, 2009). The input for SCWRL4 were simply the starting wildtype antibody structure and the sequence of the desired mutant. The output was a PDB-formatted file of the mutant ready for docking in the next step.

5.2.3 Rigid Body Docking Using a Box

PIPER, as described in Section 1.2.1, is a rigid body protein docking program. Since we already had the crystal structures of the bound antibody-antigen complexes (see Table 5.1), we were not interested in using exhaustive sampling to predict the binding interface. Instead, we adopted a focused sampling strategy in order to address the impact of mutation happening at the known interface. Therefore, we utilized a focused sampling strategy, where a box was used to restrict docking results within close proximity to the original crystal antigen bound to the antibody. The focused docking box was customized to be a cube with the diagonal equal to 75% of the antigen diameter. Based on our experience, this type of box produced better sampling of the antigen binding site. In addition, because of the assignment of the asymmetric pairwise statistical potentials (ADARS) in PIPER (Brenke et al., 2012), for all studies we consistently treated the antibody as the receptor and the antigen as the ligand. Near-native 'hits' were calculated based on how many docked poses had interface root-mean-square deviation (iRMSD) < 10 Å when compared to the ligand in the original crystal structure, among the 1000 poses with the lowest PIPER energies.

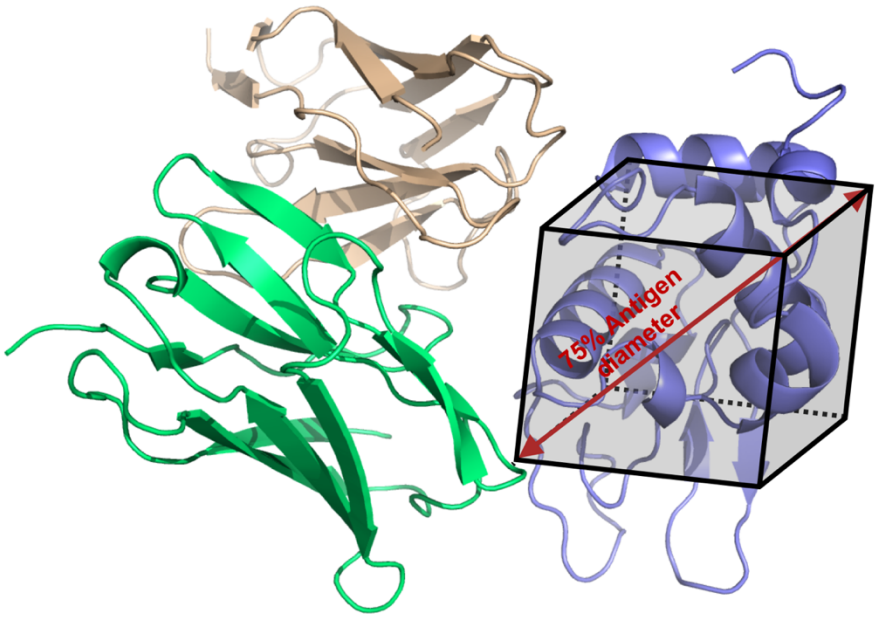

Figure 5.2: Focused docking of antigen to antibody using a box. The example (PDB ID: 1dqj) shows the antigen (purple cartoon) poses were restricted in a box, with diagonal equal to 75% of the ligand diameter. This strategy ensured the final poses of the antigen would stay at the original interface that binds to the antibody (green and wheat cartoon).

5.2.4 Connecting Docking and Binding Free Energy

The computational chemistry and biophysics community has been interested in quantitative prediction of the thermodynamics in protein-protein interactions for more than two decades. The approaches widely range from empirical scoring functions (Vajda, Sippl, & Novotny, 1997; Vajda, Weng, Rosenfeld, & Delisi, 1994; Weng, Delisi, & Vajda, 1997) to lengthy simulations (Jiang et al., 2014; Klimovich, Shirts, & Mobley, 2015) such as implementing multiple copy algorithms for molecular dynamics. Examples include, but are not limited to, molecular mechanics Poisson–Boltzmann or generalized Born surface area continuum solvation methods (MM/PBSA and MM/GBSA) (Genheden & Ryde, 2015), free energy perturbation (FEP) (Shivakumar et al., 2010), and MELD

(Modeling Employing Limited Data) -accelerated molecular dynamics (Morrone, Perez, MacCallum, & Dill, 2017). Based on previous work with ClusPro docking, we have formulated the idea that a high affinity complex will generate more near-native docked hits than a lower affinity complex (Kozakov et al., 2013; Yueh et al., 2017). To further extend this theoretical basis, the following formulation shows how docked near-native hits can be connected to the free energy of antibody-antigen binding. The same derivation has been described by Dr. Kathryn Porter (K. A. Porter, 2019).

$$F = -RTlnQ \quad [5.1]$$

$$Q = \sum_j e^{-\frac{E_j}{RT}} \quad [5.2]$$

$$Q \approx Ne^{-\frac{E}{RT}} \quad [5.3]$$

$$F \approx -RTlnN + E \quad [5.4]$$

$$\Delta G = \Delta F \approx -RT\, ln\frac{N}{N_{ref}} + (E - E_{ref}) \quad [5.5]$$

$$\Delta G = -\alpha lnN + \beta E + \gamma \quad [5.6]$$

The derivation starts with connecting the partition function Q to the Helmholtz free energy F [5.1]. The partition function describes a system consisting of three sub-states: antibody mutant-antigen complexes, wildtype antibody-antigen complexes and a dilute mixture of the free antigen and antibody. The system is also a canonical ensemble consisting of j microstates where each respective microstate has energy E_j [5.2]. We restrict our consideration of the system to those at the near-native interface (as described in the previous section, poses with iRMSD < 10 Å were considered near-native hits).

Those low energy poses docked elsewhere than the original interface are subsequently regarded as transition states or simply false positives. We also presume that high energy microstates make smaller contribution to the partition function. Consequently, the resulting Q could be written by including the most dominant N low-energy microstates. Additionally, these low energy poses already lie within a narrow energy well relative to the whole energy landscape and such a narrow range is comparable to the error of the energy function, it is then acceptable to assume that the microstates have energy values roughly equal to each other (i.e. $E_i \approx E$). The approximation $E_i \approx E$ is consistent with the ClusPro strategy behind clustering by population, which has repeatedly demonstrated success (Kozakov et al., 2013) (Vajda et al., 2017) (Kozakov et al., 2017). This assumption eventually leads to equation [5.3], including terms N and E. More specifically, the term N is the number of near-native hits, defined by counting poses within 10 Å iRMSD from the wildtype antibody-antigen in crystal form. The term E is the energy values PIPER outputs for docking, which has components of van der Waals, electrostatic interaction and statistical pairwise potentials (Kozakov et al., 2006). Connecting back to equation [5.1] establishes the linear relationship between Helmholtz free energy F, N and E [5.4]. In laboratory settings where binding affinities are measured, the volume of the liquid phase is generally constant. Therefore, we can introduce ΔG into equation [5.5]. The reference state of the Gibbs free energy is an unspecified theoretical one, which we note accordingly in [5.5] and combine as a constant value γ in [5.6]. Since there is only one single partition function Q encompassing both wildtype and mutant antibodies, it is reasonable to assume that γ is

identical for both wildtype and mutants. As explained in Section 5.2.3, we used a box to perform local sampling, and thus a scaling factor α was applied instead of using the term $-RT$ directly. In a similar fashion, β is applied to E to correct for scaling and uncertainty (see [5.6]). During sampling, we applied a range of rotational angles and weight coefficients in generating docking poses. We concluded that a set of these docking parameters were most suitable for this formulation. The average results and respective standard deviations (displayed as error bars in all figures) are therefore presented in Section 5.3.

5.2.5 MD Ensemble Docking

As described in Section 5.1, we did not aim to predict drastically changed H3 loop conformation. Instead, we aimed to gain insight into how mutations at the antibody-antigen interface may impact binding free energy, assuming the H3 loop is rigid for high affinity mutants. Simply introducing mutations into the H3 loop without optimizing the loop conformation is not the most rigorous way of modeling mutations. In order to reduce apparent strains, clashes and unfavorable interactions, we also experimented with relatively short MD simulations to apply perturbation and relaxation to the mutated loop. We presumed the high-affinity mature antibodies generally acquired the same H3 loop conformation as the wildtype, but some perturbation could more realistically represent the loop dynamics. Since MD simulations are computationally expensive and time consuming, we only applied such method to a few systems. One case, PDB ID 1vfb, will be discussed in this chapter. For each mutant, ten rounds of 100 ns MD simulations were

performed, each seeded with a random initial velocity. Dr. Istvan Kolossvary ran all MD simulations with Desmond (Bowers et al., 2006), using the AMBER99SB-ILDN (Lindorff-Larsen et al., 2010) force field. The ensemble docking strategy was very similar to what was described for PSTI in Chapter 4 Section 2 (Figure 4.3). First, to extract representative structures from the MD simulations, the ten trajectories of 100 ns MD simulations were concatenated and clustered. One special note here is that the RMSD-based clustering was carried out based on C-alphas of the H3 loop and a few surrounding residues. The H3 region was annotated with the Kabat scheme (Wu & Kabat, 1970). Based on the distribution of pairwise RMSD values, clustering radii ranging from 0.45 Å to 0.6 Å were applied to produce cluster centers. Second, the representative MD snapshots, i.e. the cluster centers, were extracted from the MD trajectories and used as individual receptors for rigid body docking. The last step was to calculate the near-native hits and the weighted average values of N and E, corresponding to equation [5.6]. The probabilities p_i for each snapshot were calculated as the cluster size divided by the total number of frames, where N_i and E_i were the values retrieved for each MD snapshot. The weighted average values were calculated for m snapshots (i.e. m cluster centers) for each antibody mutant-antigen docking in the following fashion.

$$N = \sum_{i=1}^{m} p_i \times N_i \qquad [5.7]$$

$$E = \sum_{i=1}^{m} p_i \times E_i \qquad [5.8]$$

5.3 Results and Discussions

5.3.1 Rigid Body Docking Case Studies

This section details several example cases, which showed consistent performance in terms of the numerical correlation between docked near-native hits and experimentally-measured binding free energies across the set of docking parameters. Following a description of the particular system and mutation locations, plots of docking-predicted ΔG vs. experimentally-derived ΔG will be presented to demonstrate the correlation. Note that the docking-predicted ΔG is the average value of all predictions obtained from ten sets of docking parameters, hence the error bars in the plots. The details of the ten sets of docking parameters are specified in Table B.1. The weight coefficients varied, but a few commonalities were: 1) they all included the contribution from the ADARS pairwise potentials (Brenke et al., 2012; Kozakov et al., 2017) and 2) the weights for the electrostatic interactions were relatively high. The optimal rotation angles seemed to be under 40 degrees (see Table B.1). In total, four systems showed consistent correlation using the above-mentioned sets of docking parameters. The rest of the cases mentioned in Table 5.1 failed to generate satisfactory correlation across a set of docking parameters. For clarity, the cases presented below will be numbered.

(1) Docking of antigen HEWL to antibody HyHEL-63 (wildtype PDB ID: 1dqj) (Li, Li, Smith-Gill, & Mariuzza, 2000). In addition to the wildtype antibody, six other mutants were included in the docking experiments based on the criteria mentioned in Section 5.2.1. The original binding affinity data was obtained by SPR (T = 25 °C, pH=7.5 and salt concentration = 0.150 M) and published in 2003 (Li, Urrutia, Smith-Gill, &

Mariuzza, 2003). The locations and specific mutations are summarized in Figure 5.3. All six mutants were located at the antibody-antigen binding interface and involved single residue mutations to alanine. After we computed docked near-native hits and obtained PIPER energies (N and E, respectively in equation [5.6]), we used multiple linear regression to fit N and E with experimentally-derived ΔG. Subsequently, using the coefficients from multiple linear regression we were able to calculate docking-predicted ΔG. In this case, the linear relationship was established with $R^2 = 0.60$ (Figure 5.4). Further details from docking, including number of near-native hits and PIPER energies, are reported in Table B.2.

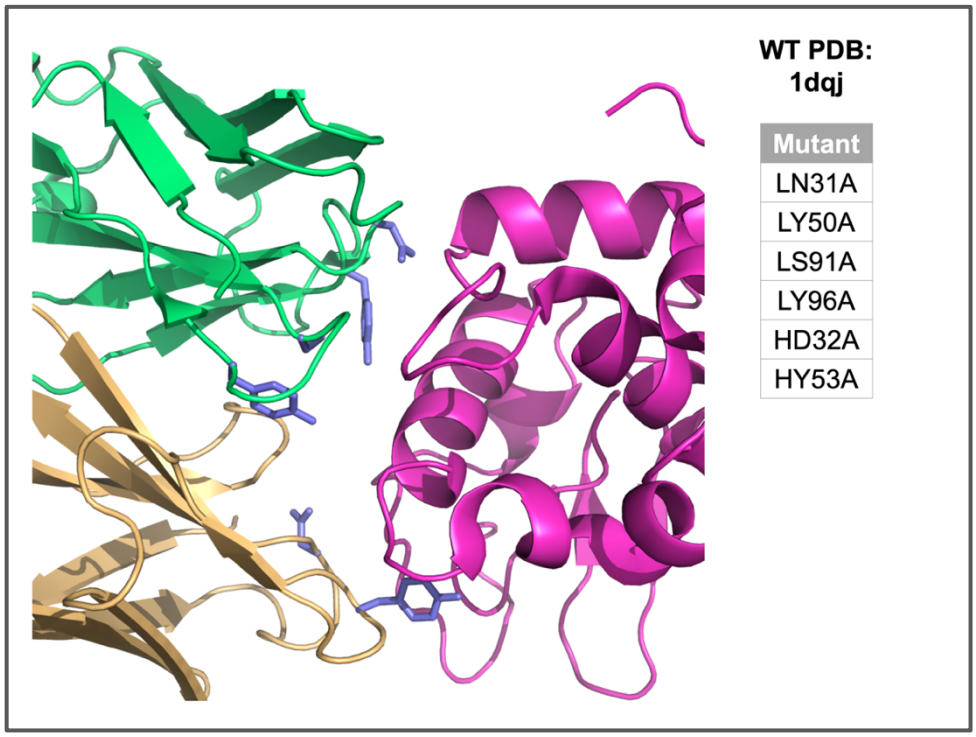

Figure 5.3: Interface of antibody HyHEL-63 and antigen HEWL (wildtype PDB ID: 1dqj). The mutations are located at the interface of the antibody (light chain: lime green, heavy chain: light orange, mutation sites: purple sticks) and the antigen (magenta). The six mutations are listed in the table at the upper right corner. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LS91A" means light chain, serine at position 91 mutated to alanine).

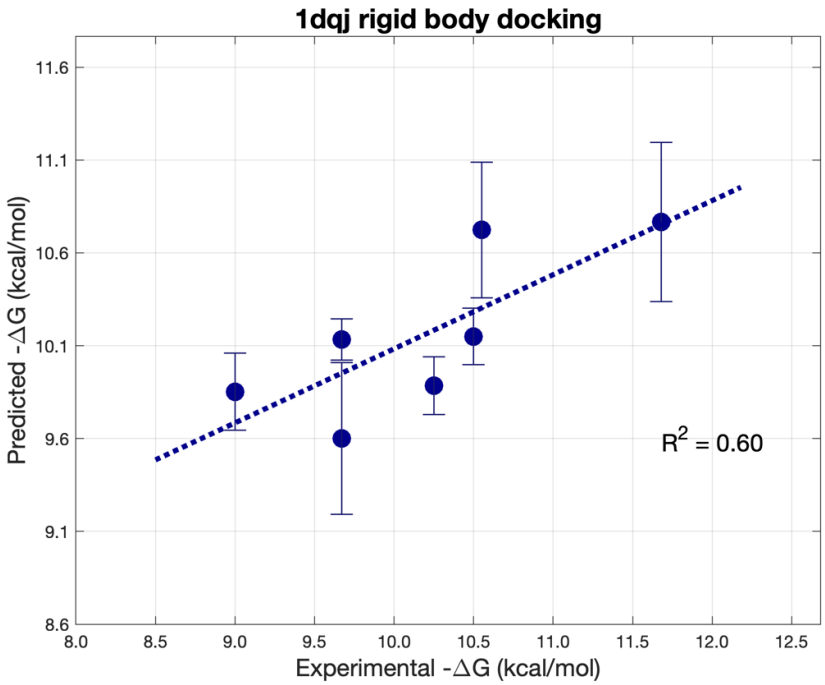

Figure 5.4: Docking-predicted $-\Delta \boldsymbol{G}$ plotted against experimentally measured $-\Delta \boldsymbol{G}$ for HyHEL-63 and HEWL (wildtype PDB ID: 1dqj). The coefficient of determination is $\boldsymbol{R^2 = 0.60}$. The standard deviation across the ten sets of parameters are displayed as error bars.

(2) Docking of antigen human HER2 to Herceptin Fab (wildtype PDB ID: 1n8z) (Cho et al., 2003). In addition to the wildtype antibody, 11 other mutants were included in the docking experiments. The experimental data was obtained using SPR at 25 °C. For each mutant, fold improvement (i.e. K_D^{mutant}/K_D^{WT}) was calculated. The wildtype dissociation constant K_D was reported to be 58 $\pm$ 18 pM (Vivcharuk et al., 2017). The ΔG values in kcal/mol were then derived accordingly. The locations and specific mutations are summarized in Figure 5.5. The specific mutations varied from positively charged lysine to hydrophobic tryptophan, but they were located in only four positions: two on the heavy chain and two on the light chain. Similarly, we calculated docking-predicted ΔG and generated the plot of predicted $-\Delta G$ vs. experimentally measured $-\Delta G$ (Figure 5.6),

with $R^2 = 0.68$. As shown in the plot, the data points are distributed in two clusters. In each cluster, there appears to be one "outlier" data point deviating from the fitted line, which are the wildtype antibody and the LS50W mutant. Detailed docking results can be found are in Table B.3.

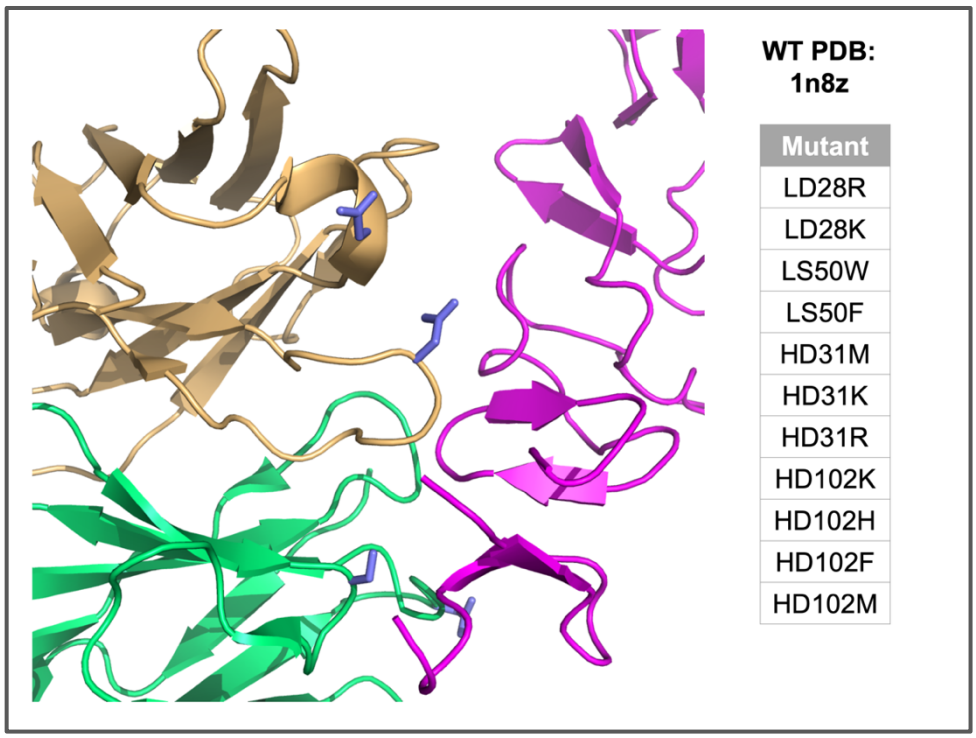

Figure 5.5: Interface of antibody Herceptin and antigen HER2 (wildtype PDB ID: 1n8z). The mutations are located at the interface of the antibody (light chain: lime green, heavy chain: light orange, mutation sites: purple sticks) and the antigen (magenta). The 11 mutations are listed in the table at the upper right corner. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LD28K" means light chain, aspartic acid at position 28 mutated to lysine).

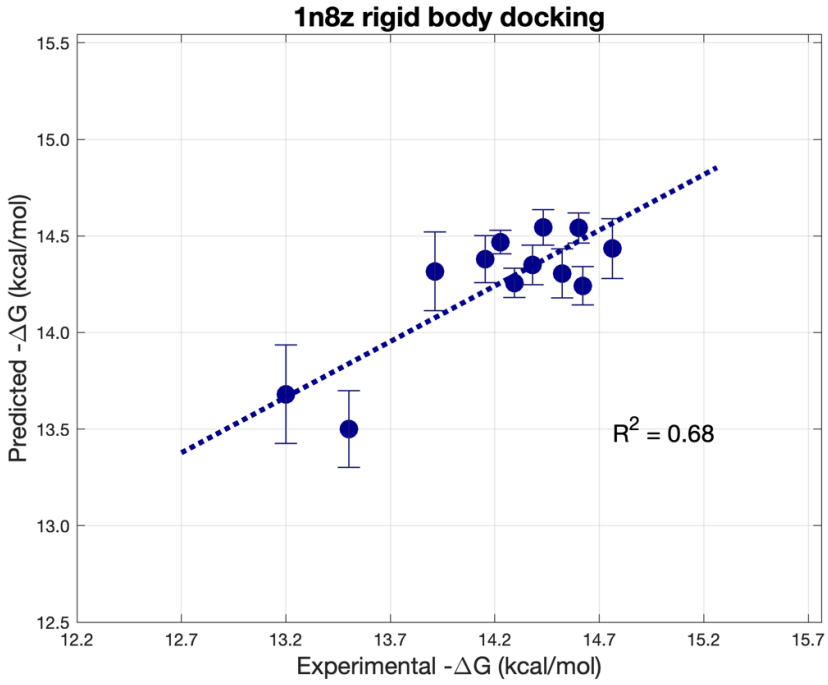

Figure 5.6: Docking-predicted $-\Delta \boldsymbol{G}$ plotted against experimentally measured $-\Delta \boldsymbol{G}$ for Herceptin-HER2 (wildtype PDB ID: 1n8z). The coefficient of determination is $\boldsymbol{R^2 = 0.68}$. The standard deviation across the ten sets of parameters are displayed as error bars.

(3) Docking of antigen HER2 to bF1 Fab (wildtype PDB ID: 3be1) (Bostrom et al., 2009). Including the wildtype antibody, 14 antibodies were docked. The experimental SPR data was provided in the ADAPT paper (Vivcharuk et al., 2017), same as the Case (2). For each mutant, the ratio of K_D^{mutant}/K_D^{WT} was provided instead of absolute values of K_D. The wildtype dissociation constant K_D was reported to be 3.6 $\pm$ 0.6 nM. The corresponding ΔG values in kcal/mol were then derived assuming T = 25 °C. The locations and specific mutations are summarized in Figure 5.7. Eight out of the 13 mutations were located on the heavy chain of the antibody, including five mutations all involving the same aspartic acid at position 98. The linear relationship between docking-predicted $-\Delta G$ and experimentally measured $-\Delta G$ is shown in Figure 5.8, with $R^2 =$

0.58. From the plot, it is evident that the data points can be grouped into one large cluster of 13 mutants plus one single mutant (HY33R) with a much lower value of experimentally measured $-\Delta G$. Removing the mutant HY33R from the data series reduced R^2 to 0.27. This case showed that our method was insufficient to establish the linear correlation in a group of structures with similar binding affinities. Further details from docking are reported in Table B.4.

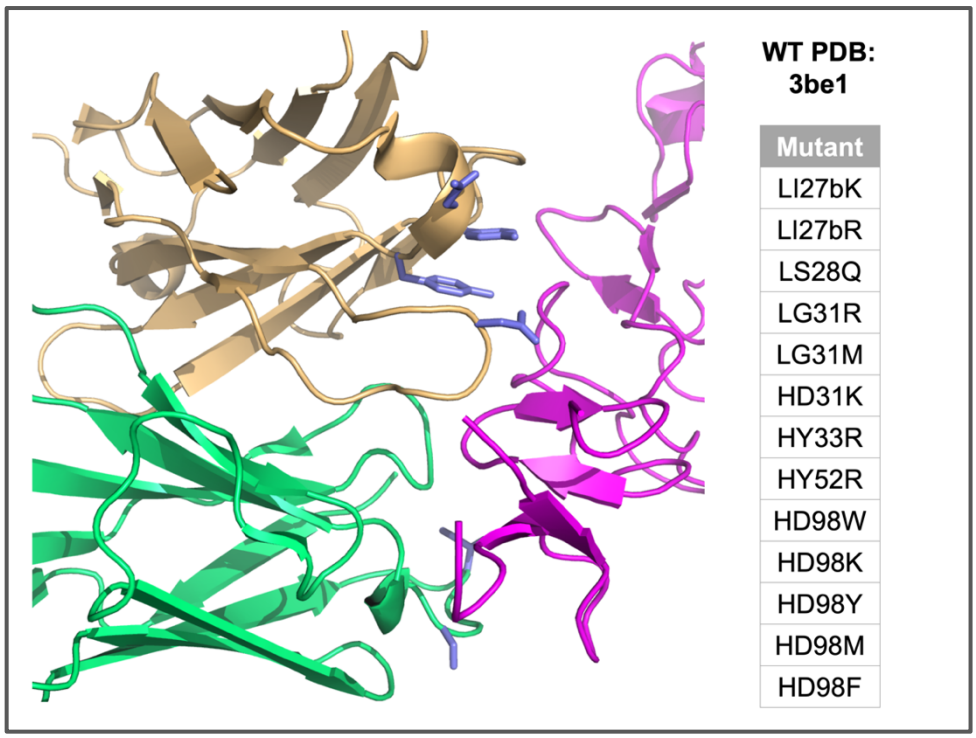

Figure 5.7: Interface of Fab bH1 and antigen HER2 (wildtype PDB ID: 3be1). The mutations are located at the interface of the antibody (light chain: lime green, heavy chain: light orange, mutations sites: purple sticks) and the antigen (magenta). The 13 mutations are listed in the table on the right. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid. The residue numbers follow the numbering in PDB structures (i.e. "27b").

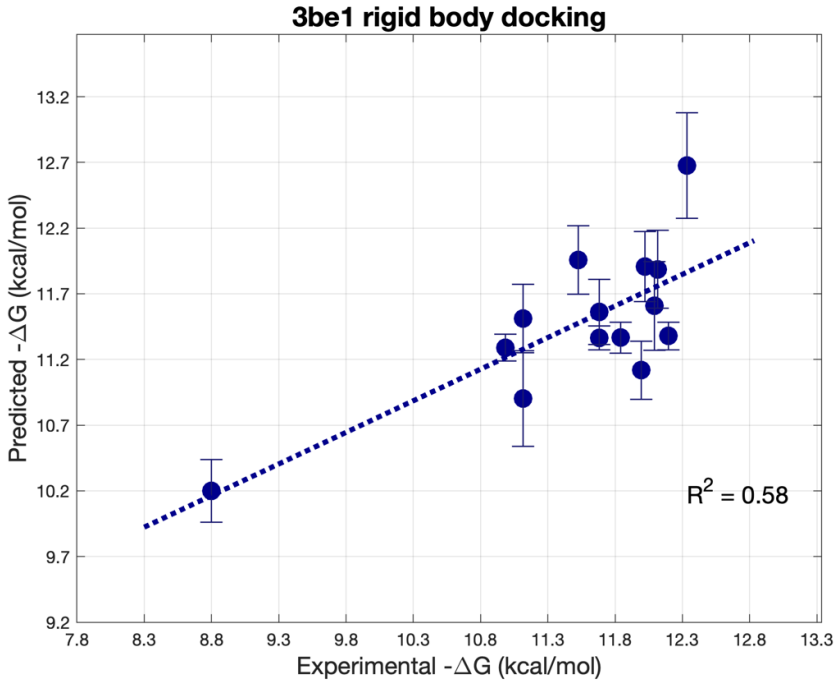

Figure 5.8: Docking-predicted $-\Delta \boldsymbol{G}$ plotted against experimentally measured $-\Delta \boldsymbol{G}$ for bH1-HER2 Fab (wildtype PDB ID: 3be1). The coefficient of determination is $\boldsymbol{R^2 = 0.58}$. The standard deviation across the ten sets of parameters are displayed as error bars.

(4) Docking of lysozyme to Igg1-Kappa D1.3 Fv (wildtype PDB ID: 1vfb) (Bhat et al., 1994). In addition to the wildtype antibody, six other mutants were included in the docking. The original binding affinity data was obtained by SPR (T = 25 °C, pH=7.5 and salt concentration = 0.150 M) and published in 1998 (Dall'Acqua et al., 1998). The locations and specific mutations are summarized in Figure 5.9. These six mutations were all mutations to alanine except one mutation to phenylalanine at various positions in both heavy and light chains. In this case, the linear relationship between docking-predicted and experimental $-\Delta G$ was established with merely $R^2 = 0.41$ (Figure 5.10). Despite the low R^2 in the plot, we still considered this system because we were filtering for R values when selecting representative cases. There is also an "outlier" point LY50A, for which

our method heavily under-predicted binding free energy even considering the error bars (predicted $-\Delta G = 9.55\ kcal/mol$ and experimental $-\Delta G = 10.25\ kcal/mol$). By removing LY50A from the data series, the new R^2 increased to 0.82. Detailed docking results including values of near-native hits and PIPER energies can be found in Table B.5.

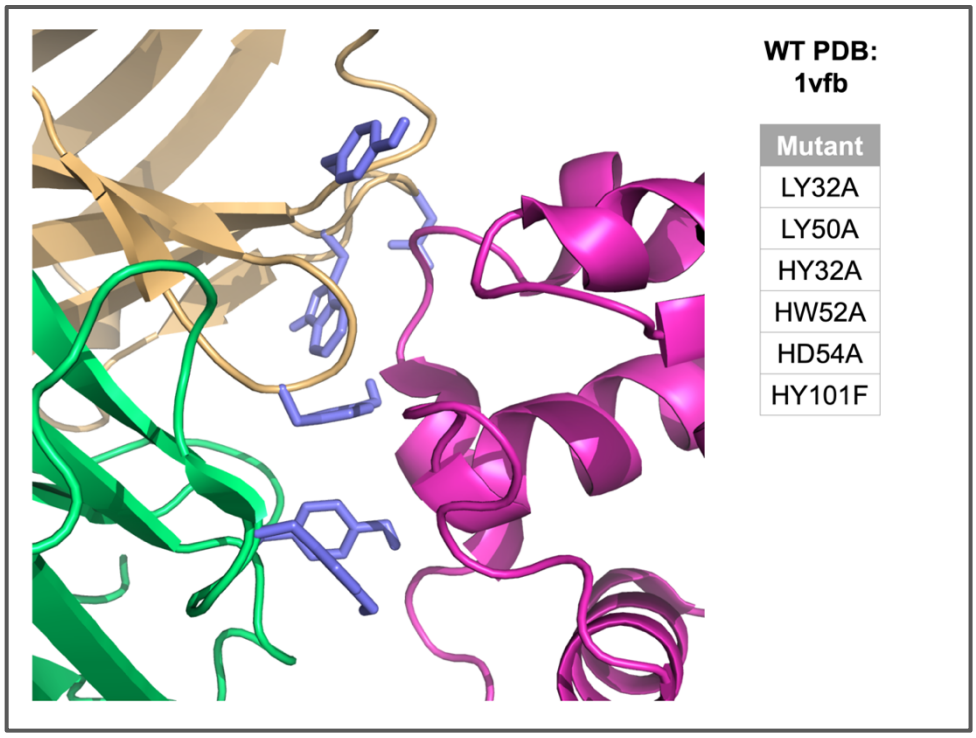

Figure 5.9: Interface of D1.3 Fv and antigen lysozyme (wildtype PDB ID: 1vfb). The mutations are located at the interface of the antibody (light chain: lime green, heavy chain: light orange, mutation sites: purple sticks) and the antigen (magenta). The six mutations are listed in the table on the top right. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid.

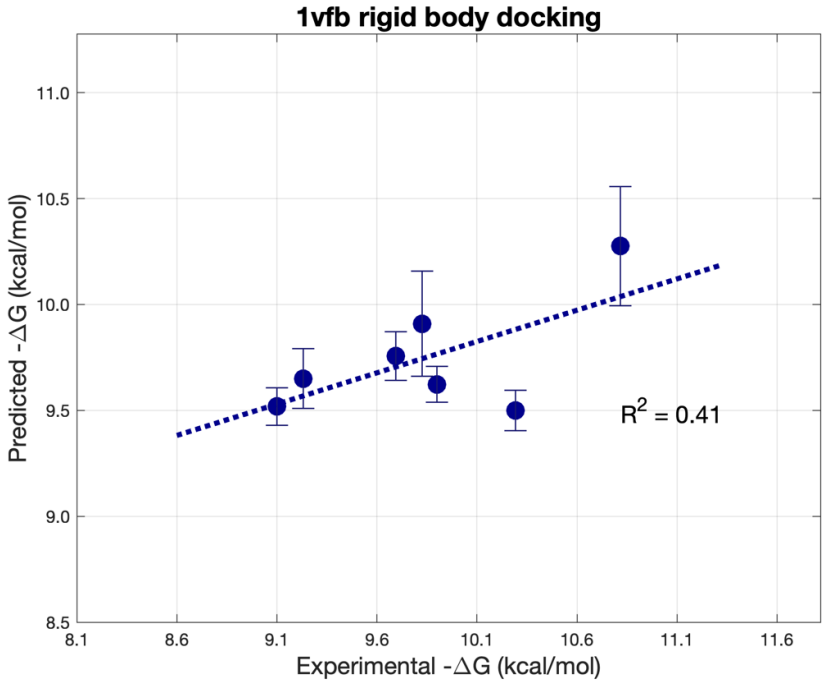

Figure 5.10: Docking predicted $-\Delta \boldsymbol{G}$ plotted against experimentally measured $-\Delta \boldsymbol{G}$ for D1.3 Fv-lysozyme binding (wildtype PDB ID: 1vfb). The coefficient of determination is $\boldsymbol{R^2 = 0.41}$. The standard deviation across the ten sets of parameters are displayed as error bars.

5.3.2 Docking of Lysozyme to Antibody MD Ensemble

As shown in Figure 5.10, the linear relationship between docking-predicted and experimentally measured $-\Delta G$'s is far from being strong ($R^2 = 0.41$). We speculated this case failed because mutations were deviating the interface conformation away from that in the wildtype crystal structure. Fixing the backbone of the H3 loop could become too restrictive in this case. On the other hand, we were not confident that lengthy MD simulations would lead to better solutions. Therefore, we attempted relatively short MD (100 ns) to refine or introduce perturbation around the crystal form of the H3 loop, in order to generate more realistic loop dynamics. Representative snapshots were extracted from the MD simulations, and ensemble docking was performed. We calculated the

average near-native hits weighted by cluster sizes of the MD snapshots. The PIPER energies were processed in a similar fashion. Finally, multiple linear regression analysis was used to fit the weighted average N and E to experimentally determined $-\Delta G$ values. The linear relationship between docking-predicted and experimental binding free energy is shown in Figure 5.11. To our surprise, the coefficient of determination was even lower than before. The range of predicted $-\Delta G$ in fact became narrower (about half of what was shown in Figure 5.10), meaning the MD perturbation dampened the differences between mutants. On the other hand, the error also seemed to become smaller. It was possible that MD simulations mitigated the arbitrariness of the docking parameters and allowed for more robust performance across a set of docking parameters. More details of the docking results are included in Table B.6.

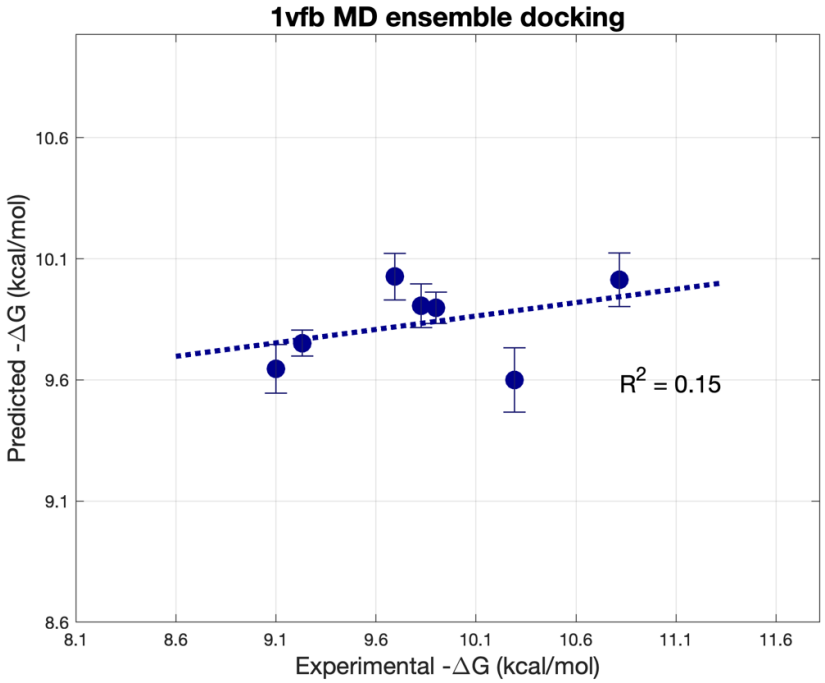

Figure 5.11: MD ensemble docking-predicted $-\Delta \boldsymbol{G}$ plotted against experimentally measured $-\Delta \boldsymbol{G}$ for D1.3 Fv binding to lysozyme (wildtype PDB ID: 1vfb). The coefficient of determination is $\boldsymbol{R^2 = 0.15}$. The standard deviation across the ten sets of parameters are displayed as error bars.

5.4 Conclusions

Computational prediction of antibody binding affinity will be a useful tool for both academic research and biologics design in industry. Protein docking is a much faster alternative to the computationally expensive methods such as MD simulations or FEP. We have established a theoretical basis for using docked near-native hits and PIPER energies to predict binding free energy, ΔG. In this chapter, we focused on high-affinity antibody series, under the premise that the most flexible CDR H3 loop has rigidified in these cases. This assumption lead to our approach of minimally modeling the mutations, except for side chain packing. The Purisima research group summarized a number of high-quality cases in their publications (Sulea et al., 2016; Vivcharuk et al., 2017), which we used as the starting point for testing our docking-based method. We were able to find a few systems that consistently demonstrated linear relationship between docking results and experimentally measured ΔG across a set of docking parameters. Nonetheless, quite a few systems only performed well under very specific rotation angles and weight coefficients, showing that the robustness of the method is limited. Furthermore, a few systems displayed two clusters of points in their plots (Figure 5.6 and Figure 5.8), and removing one cluster impaired the correlation as shown in Case (3) in the previous section. This phenomenon pointed to the "catch 22" nature of our method: we needed a small range of binding affinities in order to support our assumption of minimal modeling the loop, but a small range of binding affinities were also more challenging to establish linear relationship. The docking parameters presented in this chapter were not optimized for antibody binding affinity prediction, and the margin of error was large as shown in the

plots. In addition, PIPER energy was by no means tailored for accurately predicting protein-protein binding free energy, as we informed our users (Kozakov et al., 2017). It also became evident that the systems had different types of mutations in terms of locations and biochemical properties. Naturally, one would question if it was fair to compare different series of mutations under the same conditions. We also found mutations that were “outlier” data points in the data series. Removing such data points seemed to improve the linear correlation, but we were unsure why these mutants became “outliers”. Consequently, it is necessary to re-think the applicable cases for docking-based binding affinity prediction. Ideally, it would also be essential to systematically optimize a set of docking parameters on a large number of cases.

Appendix A: Supplemental Information for Chapter 2

Table A.1: List of 32 proteins used for analysis in Chapter 2. The columns include the PDB and chain for the apo structures, holo structures (i.e. "2cm2_B" means we used PDB ID 2cm2 and the chain B), ligand PDB ID, name of the protein and a brief description of the site. The three groups of proteins are separated by thick lines in the table.

Apo	Holo	Lig	Name	Site
2cm2_A	2h4k_A	509	PTP1B	Stronger pTyr binding site.
1pkl_B	3hqp_P	ATP	Pyruvate kinase enzyme	ATP+Oxalate binding site.
1rtc_A	1br6_A	PT1	Ricin	Pteroic acid binding at the active site.
1rhb_A	2w5k_B	NDP	Ribonuclease A.	NADPH, binding at the active site.
3cj0_A	2brl_A	POO	HCV polymerase NS5B	Between fingers and thumb domains
2f6v_A	1t49_A	892	PTP1B	Allosteric site under the C-termin. helix.
1zah_B	2ot1_D	N3P	Fructose aldolase	A competitive inhibitor.
1g24_D	1gzf_C	NIR	Rho ADP-Ribosyl. Enz.	Structure also contain a NAD and ADP
1w50_A	3ixj_C	586	BACE-1 protease	Active site, too open in apo structure.
1bsq_A	1gx8_A	RTL	Bovine Beta-lactoglobulin	Retinol binding in the central cavity.
1hag_E	1ghy_H	121	Thrombin	Pocket is too open, with flexible loops.
3f74_C	3bqm_C	BQM	Alpha-L (Integrin) domain	Active site with disordered C terminus
1my0_B	1n0t_D	AT1	Glutamate receptor 2	Stabilizes the open form of the receptor.
1xcg_B	1ow3_B	GDP	Transforming protein	GDP interacts only with RhoA.
1jwp_A	1pzo_A	CBT	TEM β-lactamase	Allosteric site between two helices.
2bls_B	3gqz_A	GF7	AMPc beta-lactamase	Peripheral allosteric site. Very weak.
2bu8_A	2bu2_A	TF1	Pyruvate dehyd. kinase	Allosteric inhibitor site.
3cj0_A	3fqk_B	79Z	HCV polymerase NS5B	Binding near the active site.
2brk_A	2gir_B	NN3	HCV polymerase NS5B	Non-nucleotide inhibitor (thumb) site.
1fxx_A	3hl8_A	BBP	Exodeoxyribonuclease I	Prevents Exol/SSB interactions.
1ok8_A	1oke_B	BOG	Dengue 2 virus envelope	Site is between two domains.
2aka_A	1yv3_A	BIT	Myosin II	Narrow pocket, planar ligand.
3mn9_A	3eks_A	CY9	Monomer. actin with toxin	Binding to the barbed end of filaments.
1nuw_A	1eyj_B	AMP	Fruct. 1,6-bisphosphatase	AMP binding site.
3puw_E	1fqc_A	GLO	Maltodextrin/maltose BP	Interdomain binding, domain flexibility.
3kqa_B	3lth_A	UD1	MurA dead-end complex	Interdomain binding, loop is closing.
3gxd_B	2wcg_A	MT5	Acid-beta-glucosidase	Active site.
1bnc_B	2v5a_A	LZL	Biotin carboxylase	ATP competitive inhibitor.
1my1_C	1ftl_A	DNQ	Glutamate receptor 2.	Interdomain binding.

2ax9_A	2piq_A	RB1	Androgen receptor	Allosteric inhibitor binds on surface.
2zb1_A	2npq_A	BOG	P38 MAP kinase	Helix 253-261 moves outward.
2air_H	1za1_D	CTP	Asp. transcarbamylase	Binds CTP at the flexible N-terminal

Table A.2: TEM β-lactamase druggability scores (DS) and mutations. The columns include the PDB ID and chain name (i.e. "4mez_B" means PDB ID is 4mez and chain ID is B), protein type ("M" for mutants and "U" for unbound wildtype), the calculated druggability sore (DS) and a brief description of the mutations.

PDB ID	DS	T	Mutation (E. Coli)
4mez_B	0.032	M	(M68L, M69T)
4mez_A	0.037	M	(M68L, M69T)
4ibx_E	0.057	M	TEM v.13 (A42G, N52A, I84V, R120G, M182T, L201A, T265M); T_m = 69.0ºC
1zg6_A	0.076	M	Catalytic residue mutation: S70G; expected to improve stability
3dtm_A	0.129	M	(P62S, V80I, E147G, M182T, L201P, A224V, I247V, R275R) T_m = 69.2 ºC
1jwp_A	0.186	M	Strong stabilization: (M182T, V184A) M182T alone yields T_m = 63.2ºC
1yt4_A	0.237	M	TEM-76: S130G; T_m = 52.3 ºC
1ck3_A	0.325	M	TEM-84: (N276D) T_m = 58.0 ºC
1zg4_A	0.390	U	WT TEM1 beta lactamase, T_m = 58.5 ºC
4gku_A	0.418	M	(I84V, V184A); V184A on its own yields T_m = 58.1 ºC
3toi_B	0.541	M	First 15 residues removed & (I56V, R120G, M182T, T195S, I208M, A224V, R241H,T265M); T_m = 59.0 ºC
1htz_E	0.571	M	TEM52: (E104K, M182T, G238S); T_m = 55.6 ºC
1htz_C	0.599	M	TEM52: (E104K, M182T, G238S); T_m = 55.6 ºC
1htz_B	0.612	M	TEM52: (E104K, M182T, G238S); T_m = 55.6 ºC
4oqg_E	0.629	U	WT TEM-1 beta-lactamase: no ligand in chain E; T_m = 58.5 ºC
1htz_A	0.640	M	TEM52: (E104K, M182T, G238S); T_m = 55.6 ºC
1htz_D	0.640	M	TEM52: (E104K, M182T, G238S); T_m = 55.6 ºC
3toi_A	0.669	M	First 15 residues removed & (I56V, R120G, M182T, T195S, I208M, A224V, R241H,T265M); T_m = 59.0 ºC
1li9_A	0.698	M	TEM-34: (M69V); T_m almost identical to or greater than that of TEM-1
1lhy_A	0.718	M	TEM-30: (R244S); Destabilizing
3cmz_A	0.849	M	(L201P); T_m = 53.4 ºC

Appendix B: Supplemental Information for Chapter 5

Table B.1: The ten sets of docking parameters. This table shows rotation angles and weight coefficients used in PIPER sampling. The weight coefficients are further broken down to details directly relevant to the PIPER program.

param. set number	rotation degree	coefficient set				
		energy term breakdown for PIPER				
		weight 0 repulsion	weight 1 attraction	weight 2 electrostatic (Coulombic)	weight 3 electrostatic (Born)	weight 4 DARS
1	20°	0.40	-0.40	700.00	70.00	1.00
2	20°	0.50	-0.40	700.00	70.00	1.00
3	30°	0.30	-0.20	500.00	50.00	1.00
4	30°	0.40	-0.30	400.00	40.00	1.00
5	30°	0.40	-0.30	500.00	50.00	1.00
6	30°	0.50	-0.30	400.00	40.00	1.00
7	30°	0.50	-0.40	400.00	40.00	1.00
8	40°	0.40	-0.30	500.00	50.00	0.25
9	40°	0.50	-0.40	400.00	40.00	0.25
10	40°	0.50	-0.40	400.00	40.00	1.00

Table B.2: Detailed docking results of antibody HyHEL-63 and antigen HEWL (wildtype PDB ID: 1dqj). The near-native hits, absolute values of PIPER energies and predicted $-\Delta \boldsymbol{G}$ values are reported for all mutants and the wildtype, across the ten sets of docking parameters. For predicted $-\Delta \boldsymbol{G}$, average and standard deviations are also reported. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LS91A" means light chain, serine at position 91 mutated to alanine).

near-native hits (N)							
param. set	LY96A	LD31A	LY50A	LS91A	HD32A	HY53A	WT
1	504	506	453	484	454	452	509
2	499	492	449	476	450	453	514
3	737	678	614	647	597	657	723
4	799	743	675	710	672	682	799
5	828	772	704	739	688	717	828
6	812	763	692	725	696	676	841
7	859	810	744	780	748	726	868
8	960	975	986	969	951	944	975
9	957	965	978	957	945	913	978
10	891	866	913	838	811	737	894
absolute values of PIPER energy (|E|)							
param. set	LY96A	LD31A	LY50A	LS91A	HD32A	HY53A	WT
1	1229.68	1318.84	1165.11	1308.55	1253.32	1091.01	1220.70
2	1188.68	1277.64	1123.91	1266.05	1214.42	1058.01	1193.48
3	801.98	879.22	756.08	865.08	832.59	732.72	824.47
4	967.94	1051.20	919.05	1039.40	1016.48	860.01	981.46
5	980.75	1064.82	924.98	1052.48	1019.99	884.82	994.47
6	926.94	1019.59	889.35	996.90	977.58	810.77	961.36
7	1150.24	1237.69	1087.95	1226.80	1203.88	1018.17	1154.46
8	672.65	724.13	651.57	701.96	660.65	640.58	671.56
9	844.02	902.71	817.67	876.28	844.54	792.83	831.56
10	1150.24	1237.69	1087.95	1226.80	1203.88	1027.54	1154.46
predicted binding free energy ($-\Delta G$)							
param. set	LY96A	LD31A	LY50A	LS91A	HD32A	HY53A	WT
1	10.84	10.31	9.79	9.76	9.24	10.24	11.03
2	10.86	10.20	9.69	9.79	9.26	10.16	11.26
3	11.04	10.31	9.59	9.94	9.31	10.15	10.87
4	11.05	10.15	9.76	9.80	9.40	10.04	11.01
5	11.02	10.21	9.77	9.87	9.34	10.02	10.98
6	11.01	10.05	9.83	9.72	9.47	9.99	11.14
7	10.99	10.07	9.90	9.76	9.46	9.96	11.07
8	10.24	10.13	9.87	10.17	10.35	10.41	10.04
9	10.18	9.93	10.31	10.04	10.16	10.35	10.25
10	10.15	10.11	10.18	10.14	10.17	10.32	10.15
average	10.74	10.15	9.87	9.90	9.61	10.16	10.78
standard deviation	0.37	0.11	0.21	0.16	0.41	0.15	0.43

Table B.3: Detailed docking results of antibody Herceptin and antigen HER2 (wildtype PDB ID: 1n8z). The near-native hits, absolute values of PIPER energies and predicted $-\Delta \boldsymbol{G}$ values are reported for all mutants and the wildtype, across the ten sets of docking parameters. For the predicted $-\Delta \boldsymbol{G}$, average (avg.) and standard deviations (sd.) are also reported. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LD28R" means light chain, aspartic acid at position 28 mutated to arginine).

near-native hits (N)												
param. set	HD102M	LD28K	HD102F	HD31M	HD31K	HD31R	LS50F	LS50W	HD102H	LD28R	WT	HD102K
1	524	530	525	529	530	526	506	511	523	530	527	522
2	530	532	528	531	531	526	517	518	530	532	529	527
3	957	962	974	969	984	948	850	823	940	957	966	929
4	827	886	858	830	858	804	740	724	797	858	828	802
5	938	952	952	946	970	933	842	811	920	944	934	930
6	872	899	879	861	899	838	774	746	835	877	854	850
7	831	878	826	827	851	810	757	731	808	847	825	816
8	988	997	979	991	992	991	990	998	988	998	994	988
9	988	992	965	989	989	988	979	969	987	994	989	985
10	930	964	936	921	940	925	858	827	918	958	912	927
absolute values of PIPER energy (|E|)												
param. set	HD102M	LD28K	HD102F	HD31M	HD31K	HD31R	LS50F	LS50W	HD102H	LD28R	WT	HD102K
1	1441.6	1584.3	1460.7	1444.4	1454.1	1458.0	1492.3	1510.4	1440.6	1583.7	1400.7	1455.6
2	1368.1	1510.7	1387.2	1370.9	1380.6	1384.5	1412.2	1415.6	1367.1	1512.1	1337.3	1382.1
3	1008.1	1066.2	1023.2	1009.6	1016.6	1017.9	1052.9	1057.9	1009.1	1064.6	973.2	1019.2
4	1135.1	1179.9	1145.6	1136.3	1141.9	1142.9	1174.2	1146.7	1135.9	1174.3	1091.0	1143.9
5	1177.6	1235.9	1188.1	1179.0	1186.0	1187.3	1214.8	1193.4	1178.5	1228.1	1132.0	1188.6
6	1074.7	1119.5	1085.2	1075.9	1081.5	1082.5	1108.9	1081.3	1075.5	1110.9	1029.1	1083.5
7	1304.6	1350.6	1316.1	1305.7	1311.3	1312.3	1343.3	1297.2	1305.3	1339.7	1249.8	1313.3
8	809.4	909.9	794.9	830.1	876.6	871.2	802.5	764.9	812.6	907.7	799.8	826.0
9	927.1	1013.2	922.9	940.1	977.5	990.4	932.6	890.1	931.3	1015.5	922.5	947.9
10	1304.6	1350.6	1316.1	1305.7	1311.3	1312.3	1343.3	1297.2	1305.3	1339.7	1249.8	1313.3
predicted binding free energy ($-\Delta G$)												
param. set	HD102M	LD28K	HD102F	HD31M	HD31K	HD31R	LS50F	LS50W	HD102H	LD28R	WT	HD102K
1	14.32	14.45	14.35	14.56	14.59	14.40	13.40	13.63	14.28	14.45	14.51	14.21
2	14.48	14.45	14.31	14.55	14.54	14.17	13.47	13.54	14.48	14.45	14.45	14.25
3	14.38	14.54	14.54	14.48	14.60	14.33	13.60	13.37	14.25	14.49	14.38	14.19
4	14.36	14.65	14.55	14.38	14.56	14.18	13.62	13.56	14.15	14.48	14.48	14.16
5	14.39	14.53	14.50	14.45	14.64	14.35	13.61	13.32	14.24	14.46	14.33	14.33
6	14.46	14.61	14.50	14.37	14.65	14.19	13.64	13.43	14.18	14.46	14.37	14.28
7	14.40	14.71	14.34	14.36	14.55	14.21	13.69	13.51	14.20	14.48	14.43	14.26
8	14.28	14.44	14.74	14.25	14.51	14.54	14.11	13.36	14.31	14.37	13.85	14.40
9	14.12	14.57	14.19	14.19	14.39	14.46	14.19	14.00	14.15	14.57	14.09	14.24

10	14.41	14.60	14.43	14.32	14.49	14.34	13.57	13.37	14.29	14.58	14.39	14.36
avg.	14.36	14.56	14.45	14.39	14.55	14.32	13.69	13.51	14.25	14.48	14.33	14.27
sd.	0.10	0.09	0.15	0.12	0.08	0.13	0.26	0.20	0.10	0.06	0.20	0.08

Table B.4: Detailed docking results of antibody bH1 and antigen HER2 (wildtype PDB ID: 3be1). The near-native hits, absolute values of PIPER energies and predicted $-\Delta \boldsymbol{G}$ values are reported for all mutants and the wildtype, across the ten sets of docking parameters. For the predicted $-\Delta \boldsymbol{G}$, average (avg.) and standard deviations (sd.) are also reported. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LS28Q" means light chain, serine at position 28 mutated to glutamine).

near-native hits (N)							
param. set	LG31R	HY33R	HD98W	LG31M	HD98K	HD98Y	HD98M
1	491	505	485	480	484	463	480
2	496	509	495	484	491	478	487
3	867	878	810	813	817	799	800
4	813	819	777	765	780	760	766
5	869	885	820	818	820	804	807
6	814	815	794	784	783	780	768
7	811	819	789	781	784	773	775
8	980	986	978	973	978	971	973
9	963	975	967	953	957	958	950
10	911	925	867	882	888	884	877
param. set	LI27bR	WT	HD51K	LS28Q	LI27bK	HY52R	HD98F
1	482	496	479	476	490	518	470
2	488	505	483	484	493	521	485
3	836	876	815	812	838	969	810
4	792	828	778	770	801	914	773
5	837	869	817	815	840	956	815
6	803	838	784	776	805	904	787
7	801	839	788	781	800	912	783
8	947	994	973	970	953	983	977
9	936	971	951	953	949	983	964
10	896	898	892	890	900	973	886
absolute values of PIPER energy (|E|)							
param. set	LG31R	HY33R	HD98W	LG31M	HD98K	HD98Y	HD98M
1	1561.60	1452.58	1575.30	1516.68	1538.25	1654.63	1536.01
2	1481.20	1394.54	1510.64	1436.58	1459.95	1569.43	1457.41
3	1184.38	1122.17	1235.28	1151.46	1189.10	1290.59	1188.22
4	1267.87	1201.20	1333.04	1232.56	1289.88	1380.56	1287.87
5	1303.38	1241.97	1361.48	1262.76	1317.60	1408.13	1316.32
6	1222.95	1138.78	1252.04	1195.44	1217.38	1313.34	1215.37
7	1413.85	1321.00	1459.24	1381.44	1418.38	1503.76	1415.97
8	758.33	712.26	742.89	719.79	730.69	749.75	730.24
9	881.45	822.95	849.40	843.25	842.51	866.34	841.78
10	1413.85	1321.00	1459.24	1381.44	1418.38	1503.76	1415.97
param.	LI27bR	WT	HD51K	LS28Q	LI27bK	HY52R	HD98F

set							
1	1528.74	1588.38	1518.83	1513.52	1531.60	1514.92	1596.11
2	1447.97	1504.32	1441.02	1432.55	1451.54	1443.14	1512.81
3	1195.64	1187.74	1193.47	1184.11	1196.60	1162.93	1235.54
4	1294.35	1309.16	1292.61	1288.24	1295.11	1258.50	1322.77
5	1323.54	1341.84	1321.37	1314.51	1324.50	1294.83	1351.20
6	1221.85	1269.06	1220.11	1215.74	1222.61	1186.80	1262.86
7	1422.25	1463.26	1420.51	1418.64	1423.01	1390.40	1453.27
8	724.88	779.04	724.86	717.98	727.49	739.43	740.23
9	841.00	897.21	835.26	834.70	843.13	846.41	850.41
10	1422.25	1463.26	1420.51	1418.64	1423.01	1390.40	1453.27
predicted binding free energy ($-\Delta G$)							
param. set	LG31R	HY33R	HD98W	LG31M	HD98K	HD98Y	HD98M
1	11.68	10.41	11.86	11.24	11.46	12.86	11.46
2	11.62	10.49	11.96	11.32	11.47	12.85	11.50
3	11.43	10.40	12.16	10.73	11.38	13.07	11.32
4	11.22	10.11	12.20	10.44	11.47	12.94	11.39
5	11.26	10.25	12.19	10.56	11.47	12.96	11.44
6	11.52	10.20	11.92	11.00	11.34	12.84	11.26
7	11.42	9.87	12.11	10.77	11.40	12.81	11.33
8	11.97	10.23	11.57	11.06	11.18	12.11	11.40
9	11.96	10.58	11.28	11.60	11.47	11.85	11.67
10	11.45	9.89	12.03	10.75	11.42	12.89	11.33
avg.	11.55	10.24	11.93	10.95	11.41	12.72	11.41
sd.	0.26	0.24	0.30	0.36	0.09	0.40	0.12
param. set	LI27bR	WT	HD51K	LS28Q	LI27bK	HY52R	HD98F
1	11.37	11.95	11.27	11.23	11.35	11.03	12.17
2	11.38	11.74	11.38	11.28	11.35	10.85	12.13
3	11.55	11.51	11.45	11.29	11.57	11.31	12.16
4	11.60	11.98	11.51	11.41	11.64	11.41	12.01
5	11.58	11.90	11.53	11.42	11.59	11.16	12.02
6	11.47	12.32	11.39	11.29	11.49	11.21	12.07
7	11.53	12.35	11.45	11.40	11.54	11.35	11.99
8	12.43	12.01	11.22	11.14	12.23	11.24	11.53
9	12.09	12.00	11.52	11.45	11.72	10.75	11.39
10	11.52	12.25	11.48	11.43	11.56	11.31	12.02
avg.	11.65	12.00	11.42	11.33	11.60	11.16	11.95
sd.	0.34	0.26	0.10	0.10	0.25	0.22	0.27

Table B.5: Detailed docking results of antibody D1.3 Fv and antigen hen egg white lysozyme (wildtype PDB ID: 1vfb). The near-native hits, absolute values of PIPER energies and predicted $-\Delta G$ values are reported for all mutants and the wildtype, across the ten sets of docking parameters. For the predicted $-\Delta G$, average and standard deviations are also reported. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LY32A" means light chain, tyrosine at position 32 mutated to alanine).

near-native hits (N)							
param. set	HW52A	HY101F	HD54A	LY32A	LY50A	HY32A	WT
1	527	530	532	532	524	512	531
2	524	525	531	528	519	503	525
3	708	766	820	897	849	707	813
4	788	816	841	934	870	732	842
5	773	800	833	917	864	727	833
6	762	761	812	901	839	701	819
7	833	836	855	944	878	756	855
8	725	638	666	705	688	667	687
9	788	688	706	748	738	717	756
10	723	696	683	855	773	583	752
absolute values of PIPER energy (|E|)							
param. set	HW52A	HY101F	HD54A	LY32A	LY50A	HY32A	WT
1	976.89	1003.56	1052.73	917.80	937.10	1035.33	1079.57
2	924.02	963.56	1006.63	881.31	900.00	989.53	1024.99
3	658.07	660.95	719.99	607.54	604.99	699.13	736.58
4	774.29	789.13	846.08	738.56	728.69	820.53	862.37
5	778.48	800.37	854.69	738.71	740.59	832.43	871.58
6	743.51	750.64	814.68	711.86	693.09	791.36	839.87
7	900.08	929.83	980.78	871.16	864.29	953.83	997.37
8	486.51	528.84	514.49	492.25	507.07	520.99	545.19
9	600.52	658.29	640.57	611.22	630.76	640.46	679.16
10	900.08	929.83	980.78	871.16	864.29	953.83	997.37
predicted binding free energy ($-\Delta G$)							
param. set	HW52A	HY101F	HD54A	LY32A	LY50A	HY32A	WT
1	9.69	9.81	10.04	9.41	9.50	9.96	10.16
2	9.65	9.83	10.03	9.45	9.54	9.97	10.12
3	9.60	9.71	10.13	9.58	9.50	9.84	10.22
4	9.65	9.76	10.08	9.59	9.49	9.84	10.17
5	9.59	9.75	10.11	9.58	9.52	9.81	10.21
6	9.63	9.67	10.10	9.68	9.48	9.78	10.25
7	9.65	9.79	10.06	9.58	9.50	9.85	10.14
8	9.78	9.39	9.52	9.59	9.72	9.73	10.85
9	9.86	9.52	9.47	9.52	9.74	9.62	10.85
10	9.63	9.76	10.05	9.69	9.50	9.65	10.29
avg.	9.67	9.70	9.96	9.57	9.55	9.81	10.32
sd.	0.08	0.14	0.25	0.09	0.10	0.12	0.28

Table B.6: Detailed MD ensemble docking results of antibody D1.3 Fv and antigen hen egg white lysozyme (parent PDB ID: 1vfb). The weighted average near-native hits, absolute values of weighted average PIPER energies and predicted $-\Delta \boldsymbol{G}$ values are reported for all mutants and the wildtype, across the ten sets of docking parameters. For the predicted $-\Delta \boldsymbol{G}$, average and standard deviations are also reported. The mutation code consists of chain, wildtype amino acid, residue number and new amino acid (i.e. "LY32A" means light chain, tyrosine at position 32 mutated to alanine).

weighted average near-native hits (N)							
param. set	HD54A	WT	HW52A	HY101F	HY32A	LY32A	LY50A
1	470.03	478.81	431.89	490.48	380.15	502.29	481.68
2	455.42	459.58	420.98	477.71	365.53	495.99	473.22
3	780.60	706.49	665.86	842.04	638.10	857.99	838.16
4	825.45	732.33	709.17	860.62	668.16	878.33	874.37
5	817.03	734.33	697.64	861.33	660.24	876.31	866.41
6	804.05	710.04	686.63	836.11	648.57	858.62	861.88
7	848.50	762.65	728.55	874.06	693.99	886.96	888.38
8	721.41	705.03	656.86	728.10	548.62	754.25	723.59
9	747.17	708.88	683.18	750.92	558.04	760.45	740.81
10	770.28	652.93	659.81	794.08	586.56	850.26	841.20
absolute values of weighted average PIPER energy (|E|)							
param. set	HD54A	WT	HW52A	HY101F	HY32A	LY32A	LY50A
1	995.01	996.83	908.05	925.14	947.82	872.40	835.64
2	934.39	943.63	857.56	873.59	882.17	822.53	781.24
3	696.73	709.68	632.89	617.20	671.16	574.06	554.44
4	807.25	820.03	746.16	739.94	780.77	693.60	664.35
5	812.87	825.65	744.44	750.11	788.29	699.07	672.57
6	757.41	780.62	697.50	705.07	743.81	653.49	630.93
7	933.51	944.85	863.02	877.14	901.87	826.53	785.48
8	513.73	511.97	472.03	525.00	488.53	473.77	462.33
9	650.58	637.48	593.89	655.39	605.10	602.66	579.25
10	962.37	945.10	880.16	880.38	903.01	827.66	786.92
predicted binding free energy ($-\Delta G$)							
param. set	HD54A	WT	HW52A	HY101F	HY32A	LY32A	LY50A
1	10.00	10.01	9.74	9.80	9.85	9.64	9.53
2	9.98	10.01	9.77	9.79	9.88	9.63	9.52
3	9.94	10.07	9.91	9.64	10.06	9.50	9.46
4	9.87	10.04	9.91	9.67	10.06	9.54	9.48
5	9.88	10.03	9.90	9.68	10.06	9.54	9.49
6	9.89	10.07	9.87	9.70	10.05	9.54	9.47
7	9.83	9.99	9.91	9.69	10.05	9.58	9.51
8	9.77	9.78	9.82	9.77	9.88	9.77	9.78
9	9.72	9.76	9.84	9.72	9.91	9.79	9.83
10	9.79	9.99	9.93	9.68	10.10	9.55	9.54
avg.	9.87	9.98	9.86	9.71	9.99	9.61	9.56
sd.	0.09	0.11	0.06	0.05	0.10	0.10	0.13

BIBLIOGRAPHY

Abriata, L. A., Salverda, M. L., & Tomatis, P. E. (2012). Sequence-function-stability relationships in proteins from datasets of functionally annotated variants: the case of TEM beta-lactamases. *FEBS Letters, 586*(19), 3330-3335.

Acker, T. M., Gable, J. E., Bohn, M. F., Jaishankar, P., Thompson, M. C., Fraser, J. S., . . . Craik, C. S. (2017). Allosteric Inhibitors, Crystallography, and Comparative Analysis Reveal Network of Coordinated Movement across Human Herpesvirus Proteases. *Journal of the American Chemical Society, 139*(34), 11650-11653.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Generation of Antibody Diversity. In *Molecular Biology of the Cell.* (4th ed.). New York: Garland Science.

Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A., & Ringe, D. (1996). An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *Journal of Physical Chemistry, 100*(7), 2605-2611.

Aoun, E., Chang, C. C., Greer, J. B., Papachristou, G. I., Barmada, M. M., & Whitcomb, D. C. (2008). Pathways to injury in chronic pancreatitis: decoding the role of the high-risk SPINK1 N34S haplotype using meta-analysis. *PLoS One, 3*(4), e2003.

Bakan, A., Nevins, N., Lakdawala, A. S., & Bahar, I. (2012). Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *Journal of Chemical Theory and Computation, 8*(7), 2435-2447.

Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., . . . Levy, R. M. (2005). Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry, 26*(16), 1752-1780.

Banner, D. W., & Hadvary, P. (1991). Crystallographic analysis at 3.0-A resolution of the binding to human thrombin of four active site-directed inhibitors. *Journal of Biological Chemistry, 266*(30), 20085-20093.

Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., . . . Vajda, S. (2018). Exploring the structural origins of cryptic sites on proteins. *Proceedings of the National Academy of Sciences of the United States of America, 115*(15), E3416-E3425.

Berg, J., Tymoczko, J., & Stryer, L. (2002). Section 10.5, Many Enzymes Are Activated by Specific Proteolytic Cleavage. In *Biochemistry.* 5th edition. New York: W H Freeman.

Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D., Dall'Acqua, W., . . . Poljak, R. J. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences of the United States of America, 91*(3), 1089-1093.

Bjorklund, C., Oscarson, S., Benkestock, K., Borkakoti, N., Jansson, K., Lindberg, J., . . . Samuelsson, B. (2010). Design and synthesis of potent and selective BACE-1 inhibitors. *Journal of Medicinal Chemistry, 53*(4), 1458-1464.

Bohnuud, T., Kozakov, D., & Vajda, S. (2014). Evidence of conformational selection driving the formation of ligand binding sites in protein-protein interfaces. *PLoS Computational Biology, 10*(10), e1003872.

Bostrom, J., Yu, S. F., Kan, D., Appleton, B. A., Lee, C. V., Billeci, K., . . . Fuh, G. (2009). Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science, 323*(5921), 1610-1614.

Boulling, A., Chen, J., Callebaut, I., & Férec, C. (2012). Is the SPINK1 p.Asn34Ser Missense Mutation Per se the True Culprit within its Associated Haplotype? *WebmedCentral Genetics, 3*(2).

Boulling, A., Le Marechal, C., Trouve, P., Raguenes, O., Chen, J. M., & Ferec, C. (2007). Functional analysis of pancreatitis-associated missense mutations in the pancreatic secretory trypsin inhibitor (SPINK1) gene. *European Journal of Human Genetics, 15*(9), 936-942.

Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., . . . Sacerdoti, F. D. (2006). *Scalable algorithms for molecular dynamics simulations on commodity clusters.* Paper presented at the Proceedings of the 2006 ACM/IEEE Conference on Supercomputing (SC'06). Tampa, Florida.

Bowman, G. R., Bolin, E. R., Hart, K. M., Maguire, B. C., & Marqusee, S. (2015). Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proceedings of the National Academy of Sciences of the United States of America, 112*(9), 2734-2739.

Bowman, G. R., & Geissler, P. L. (2012). Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proceedings of the National Academy of Sciences of the United States of America, 109*(29), 11681-11686.

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.

Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D. R., Landon, M. R., . . . Vajda, S. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics, 25*, 621-62710.

Brooks, B. R., Bruccoleri, E. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry, 4*, 187-217.

Brown, N. G., Pennington, J. M., Huang, W., Ayvaz, T., & Palzkill, T. (2010). Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM beta-lactamases. *Journal of Molecular Biology, 404*(5), 832-846.

Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., . . . Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry, 26*(16), 1668-1688.

Chen, C. C., Rahil, J., Pratt, R. F., & Herzberg, O. (1993). Structure of a phosphonate-inhibited beta-lactamase. An analog of the tetrahedral transition state/intermediate of beta-lactam hydrolysis. *Journal of Molecular Biology, 234*(1), 165-178.

Chen, C. C., Smith, T. J., Kapadia, G., Wasch, S., Zawadzke, L. E., Coulson, A., & Herzberg, O. (1996). Structure and kinetics of the beta-lactamase mutants S70A and K73H from Staphylococcus aureus PC1. *Biochemistry, 35*(38), 12251-12258.

Childers, M. C., & Daggett, V. (2018). Validating Molecular Dynamics Simulations against Experimental Observables in Light of Underlying Conformational Ensembles. *Journal of Physical Chemistry B, 122*(26), 6673-6689.

Cho, H. S., Mason, K., Ramyar, K. X., Stanley, A. M., Gabelli, S. B., Denney, D. W., Jr., & Leahy, D. J. (2003). Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature, 421*(6924), 756-760.

Chodera, J. D., & Noé, F. (2014). Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology, 25*, 135-144.

Chothia, C., & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology, 196*(4), 901-917.

Chuang, G., Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2008). DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophysical Journal, 95*, 4217-4227.

Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., . . . Sali, A. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology, 428*(4), 709-719.

Ciulli, A., Williams, G., Smith, A. G., Blundell, T. L., & Abell, C. (2006). Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. *Journal of Medicinal Chemistry, 49*(16), 4992-5000.

Clark, M., & Cramer, R. (1989). Validation of the general purpose Tripos 5.2 force field. *Journal of Computational Chemistry, 10*(8), 982-1012.

Colonna-Cesari, F., Perahia, D., Karplus, M., Eklund, H., Braden, C. I., & Tapia, O. (1986). Interdomain motion in liver alcohol dehydrogenase. Structural and energetic analysis of the hinge bending mode. *Journal of Biological Chemistry, 261*(32), 15273-15280.

Dall'Acqua, W., Goldman, E. R., Lin, W., Teng, C., Tsuchiya, D., Li, H., . . . Mariuzza, R. A. (1998). A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry, 37*(22), 7981-7991.

Deane, C. M., Allen, F. H., Taylor, R., & Blundell, T. L. (1999). Carbonyl-carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. *Protein Engineering, 12*(12), 1025-1028.

DeLano, W. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Current Opinion in Structural Biology, 12*(1), 14-20.

DeLano, W., Ultsch, M. H., de Vos, A. M., & Wells, J. A. (2000). Convergent Solutions to Binding at a Protein-Protein Interface. *Science, 287*(5456), 1279-1283.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., & Tawfik, D. S. (2013). What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *Journal of Molecular Biology, 425*(14), 2609-2621.

Dennis, S., Kortvelyesi, T., & Vajda, S. (2002). Computational mapping identifies the binding sites of organic solvents on proteins. *Proceedings of the National Academy of Sciences of the United States of America, 99*(7), 4290-4295.

Durrant, J. D., Keranen, H., Wilson, B. A., & McCammon, J. A. (2010). Computational identification of uncharacterized cruzain binding sites. *PLoS Neglected Tropical Diseases, 4*(5), e676.

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology, 9*, 71.

Ecker, D. M., Jones, S. D., & Levine, H. L. (2015). The therapeutic monoclonal antibody market. *MAbs, 7*(1), 9-14.

Erlanson, D. A., McDowell, R. S., & O'Brien, T. (2004). Fragment-based drug discovery. *Journal of Medicinal Chemistry, 47*(14), 3463-3482.

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery, 10*(5), 449-461.

Goodford, P. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*.

Grant, B. J., Lukman, S., Hocker, H. J., Sayyah, J., Brown, J. H., McCammon, J. A., & Gorfe, A. A. (2011). Novel allosteric sites on Ras for lead generation. *PLoS One, 6*(10), e25711.

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.

Hajduk, P., & Greer, J. (2007). A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews. Drug Discovery, 6*, 211-219.

Hajduk, P., Huth, J., & Fesik, S. (2005). Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry, 48*(7), 2518-2525.

Hall, D. R., & Enyedy, I. J. (2015). Computational solvent mapping in structure-based drug design. *Future Medicinal Chemistry, 7*(3), 337-353.

Hall, D. R., Kozakov, D., Whitty, A., & Vajda, S. (2015). Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery. *Trends in Pharmacological Sciences, 36*(11), 724-736.

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*, 409-443.

Hart, K. M., Moeder, K. E., Ho, C. M. W., Zimmerman, M. I., Frederick, T. E., & Bowman, G. R. (2017). Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLoS One, 12*(6), e0178678.

Harvey, S. C., & Gabb, H. A. (1993). Conformational Transitions Using Molecular-Dynamics with Minimum Biasing. *Biopolymers, 33*(8), 1167-1172.

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

Hecht, H. J., Szardenings, M., Collins, J., & Schomburg, D. (1991). Three-dimensional structure of the complexes between bovine chymotrypsinogen A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). *Journal of Molecular Biology, 220*(3), 711-722.

Hecht, H. J., Szardenings, M., Collins, J., & Schomburg, D. (1992). Three-dimensional structure of a recombinant variant of human pancreatic secretory trypsin inhibitor (Kazal type). *Journal of Molecular Biology, 225*(4), 1095-1103.

Hegyi, E., & Sahin-Toth, M. (2017). Genetic Risk in Chronic Pancreatitis: The Trypsin-Dependent Pathway. *Digestive Diseases and Sciences, 62*(7), 1692-1701.

Hilser, V. J., Garcia-Moreno, E. B., Oas, T. G., Kapp, G., & Whitten, S. T. (2006). A statistical thermodynamic model of the protein ensemble. *Chemical Reviews, 106*(5), 1545-1558.

Hilser, V. J., Wrabl, J. O., & Motlagh, H. N. (2012). Structural and energetic basis of allostery. *Annual Review of Biophysics, 41*, 585-609.

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron, 99*(6), 1129-1143.

Horn, J. R., & Shoichet, B. K. (2004). Allosteric inhibition through core disruption. *Journal of Molecular Biology, 336*(5), 1283-1291.

Huang, B., & Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology, 6*, 19.

Huang, M., Song, K., Liu, X., Lu, S., Shen, Q., Wang, R., . . . Zhang, J. (2018). AlloFinder: a strategy for allosteric modulator discovery and allosterome analyses. *Nucleic Acids Research, 46*(W1), W451-W458.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry, 28*, 1145-1152.

Hynes, T. R., & Fox, R. O. (1991). The crystal structure of staphylococcal nuclease refined at 1.7 A resolution. *Proteins, 10*(2), 92-105.

Ignatov, M., Liu, C., Alekseenko, A., Sun, Z., Padhorny, D., Kotelnikov, S., . . . Kozakov, D. (2019). Monte Carlo on the manifold and MD refinement for binding pose prediction of protein-ligand complexes: 2017 D3R Grand Challenge. *Journal of Computer-Aided Molecular Design, 33*(1), 119-127.

Irani, V., Guy, A. J., Andrew, D., Beeson, J. G., Ramsland, P. A., & Richards, J. S. (2015). Molecular properties of human IgG subclasses and their implications for designing therapeutic monoclonal antibodies against infectious diseases. *Molecular Immunology, 67*(2 Pt A), 171-182.

Jakalian, A., Jack, D. B., & Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry, 23*(16), 1623-1641.

Jiang, W., Phillips, J. C., Huang, L., Fajer, M., Meng, Y., Gumbart, J. C., . . . Roux, B. (2014). Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Computer Physics Communications, 185*, 908-916.

Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M., & Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Protein Science, 20*(3), 482-491.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., & Vakser, I. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques (protein-protein interaction/surface complementarity/macromolecular complex prediction/molecular docking). *Proceedings of the National Academy of Sciences of the United States of America, 89*(6), 2195-2199.

Kather, I., Jakob, R. P., Dobbek, H., & Schmid, F. X. (2008). Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *Journal of Molecular Biology, 383*(1), 238-251.

Kim, H., & Lipscomb, W. N. (1991). Comparison of the structures of three carboxypeptidase A-phosphonate complexes determined by X-ray crystallography. *Biochemistry, 30*(33), 8171-8180.

Kiraly, O., Wartmann, T., & Sahin-Toth, M. (2007). Missense mutations in pancreatic secretory trypsin inhibitor (SPINK1) cause intracellular retention and degradation. *Gut, 56*(10), 1433-1438.

Klimovich, P. V., Shirts, M. R., & Mobley, D. L. (2015). Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design, 29*, 397-411.

Knies, J. L., Cai, F., & Weinreich, D. M. (2017). Enzyme Efficiency but Not Thermostability Drives Cefotaxime Resistance Evolution in TEM-1 beta-Lactamase. *Molecular Biology and Evolution, 34*(5), 1040-1054.

Knoverek, C. R., Amarasinghe, G. K., & Bowman, G. R. (2019). Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends in Biochemical Sciences, 44*(4), 351-364.

Kolossváry, I., & Guida, W. C. (1996). Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *Journal of the American Chemical Society, 118*(21), 5011-5019.

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*, 2159-2166.

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins, 65*, 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*, 867-875.

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., . . . Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols, 10*(5), 733-755.

Kozakov, D., Hall, D. R., Chuang, G.-Y., Cencic, R., Brenke, R., Grove, L. E., . . . Vajda, S. (2011). Structural conservation of druggable hot spots in protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 13528-13533.

Kozakov, D., Hall, D. R., Jehle, S., Luo, L., Ochiana, S. O., Jones, E. V., . . . Vajda, S. (2015). Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proceedings of the National Academy of Sciences of the United States of America, 112*(20), E2585-2594.

Kozakov, D., Hall, D. R., Napoleon, R. L., Yueh, C., Whitty, A., & Vajda, S. (2015). New Frontiers in Druggability. *Journal of Medicinal Chemistry, 58*(23), 9063-9088.

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*, 255-278.

Kozakov, D., Li, K., Hall, D. R., Beglov, D., Zheng, J., Vakili, P., . . . Vajda, S. (2014). Encounter complexes and dimensionality reduction in protein-protein association. *Elife, 3*, e01370.

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.

Kuttner, Y. Y., & Engel, S. (2012). Protein Hot Spots: The Islands of Stability. *Journal of Molecular Biology, 415*, 419-428.

Kuwata, K., Hirota, M., Nishimori, I., Otsuki, M., & Ogawa, M. (2003). Mutational analysis of the pancreatic secretory trypsin inhibitor gene in familial and juvenile pancreatitis in Japan. *Journal of Gastroenterology, 38*(4), 365-370.

Kuwata, K., Hirota, M., Shimizu, H., Nakae, M., Nishihara, S., Takimoto, A., . . . Ogawa, M. (2002). Functional analysis of recombinant pancreatic secretory trypsin inhibitor protein with amino-acid substitution. *Journal of Gastroenterology, 37*(11), 928-934.

Lagasse, H. A., Alexaki, A., Simhadri, V. L., Katagiri, N. H., Jankowski, W., Sauna, Z. E., & Kimchi-Sarfaty, C. (2017). Recent advances in (therapeutic protein) drug development. *F1000Res, 6*, 113.

Landon, M. R., Lieberman, R. L., Hoang, Q. Q., Ju, S., Caaveiro, J. M. M., Orwig, S. D., . . . Ringe, D. (2009). Detection of ligand binding hot spots on protein surfaces via fragment-based methods: application to DJ-1 and glucocerebrosidase. *Journal of Computer-Aided Molecular Design, 23*, 491-500.

Lane, T. J., Bowman, G. R., Beauchamp, K., Voelz, V. A., & Pande, V. S. (2011). Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *Journal of the American Chemical Society, 133*(45), 18413-18419.

Latallo, M. J., Cortina, G. A., Faham, S., Nakamoto, R. K., & Kasson, P. M. (2017). Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme. *Chemical Science (Royal Society of Chemistry: 2010), 8*(9), 6484-6492.

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics, 10*, 168.

Lexa, K. W., & Carlson, H. A. (2013). Improving protocols for protein mapping through proper comparison to crystallography data. *Journal of Chemical Information and Modeling, 53*(2), 391-402.

Li, Y., Li, H., Smith-Gill, S. J., & Mariuzza, R. A. (2000). Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63(,). *Biochemistry, 39*(21), 6296-6309.

Li, Y., Urrutia, M., Smith-Gill, S. J., & Mariuzza, R. A. (2003). Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. *Biochemistry, 42*(1), 11-22.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., & Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins, 78*(8), 1950-1958.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., . . . Wang, R. (2015). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics, 31*, 405-412.

Loll, P. J., & Lattman, E. E. (1989). The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 A. *Proteins, 5*(3), 183-201.

Ma, W., Tang, C., & Lai, L. (2005). Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant. *Biophysical Journal, 89*(2), 1183-1193.

Marchi, M., & Ballone, P. (1999). Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *Journal of Chemical Physics, 110*(8), 3697-3702.

Marciano, D. C., Pennington, J. M., Wang, X., Wang, J., Chen, Y., Thomas, V. L., . . . Palzkill, T. (2008). Genetic and structural characterization of an L201P global suppressor substitution in TEM-1 beta-lactamase. *Journal of Molecular Biology, 384*(1), 151-164.

Marks, C., & Deane, C. M. (2017). Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal, 15*, 222-231.

MathWorks (2019). Two-sample Kolmogorov-Smirnov test - MATLAB kstest2. *MATLAB documentation.* Retrieved December 18, 2019 from https://www.mathworks.com/help/stats/kstest2.html

Mattos, C., & Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature Biotechnology, 14*(5), 595-599.

Maurer, T., Garrenton, L. S., Oh, A., Pitts, K., Anderson, D. J., Skelton, N. J., . . . Fang, G. (2012). Small-molecule ligands bind to a distinct pocket in Ras and inhibit

SOS-mediated nucleotide exchange activity. *Proceedings of the National Academy of Sciences of the United States of America, 109*(14), 5299-5304.

Maynard, J., & Georgiou, G. (2000). Antibody engineering. *Annual Review of Biomedical Engineering, 2*, 339-376.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature, 267*(5612), 585-590.

Miranker, A., & Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins, 11*(1), 29-34.

Modi, T., & Ozkan, S. B. (2018). Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 beta-lactamase. *International Journal of Molecular Sciences, 19*(12).

Morgan, H. P., McNae, I. W., Nowicki, M. W., Hannaert, V., Michels, P. A., Fothergill-Gilmore, L. A., & Walkinshaw, M. D. (2010). Allosteric mechanism of pyruvate kinase from Leishmania mexicana uses a rock and lock model. *Journal of Biological Chemistry, 285*(17), 12892-12898.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry, 30*(16), 2785-2791.

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.

Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature, 508*(7496), 331-339.

Mottarella, S. E., Beglov, D., Beglova, N., Nugent, M. A., Kozakov, D., & Vajda, S. (2014). Docking server for the identification of heparin binding sites on proteins. *Journal of Chemical Information and Modeling, 54*(7), 2068-2078.

Muegge, I. (2006). PMF scoring revisited. *Journal of Medicinal Chemistry, 49*(20), 5895-5902.

Murray, C. W., Verdonk, M. L., & Rees, D. C. (2012). Experiences in fragment-based drug discovery. *Trends in Pharmacological Sciences, 33*(5), 224-232.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443-453.

Ngan, C. H., Hall, D. R., Zerbe, B., Grove, L. E., Kozakov, D., & Vajda, S. (2012). FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics, 28*(2), 286-287.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics, 3*, 33.

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal, 2*, 5.

Oleinikovas, V., Saladino, G., Cossins, B. P., & Gervasio, F. L. (2016). Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *Journal of the American Chemical Society, 138*(43), 14257-14263.

Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P., & Stevens, R. C. (2001). Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature Structural Biology, 8*(3), 238-242.

Paci, E., & Karplus, M. (1999). Forced unfolding of fibronectin type 3 modules: An analysis by biased molecular dynamics simulations. *Journal of Molecular Biology, 288*(3), 441-459.

Perot, S., Sperandio, O., Miteva, M. A., Camproux, A. C., & Villoutreix, B. O. (2010). Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today, 15*(15-16), 656-667.

Pfutzer, R. H., Barmada, M. M., Brunskill, A. P., Finch, R., Hart, P. S., Neoptolemos, J., . . . Whitcomb, D. C. (2000). SPINK1/PSTI polymorphisms act as disease modifiers in familial and idiopathic chronic pancreatitis. *Gastroenterology, 119*(3), 615-623.

Porter, J. R., Moeder, K. E., Sibbald, C. A., Zimmerman, M. I., Hart, K. M., Greenberg, M. J., & Bowman, G. R. (2019). Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. *Biophysical Journal, 116*(5), 818-830.

Porter, K. A. (2019). *Computational Modeling of Protein-Protein And Protein-Peptide Interactions (Doctoral disseration)*. (PhD in Biomedical Engineering), Boston University, Boston, United States. Retrieved from https://open.bu.edu/handle/2144/37989

Porter, K. A., Xia, B., Beglov, D., Bohnuud, T., Alam, N., Schueler-Furman, O., & Kozakov, D. (2017). ClusPro PeptiDock: efficient global docking of peptide recognition motifs using FFT. *Bioinformatics, 33*(20), 3299-3301.

Puius, Y. A., Zhao, Y., Sullivan, M., Lawrence, D. S., Almo, S. C., & Zhang, Z. Y. (1997). Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proceedings of the National Academy of Sciences of the United States of America, 94*(25), 13420-13425.

Raman, E. P., Yu, W., Lakkaraju, S. K., & MacKerell, A. D., Jr. (2013). Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach. *Journal of Chemical Information and Modeling, 53*(12), 3384-3398.

Ramaswamy, S., el Ahmad, M., Danielsson, O., Jornvall, H., & Eklund, H. (1996). Crystal structure of cod liver class I alcohol dehydrogenase: substrate pocket and structurally variable segments. *Protein Science, 5*(4), 663-671.

Rees, D. C., Lewis, M., & Lipscomb, W. N. (1983). Refined crystal structure of carboxypeptidase A at 1.54 A resolution. *Journal of Molecular Biology, 168*(2), 367-387.

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

Salameh, M. A., Soares, A. S., Hockla, A., & Radisky, E. S. (2008). Structural basis for accelerated cleavage of bovine pancreatic trypsin inhibitor (BPTI) by human mesotrypsin. *Journal of Biological Chemistry, 283*(7), 4115-4123.

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design, 27*(3), 221-234.

Schlitter, J., Engels, M., & Kruger, P. (1994). Targeted Molecular-Dynamics - a New Approach for Searching Pathways of Conformational Transitions. *Journal of Molecular Graphics, 12*(2), 84-89.

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*, 264-269.

Schmidtke, P., & Barril, X. (2010). Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *Journal of Medicinal Chemistry, 53*(15), 5858-5867.

Schmidtke, P., Le Guilloux, V., Maupetit, J., & Tuffery, P. (2010). fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Research, 38*(Web Server issue), W582-589.

Schrodinger, LLC. (2017). *Maestro*. New York, NY.

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

Shivakumar, D., Williams, J., Wu, Y., Damm, W., Shelley, J., & Sherman, W. (2010). Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *Journal of Chemical Theory and Computation, 6*(5), 1509-1519.

Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter, K., & Vajda, S. (2003). Identification of substrate binding sites in enzymes by computational solvent mapping. *Journal of Molecular Biology, 332*(5), 1095-1113.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology, 213*(4), 859-883.

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

Speck, J., Hecky, J., Tam, H. K., Arndt, K. M., Einsle, O., & Muller, K. M. (2012). Exploring the Molecular Linkage of Protein Stability Traits for Enzyme Optimization by Iterative Truncation and Evolution. *Biochemistry, 51*(24), 4850-4867.

Stec, B., Holtz, K. M., Wojciechowski, C. L., & Kantrowitz, E. R. (2005). Structure of the wild-type TEM-1 beta-lactamase at 1.55 angstrom and the mutant enzyme Ser70Ala at 2.1 angstrom suggest the mode of noncovalent catalysis for the mutant enzyme. *Acta Crystallographica Section D-Biological Crystallography, 61*, 1072-1079.

Sternberg, M. J., Gabb, H. A., Jackson, R. M., & Moont, G. (2000). Protein-protein docking. Generation and filtering of complexes. *Methods in Molecular Biology, 143*, 399-415.

Sulea, T., Vivcharuk, V., Corbeil, C. R., Deprez, C., & Purisima, E. O. (2016). Assessment of Solvated Interaction Energy Function for Ranking Antibody-Antigen Binding Affinities. *Journal of Chemical Information and Modeling, 56*(7), 1292-1303.

Sun, Z., Wakefield, A. E., Kolossvary, I., Beglov, D., & Vajda, S. (2019). Structure-Based Analysis of Cryptic-Site Opening. *Structure*. https://doi.org/10.1016/j.str.2019.11.007

Thanos, C. D., DeLano, W. L., & Wells, J. A. (2006). Hot-spot mimicry of a cytokine receptor by a small molecule. *Proceedings of the National Academy of Sciences of the United States of America, 103*(42), 15422-15427.

Thomas, V. L., Golemi-Kotra, D., Kim, C., Vakulenko, S. B., Mobashery, S., & Shoichet, B. K. (2005). Structural consequences of the inhibitor-resistant Ser130Gly substitution in TEM beta-lactamase. *Biochemistry, 44*(26), 9330-9338.

Vajda, S., Sippl, M., & Novotny, J. (1997). Empirical potentials and functions for protein folding and binding. *Current Opinion in Structural Biology, 7*(2), 222-228.

Vajda, S., Weng, Z., Rosenfeld, R., & Delisi, C. (1994). Effect of Conformational Flexibility and Solvation on Receptor-Ligand Binding Free Energies+. *Biochemistry, 33*, 13977-13988.

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*, 435-444.

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

Vidarsson, G., Dekkers, G., & Rispens, T. (2014). IgG subclasses and allotypes: from structure to effector functions. *Frontiers in Immunology, 5*, 520.

Vivcharuk, V., Baardsnes, J., Deprez, C., Sulea, T., Jaramillo, M., Corbeil, C. R., . . . Purisima, E. O. (2017). Assisted Design of Antibody and Protein Therapeutics (ADAPT). *PLoS One, 12*(7), e0181490.

Wagner, J. R., Lee, C. T., Durrant, J. D., Malmstrom, R. D., Feher, V. A., & Amaro, R. E. (2016). Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical Reviews, 116*(11), 6370-6390.

Wang, J., Wang, W., Kollman, P. A., & Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling, 25*(2), 247-260.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry, 25*(9), 1157-1174.

Wang, R., Lai, L., & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design, 16*(1), 11-26.

Wang, X., Minasov, G., & Shoichet, B. K. (2002a). Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of Molecular Biology, 320*(1), 85-95.

Wang, X., Minasov, G., & Shoichet, B. K. (2002b). The structural bases of antibiotic resistance in the clinically derived mutant beta-lactamases TEM-30, TEM-32, and TEM-34. *Journal of Biological Chemistry, 277*(35), 32149-32156.

Wassman, C. D., Baronio, R., Demir, O., Wallentine, B. D., Chen, C. K., Hall, L. V., . . . Amaro, R. E. (2013). Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nature Communications, 4*, 1407.

Weikl, T. R., & von Deuster, C. (2009). Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. *Proteins, 75*(1), 104-110.

Weitzner, B. D., Dunbrack, R. L., Jr., & Gray, J. J. (2015). The origin of CDR H3 structural diversity. *Structure, 23*(2), 302-311.

Weng, Z., Delisi, C., & Vajda, S. (1997). Empirical free energy calculation: comparison to calorimetric data. *Protein Science, 6*(9), 1976-1984.

Whitcomb, D. C. (2013). Genetic risk factors for pancreatic disorders. *Gastroenterology, 144*(6), 1292-1302.

Wilson, I. A., & Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Current Opinion in Structural Biology, 4*(6), 857-867.

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

Wrabl, J. O., Gu, J., Liu, T., Schrank, T. P., Whitten, S. T., & Hilser, V. J. (2011). The role of protein conformational fluctuations in allostery, function, and evolution. *Biophysical Chemistry, 159*(1), 129-141.

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., 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.

Zhang, C., Vasmatzis, G., Cornette, J. L., & DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology, 267*(3), 707-726.

Zhang, E., & Tulinsky, A. (1997). The molecular environment of the Na+ binding site of thrombin. *Biophysical Chemistry, 63*(2-3), 185-200.

Zimmerman, M. I., Hart, K. M., Sibbald, C. A., Frederick, T. E., Jimah, J. R., Knoverek, C. R., . . . Bowman, G. R. (2017). Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Central Science, 3*(12), 1311-1321.

Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R., & Bowman, G. R. (2018). Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *Journal of Chemical Theory and Computation, 14*(11), 5459-5475.

CURRICULUM VITAE