Development of computational approaches for the design of antibodies and other binding proteins for target epitopes

by

Varun Mahendra Chauhan

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2022

Keywords: antibody, fixed-backbone binding protein, de novo structure design, hydrogen bonds, conformational stability, force fields

Approved by

Dr. Robert Pantazes, Chair, Assistant Professor, Department of Chemical Engineering
Dr. Andrew Adamczyk, Assistant Professor, Department of Chemical Engineering
Dr. Rafael Bernardi, Assistant Professor, Department of Physics
Dr. Allan David, John W. Brown Associate Professor, Department of Chemical Engineering

Abstract

Antibodies are the most studied and successful therapeutic binding protein. Alternative, smaller-sized binding proteins are being modified for therapeutic applications. Computational energy models i.e. forcefields, protein analysis and design tools have been developed to aid in the discovery and engineering of antibodies and other binding proteins. Current binding protein design tools employ iterative cycles of energy calculations and optimizations, which are computationally expensive. In this work, I have gained a deeper understanding of features of binding interface residues and make an attempt towards developing novel computational design approaches that use sequential forcefield-independent steps.

The dissertation begins with the development and energetic analysis of a database of non-redundant antibody-antigen complexes. Building upon this analysis, we developed AUBIE, a novel computational tool for the de novo design of binding proteins like antibodies for any target epitope. AUBIE identifies groups of compatible binding loops that can simultaneously make strong interactions with the target epitope. AUBIE-generated antibodies for a HER2 epitope were tested experimentally. In order to get a better understanding of flexible nature of binding surfaces, features related to the conformational stability of binding epitopes and paratopes were studied. Learning from these features, improvements were made to the AUBIE approach to better replicate epitope stabilities in AUBIE designed antibodies. We then developed a novel MutDock, docking approach that can simultaneously perform mutations for fixed backbone binding protein scaffolds for any target epitope. MutDock was benchmarked against commonly used fixed-sequence

docking tools. The dissertation will provide in-depth details about the methods and results regarding database generation and analysis, AUBIE and MutDock.

Acknowledgments

This dissertation would not have been possible without the support, guidance and efforts of several people. I have been extremely fortunate to have Dr. Pantazes as my Ph.D. advisor. His guidance, patience, expertise and belief in me have been the catalyst in my transformation into a researcher. I would like to thank my thesis committee members: Dr. Bernardi, Dr. Adamczyk, Dr. David and Dr. Montgomery for their support and insights. Special thanks to Dr. Bernardi and Dr. Montgomery for agreeing to serve on my committee late during my studies. I am also thankful to collaborators from University of Texas, Austin and Johns Hopkins University for experimentally testing predictions from this work and helping us improve our design methods. I am grateful to Dr. Eden and the Chemical Engineering department staff for giving me much needed administrative help over the past six years. I would like to thank my friends, including Pantazes group members, for their feedback, help and encouragement and for cheering me up when my Ph.D. seemed impossible to complete. Lastly, I would like to thank my family for supporting my decision to pursue a Ph.D. and being my emotional support throughout the past six years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

vdW	van der Waals
H-bond	Hydrogen bond
BE	Binding energy
WT	Wild-type
Ab	Antibody
Ag	Antigen
SASA	Solvent Accessible Surface Area
Sc	Shape complementarity
CDR	Complementarity Determining Region
RL-SASA	Rotamers Lost per Solvent Accessible Surface Area buried
SBR	Strong Binding Region
PDB	Protein Data Bank
EBA	Epitope Binding Atom
PBR	Paratope Binding Region
RAbD	Rosetta Antibody Designer

1. Chapter 1 - Introduction

Proteins are a class of biomolecules made of repeating units called amino acids. There are twenty common naturally occurring amino acids which vary in the group of atoms attached to the alpha carbon of the amide backbone atoms. These side chain atoms are responsible for the unique properties of each amino acid such as size, polarity, charge and hydrophobicity. The diverse arraignment of these amino acids in protein sequences allows them to attain structures that are able to perform a wide array of functions in living organisms. In humans, proteins catalyze biochemical reactions (enzymes) [1], act as messengers between cells, tissues and organs (hormones) [2], perform vital structural roles (collagen) [3], maintain the required pH levels (hemoglobin) [4] and supply and store nutrients (hemoglobin, ferritin) [5] among other vital roles. Furthermore, in the immune system, proteins are responsible for aiding in the recognition and elimination of foreign substances through surface receptors on various immune cells and antibodies. Parts of this chapter have been adopted from our article titled "MutDock: A Computational Docking Approach for Fixed-Backbone Protein Scaffold Design", which is under review.

Antibodies

Antibodies (i.e., immunoglobulins (Ig)) are binding proteins that can attach to immune cell receptors and harmful foreign microbes, viruses and molecules at the same time, thus functioning as signaling proteins that increase the effectiveness of the immune response. After the innate immune system is unable to control the growth and spread of a foreign pathogen, the adaptive immune system employs B-cells to secret antibodies which will bind to that pathogen with high specificity and affinity [6]. This production of antibodies occurs after naïve B-cells are activated by either helper T cells that present a surface protein of the pathogen or B-cell surface receptors

that bind to the pathogen itself. Once the antibodies are secreted into the blood stream, they aid in deletion of their target pathogen (i.e., antigen) through various mechanisms. By binding to their target antigens, antibodies may either block the target's mode of action to enter host cells (viruses) or tag the target for destruction by macrophages, natural killer cells or complement activation [7]. During and after the antibody-mediated attack on the pathogen, some activated B-cells undergo clonal expansion and differentiation to become affinity matured memory B-cells in the lymph nodes. The affinity maturation of B-cells in lymph nodes occurs through the somatic hypermutation of the antibody-coding genetic code in the B-cells and the selection of higher affinity antibodies expressed on the surface of the mutated B-cells [6]. These memory B-cells exist in the bloodstream for significantly longer durations than non-memory B-cells and lead a stronger immune response during future exposures to the same pathogen [8]. The large diversity of antibodies and their strong binding affinity and specificity towards their target antigens can be understood in their structure and mode of generation.

Among the five classes of antibodies (i.e., IgA, IgD, IgE, IgG and IgM), which differ from one another in their structure and role during an immune response, IgG is the most abundant type [9]. IgG antibodies fold into a 'Y' shape that consists of pairs of heavy and light chains as shown in Figure 1.1. The stem of the 'Y' structure consists of two identical heavy chains that extend into and occupy half of the two branches of the antibody. The other halves of the branches are the identical light chains. The stem and lower halves of the branches constitute the constant region while the upper halves are known as the variable regions of the antibody. The constant region of the antibody (Fc) is responsible for binding to the immunogenic cell receptor [9]. The variable region consists of a highly consistent segment known as the framework and a hypervariable region known as the Complementarity Determining Regions (CDRs) as shown in Figure 1.2. The six

CDR loops (H1, H2, H3, L1, L2, L3), three from each chain, form the antigen binding region of the antibody and are capable of binding to a variety of proteins, peptides and small molecules [10]. The region of the antibody that binds to the antigen is called the paratope and the region of the antigen that binds to the antibody is known as the epitope. In humans, the IgG variable region is created through the recombination of three germline gene regions: variable (V), diversity (D) and joining (J). The recombination of three genes (V, D and J) is used to generate heavy variable regions while the recombination of two genes (V and J) is used from two gene loci (κ and λ) to create the light variable regions. There are 65 V, 27 D and 6 J genes for the heavy variable regions, resulting in a diversity of 10,530 combinations. For the κ light chains, 40 V and 5 J gene segments can provide a diversity of 200 combinations. For the λ light chains, 30 V and 4 J gene segments can provide a diversity of 120 combinations. Hence a total of 320 different light chains can combine with 10,530 types of heavy chains to provide a theoretical diversity of 3,369,600 different variable regions [11]. The process of joining V, D or J segments involves the deletion of nucleotides by the Recombination-Activation Genes (RAG) protein complex and the random addition of nucleotides by the enzyme terminal deoxynucleotidyl transferase (TdT), leading to junctional diversity at gene joints only [11]. This junctional diversity is located entirely in CDR3 since all the gene junctions occur in this CDR. The combination of genetic combinatorial diversity, junctional diversity and B-cell somatic hypermutations enables the human immune system to produce billions of distinct antibodies [11] [12].

Alternative binding proteins

Although antibodies have been very successful, they may not be the best choice for all contexts where protein binding is needed. An alternative to their use are smaller sized protein

domains, including Knottins [13], Kunitz domains [14], Fynomers [15], and Fibronectin domains [16] among others. Benefits of using smaller sized scaffolds over antibodies include better thermostability, higher tumor penetration, lower cost of production and decreased chance of denaturation [17]–[19]. These smaller scaffolds have seen success as therapeutics: drugs developed using Kunitz domains and Knottins have been approved by the FDA while other alternative scaffold-based drugs are in different phases of clinical trials [20].

Some of these alternative scaffolds bind to target molecules with modular loops, comparable to the CDRs of antibodies, and such proteins can be designed with computational protocols similar to those developed for antibodies. However, binding for a subclass of these alternative scaffolds is governed by point mutations to surface residues in highly stable secondary structures, and these mutations do not alter the proteins' structures. Such proteins will be referred to as fixed-backbone binding proteins. Examples of such binding proteins include affibodies and designed ankyrin repeats (DARPins) [21]. An affibody consists of 58 amino acids and is arranged in a three alpha helix bundle framework. Design of antigen-specific affibodies has been primarily done through combinatorial mutations of 13 surface residues on two helices [22]. Affibodies have been designed to bind over 40 antigens and the HER-2 binding affibody ABY-025 has reached phase 2/3 clinical trial [23]. A DARPin molecule consists of 33 residue long motifs that are typically repeated two to four times, along with N and C terminal motifs. Similar to affibodies, binders are designed by mutating six residues in each motif barring the terminal motifs [24]. Abicipar, a VEGF-A binding therapeutic DARPin drug, has reached phase III clinical trial [20]. The structures and variable residues of affibodies and DARPins are shown in Figure 1.3.

Each antibody has a potential to bind to its respective target region (i.e., epitopes) with high affinity and specificity, a feature that makes them extremely useful in biomedical research

and powerful tools in the study of proteins [25]. Since 1975, monoclonal antibodies (mABs) have been synthesized in the lab to bind to specific epitopes [26]. mABs have been used for numerous therapeutic applications that mainly include serving as diagnostic reagents for imaging and analysis of diseases and as direct treatments for several cancers, autoimmune diseases and cardiovascular diseases. These applications have led to the development of a $130 billion global market of more than 80 FDA approved therapeutic mABs [27], [28]. All these mABs have been initially identified using experimental means. The two conventional approaches of generating mABs in lab, in-vitro evolution and hybridoma technology, while fairly successful, are laborious and inefficient in binding to the target epitope [29], [30]. Hence there is a need for a paradigm shift in antibody design approaches to generate novel antibodies faster than purely experimental means.

Force fields

The fundamental challenge of rational antibody design and in protein design in general, is predicting the structure and function of a protein from its amino acid sequence [31]. Proteins are made of hundreds or thousands of amino acids and there exist a vast number of possible backbone and side chain angles of each amino acid, diverse non-bonded interactions and various folding mechanisms and secondary structures. Such large degrees of freedom make overcoming this challenge virtually impossible without computational means [31]. Using computational methods allows for the quantification of the forces that dictate how atoms interact with one another within and between amino acids and eventually, allow us to make reasonable predictions of the protein structure and behavior. This quantification of forces requires the calculation of various forms of energies through mathematical formulations (i.e., energy force fields). Energy force fields are used to calculate the change in the Gibbs free energy between the initial and final states of proteins since

proteins adopt structures that allow them to minimize their total Gibbs free energy [32]. The change in Gibbs free energy is given by Equation 1.

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

Here, ΔH is the enthalpic energy change and $T\Delta S$ represents the free energy change due to entropic disorder of the system, mainly from the protein's interaction with the solvent (i.e., solvation) [32]. In the context of this work, Gibbs free energy calculations need to be performed to determine energetically stable protein structures obtained from experimental methods, to analyze antibody-antigen complex interfaces and for molecular dynamics (MD) simulations. Several force fields have been developed for protein analysis such as CHARMM [33], AMBER [34], Rosetta [32], GROMOS [35], FoldX [36] among others. CHARMM, AMBER and Rosetta have often been used in other work concerning antibody design [10] [37] or analysis [38] and hence are used in this work.

The AMBER force field calculates the total change in Gibbs free energy using Equation 2 [39].

$$\begin{aligned}\Delta G_{total} = & \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \\ & + \sum_{dihedrals} \frac{V_n}{2}((1 + \cos(n\phi - \gamma)) + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \\ & + \sum_{H\ bonds}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] + E_{solvation}\end{aligned} \tag{2}$$

In the first and second term in Equation 2, separate energy functions treat bonds and angles as springs and use the quadratic form of Hooke's law to calculate the respective bond and angle energies. The third term calculates the bond rotation energy i.e., torsion energy by using a Fourier

series equation. All the fitted coefficients in these equations are determined through experimental data [39]. The non-bonded energies (the fourth term) comprise of Van der Waals (vdW) and electrostatic energies. The vdW energies were calculated using the 12-6 Lennard Jones potential while the electrostatic energies were calculated using Coloumbs law with a distance dependent dielectric. The penultimate term calculates the hydrogen bond energies using a 10-12 potential. The modified implicit Generalized Born Surface Area (GBSA) model by A. Onufriev, D. Bashford and D.A. Case divides the solvation energy into two components: columbic and non-columbic [34]. The columbic solvation energy accounts for the electrostatic energy between solute atoms in a continuum solvent based on the position of the solute atom in the protein. It is calculated using the generalized Generalized Born model. The non-columbic solvation energy accounts for the charge-less vdW effect between the solvent and the solute and the energetic cost of breaking this solute-solvent complex. It is proportional to the solvent exposed surface area of each atom and is calculated by the Linear Combinations of Pairwise Overlaps (LCPO) model [40]. The AMBER force field ff14SB was used in this work.

Similar to the energy calculation in AMBER, the total free energy change in CHARMM was calculated as a sum of individual energies that accounts for bond lengths, bond angles, bond torsions, improper dihedral angles, Urey-Bradley energy, vdW, electrostatics and solvation energies as shown in the Equation 3 [33].

$$\Delta G_{total} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi\left((1 + \cos(n\phi - \delta)\right) + \sum_{improper\ dihedrals} K_\varphi(\varphi - \varphi_0)^2 + \sum_{Urey-Bradley} K_{UB}(r_{1,3} - r_{1,3;0})^2 + \sum_{nonbonded} \left[\frac{q_i q_j}{4\pi D r_{ij}} + \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right]\right] + E_{solvation} \quad (3)$$

The additional fourth term accounts for out-of-plane bending and the Urey Bradley energy ensures the right distances between two atoms bonded to a common atom (1-3 atom pair). Similar to Equation 2, energies for bond lengths, angles, improper dihedrals and Urey-Bradley energies are accounted for using a quadratic function and angle torsions are calculated by a Fourier series form. The Fast Analytical Continuum treatment of Solvation (FACTS) model was used to calculate implicit solvation energy in CHARMM [41]. FACTS is a Generalized Born model that differs from AMBER's GBSA model in its approach to calculate the atomic electrostatic solvation free energy and the solvent exposed surface area. Both the columbic and non-columbic energies in FACTS are estimated using a geometrical approach that accounts for the symmetrical manner in which solute atoms displace water around another atom. Explicit solvation models are computationally expensive and slow to implement for the purposes of this work even though they are more accurate than the implicit solvation models used in this work [42]. The CHARMM force field models CHARMM22 and CHARMM36 have been used in this work. An additional

correction term for the backbone angular dependence of the total energy is added to Equation 3 in CHARMM 22.

Rosetta uses a linear combination of several energies to calculate the total change in Gibbs free energy given by Equation 4 [32].

$$\Delta G_{total} = \sum_{i} w_i E_i(\Theta_i, aa_i) \tag{4}$$

Here, w_i is a weight factor for energy E_i which is a function of degrees of freedom Θ_i and amino acid identity aa_i. Rosetta uses segmented versions of the Lennard-Jones potential and Coulomb's law to simulate vdW and electrostatic energies respectively. Rosetta uses the Gaussian solvent-exclusion model EEF1 developed by Lazardis and Karplus [43]. This implicit solvent model has two components. The first component estimates the free energy change associated with the loss in volume occupied by the solvent around the solute with a Gaussian function. The second component calculates the energy associated with solvent perturbation due to secondary solute presence by using neutralized ionic side chains and a distance-dependent dielectric constant. Rosetta adds a third anisotropic orientation-dependent term that accounts for the destabilizing presence of solutes at possible hydrogen bonding regions around ionic atoms. Hydrogen bond energies are calculated by combining electrostatics and orientation preferences observed in experimentally obtained high resolution protein crystal structures. Moreover, hydrogen bond energies are divided into four components based on the side chain or backbone identity and distance between the participating atoms [32]. Disulfide bond energies are calculated based on structural features observed in crystal protein structures. Energies associated with backbone and side chain torsions are calculated based on parameters obtained from probability estimates of backbone dihedral angles and rotamer conformations given the amino acid type. All probability

calculations are based on a library of high-resolution, non-redundant protein structures [32]. Additional terms account for reference amino acid energies in the unfolded state, unstable open proline rings and tyrosine hydroxyl hydrogens which are not in the aromatic plane. The weight factor for each energy term has been adjusted to match calculated energies with small molecule thermodynamic data and experimental observations [32]. Contrary to the 'first principles' based approach of AMBER and CHARMM, Rosetta is highly dependent on statistical coefficients obtained from protein structures. Rosetta is the most commonly used force field for protein engineering and is used specifically for antibody design in approaches such as RosettaAntibodyDesign (RAbD) [37] and hotspot-centric approach [44]. However, its electrostatics potential is parameterized for the hydrophobic interiors of proteins rather than the polar exteriors [45]. Other design software developed by the Maranas group [10] [12] [46] employ CHARMM for antibody design as discussed later.

Computational antibody design

The computational design of antibodies has made several advances in the past two decades. Analysis of libraries of antibody structures has revealed that the backbone structure of antibody framework and constant regions is mostly conserved [10]. Combined with the fact that antibody binding is dictated by CDR loops, antibody design efforts have focused at engineering CDR loop shapes and sequences to bind to the target antigen. Various types of computational approaches have been used to generate novel antibodies. One of these approaches includes grafting non-Ig peptides into CDRs to mimic natural interactions from a non-Ig complex. Williamson et al. grafted binding regions of the misfolded Prion protein (PrP^{sc}) into HCDR3 and generated antibodies that bound both PrP^{sc} and cellular Prion proteins [47]. Peter el al. grafted hydrophobic regions of Aβ

peptide into HCDR3 and generated antibodies that bound Aβ fibrils with sub-micromolar affinity [48]. Another approach combines experimental and computational design by conserving residues of interest and mutating neighboring residues in CDRs to generate libraries that are screened for the desired property. Smith et al. [49] and Taussig and Stoevesandt [50] used this hybrid approach to design integrin binding antibodies and phosphorylated peptide binding antibodies respectively. A common design approach is to redesign CDRs through point mutations. Lippow, Wittrup and Tidor improved binding affinity of anti-lysozyme antibody by computationally evaluating point mutations in the CDRs and selecting those that increased electrostatic interactions [51]. Sasisekharan et al. mutated residues at the periphery of the binding interface of a dengue virus antibody to widen its specificity to other variants of the virus [52]. Further information on computational antibody design approaches can be obtained from reviews performed by Tessier and Tiller [53], Fischman and Ofran [54], Nakamura et al. [55] and Saven et al. [31]

The most challenging approach to designing antibodies is to determine the sequence and structure of CDR loops or variable domains that produces a desired function, such as binding to a target epitope, without an initial antibody-antigen complex structure as a starting point. Several methodologies and software have been proposed for the computational design of CDRs and antibody variable domains that bind to a specific antigen epitope. These are OptCDR [10], OptMAVEn [12], OptMAVEN 2.0 [46], RAbD [37], Hotspot-centric design [44] and AbDesign [56] among others. These are reviewed in brief below.

OptCDR follows a four step workflow. In step 1, optimal canonical CDR structures are selected to bind with the selected antigen epitope. This is done through the use of a library of canonical structures for all the CDRs and MILP optimization to determine the best group of non-clashing canonical structures. In step 2, the amino acid sequences of the canonical CDR structures

from step 1 are initialized through the use of energy functions, a rotamer library and a MILP rotamer optimization formulation. In step 3, over thousand iterations of the IPRO program [57] are performed to determine the optimal CDR structures and sequences. Each IPRO iteration consists of steps that perturb the backbone angles of a randomly chosen CDR, reselect the CDR amino acid sequence, minimize the total energy of the complex and dock the antibody to the epitope followed by another energy minimization. In step 4, several ranked solutions with different amino acid sequences are obtained using the optimal complexes from step 3 through rotamer selection MILP. OptCDR was used to design antibodies that bind to a peptide from the capsid of hepatitis C, fluorescein and VEGF. In all the three cases, the OptCDR designs for antibody and nanobody scaffolds had binding energies similar to or stronger than the wild type and mutated antibodies designed to bind the same epitope. All the simulations were run on 3.0 GHz Intel Xeon processors with 4 GB of RAM for an average of 9 days [10]. Out of 50 designs generated by OptCDR to bind a tetra peptide, four displayed strong binding as they bound with nanomolar affinity in experimental tests [58].

OptMAVEn follows a three step workflow. In step 1, an antigen binding site is defined as a rectangular box near the CDRs that occupies all the geometric centers of 750 known epitope structures. The target epitope mean is then positioned in 539 equally spaced grid points inside the box and rotated around the z axis with intervals of 60 degrees while the epitope is positioned facing the CDRs. In step 2, a MILP optimization formulation was used to identify combinations of six modular antibody parts from the MAPS database [59] with low interaction energies calculated using the CHARMM force field. The MAPS database, developed by the same group, consists of modular variable domain loops analogous to the V, D, and J germline regions. In step 3, the IPRO protocol is used for affinity maturation of the designed variable domains from step 2. Furthermore,

in each iteration, all 9-mer sequences from the designs are compared with 9-mer sequences from human antibodies to assess their immunogenicity. Designs with lower dissimilarity scores are selected for the next IPRO iteration. The authors computationally benchmarked each step of the OptMAVEn workflow and designed antibody libraries to bind to two antigens, envelope glycoprotein gp120 and hemagglutunin. All the final affinity matured designs displayed interaction energies lower than the native complexes and the pre-affinity maturation designs. A design run using ten processors in parallel with an antigen of about 100 amino acids is expected to take less than three weeks of computational time [12]. Five designs from this set were chosen for experimental validation and three of the designs showed <30 nM binding affinities [60].

The OptMAVEN approach was improved in its second version, OptMAVEN 2.0. OptMAVEn 2.0 improved the handling of the vast amount of sequence and structural data, while the core OptMAVEn procedures of using a rectangular box of grid points for the epitope positioning, searching through the MAPS database, IPRO affinity maturation and humanization remained the same. Initially, for each MAPs loop type, a matrix of pairwise sequence similarity score is developed. Stojmirovic's method is then used to convert the similarity scores into metric distances. This is followed by determining a 3D coordinate for each loop using Distance Geometry Optimization Software and the metric distances. Each OptMAVEn-2.0 design is represented as a 23-dimensional vector, each dimension for a geometric identity for the MAPs parts, antigen position and antigen rotation angle. The 23-dimensional vectors are transformed into 3-dimensional vectors using Principal Component Analysis and clustered using k-means clustering. The clustered designs are then ranked from most to least promising. OptMAVEN 2.0 provides the user with an additional option of performing a 50 ns MD simulation using QwikMD to assess the binding stability of the designs. The authors showed that OptMAVEn 2.0 was about 80% more

efficient in terms of CPU time and memory than OptMAVEn. Moreover, the authors benchmarked OptMAVEn 2.0's performance computationally through 64 antigens [46] and reported runtimes ranging from 4 to 55 hours. No experimental verification has been reported.

RAbD performs the *de novo* CDR redesigns of input antibody-antigen complex structures. RosettaDock can be used prior to RAbD for complex generation. The RAbD design protocol contains of inner and outer loops that are executed repeatedly. The outer loop is responsible for selecting a CDR structure from a CDR cluster and CDR type, all of which are randomly chosen. The inner loop is responsible for performing several tasks such as the sequence design, side-chain repacking and energy minimization of the CDR. The sequence design and side-chain repacking are done using the Monte Carlo side-chain repacking procedure available in Rosetta. Rosetta is also used to perform local steepest-design energy minimization. The Metropolis Monte Carlo criterion is applied to the antibody structure after each inner and outer loop based on the total energy or the interface energy. The structure with the lowest observed energy is produced from the outer loop as the final solution. Users can select the number of inner and outer loop design steps. The authors did not report any runtimes for the benchmarking simulations. The affinity maturation ability of RAbD was experimentally tested. Out of 30 RAbD designs for HIV-CD4 site binding antibody, 20 designs showed some degree of binding. Similarly, out of 27 RAbD designs that were predicted to improve hyalunoridase binding antibody, 7 designs showed some degree of binding. Moreover, four of these designs improved binding affinities 10 to 50 fold over their native structures [37].

The hotspot-centric approach can been used to redesign any scaffold, including antibodies, to bind a target epitope. The approach builds on the idea that protein binding interface consists of hotspots, a cluster of few residues that make a significant contribution to binding energy through

short range hydrogen bonding, vdW packing or electrostatic interactions. The design protocol consists of initially designing a library of single amino acid hotspot positions on the antigen surface using RosettaDock. Parallel to the hotspot library design, the chosen scaffold is docked to the target surface using PatchDock and then the low resolution RosettaDock. Binding poses from the second step that incorporate native or mutated resides from the hotspot library are passed through hotspot energy evaluations and target-specific filters in iterative steps. Successful complexes from the previous step are optimized through RosettaDesign and energy minimization before passing through another set of binding energy and shape complementarity filters [44]. The authors tested the protocol by designing 88 proteins of two or three residue hotspots from various scaffolds to bind the stem region of Hemagglutinin. Out of the 88 designs, only two designs displayed weak binding affinities. The authors then used designed libraries of variants of the two designs using single site-saturation mutagenesis and error-prone polymerase chain reaction. The authors were able to identify mutations that improved the binding affinities to single digit nanomolar values. The authors did not mention any runtimes since the approach consists of iterative steps that require manual analysis [61].

The AbDesign algorithm is built upon the idea of combinatorial backbone and sequence sampling from a database of known structures from the same protein family. The initial steps of the workflow consist of constructing a Position Specific Scoring Matrix (PSSM) based on the structures of proteins from the same family as the chosen scaffold. The chosen scaffold is then segmented based on structurally conserved positions. For antibodies, the segmentation is inspired from the V-D-J genomic segmentation. For each antibody segment, a library of backbone confirmations is then developed. Canonical structures from the clustered database of backbone confirmations are determined and inserted into the chosen scaffold to obtain combinatorial

confirmations. Each of these conformations is docked using low resolution RosettaDock, repacked with side chains using the PSSM matrix, minimized and passed through energy and structural filters. The solution structure from the previous step is then refined using randomly selected different backbone conformations and multi-constrained optimization. The optimization formulation attempts to optimize the binding energy and the energy of the unbound scaffold. A final set of energy and structural filters are used to determine the final solution structures. The authors used AbDesign to improve a set of nine antibody structures binding to their respective epitopes. The generated designs displayed comparable interaction energies and buried surface areas and lower shape complementarity scores than a set of natural protein-binding antibodies. The authors reported runtimes of 4.6 hours for the precomputation steps (PSSM and backbone conformation database generation) and 7 hrs for each design trajectory respectively [56]. An epitope-specific design approach can consist of multiple such design trajectories.

The approaches reviewed represent the current state-of-the-art in computational antibody design. These methodologies employ computationally burdensome and time-consuming algorithms and can design only antibodies, even though have been successful in designing novel high-affinity antibodies with experimental validation. All the approaches employ iterative routines of either protein-protein docking and energy calculations and minimizations which increase memory usage and runtimes. This can be seen in the workflows of the reviewed approaches as shown in Figure 1.4. OptCDR, OptMAVEn and OptMAVEN 2.0 employ CHARMM while RAbD, AbDesign and Hotspot-centric approach use Rosetta for their energy calculations. RAbD, OptMAVEn 2.0 and OptCDR use databases of antibody structures, thus excluding their use for other non-Ig scaffolds discussed earlier. AbDesign depends on the availability of a family of protein structures similar to the chosen scaffold, hence making it difficult to use for other binding

scaffolds. The hotspot-centric approach does not guarantee the generation of designs for a selected scaffold. Previous work that has validated the approach have tested multiple scaffolds [61] [62].

Computational design of fixed backbone scaffolds

Computational methods to design fixed backbone protein scaffolds, such as affibodies and DARPins, require different approaches than those for antibodies due to their lack of loops analogous to CDRs. One strategy is to use a docking program [63]–[67] to create an initial complex followed by iterative cycles of point mutations [68], [69]. This approach is analogous to and inspired by the affinity maturation process of antibodies by the immune system [70] as well as the experimental directed evolution protocol [71]. Various docking approaches have been developed over the past two decades. ZDOCK uses stepwise movements and rotations of rigid body representation of the ligand around the receptor and uses fast Fourier transforms to quickly identify poses with good shape complementarity features. Poses are then ranked based on energy potentials [63]. Other approaches like ClusPro use ZDOCK for good quality pose identification followed by further pose refinement and binding energy evaluations [66]. RosettaDock uses a coarse-grained rigid body Monte Carlo search for high scoring poses which are later refined through local docking accompanied with side chain and backbone packing and energy minimizations. The Rosetta energy function is used in the Monte Carlo search and to rank poses [65]. Swarmdock, a population-based metaheuristic approach, starts with a group of random initial poses and uses a particle swarm optimization algorithm to minimize electrostatic and vdW potentials between two proteins [67]. HADDOCK uses a combination of rigid body energy minimizations of randomly generated starting poses and flexible energy minimizations of the best 1000 initial complexes [64].

Docking approaches use rigid-structure and/or fixed-sequence representations of scaffolds since the original purpose of such tools was to predict native binding orientations. This feature, while appropriate for replicating native complexes, imposes limitations for designing binding proteins because it will reject protein poses with clashes between native side chains that could be rectified through mutations. A design approach that can dock scaffolds with mutable binding surfaces would be likely to identify higher quality complexes than methods that cannot. One strategy to do this would be to dock proteins in a manner where they form strong interactions which are known to be abundant in protein-protein binding interfaces, such as hydrogen bonds (H-bonds) and hydrophobic interactions. The Baker lab has developed RIFdock, an approach that docks proteins to a ligand or another protein while simultaneously making mutations. RIFdock does this by docking individual amino acids to the epitope, generating a large library of reverse rotamers for the well docked amino acids and identifying scaffold positions that can hold multiple reverse rotamers [72]. Cao et al. designed SARS-CoV-2 binding miniproteins by initially scanning a library of 19000 miniproteins through RIFdock against the ACE-2 binding epitope of the RBD. High quality poses were then affinity matured to bind with picomolar affinity with the ACE-2 binding epitope [73]. Similar to other docking approaches like ZDOCK and HADDOCK, RIFdock makes use of grid based movements of the scaffold around the target protein. As of the time of writing of this dissertation, a detailed description of the RIFdock methodology is not available in a peer-reviewed article. Further, RIFdock is not available in the Rosetta Commons.

Objective

The design paradigm that is followed in current approaches is to identify an initial structure with the potential to have a desired function and then identify mutations to improve that function.

While effective, the iterative nature of this paradigm makes it computationally expensive to implement. There is a need for a novel paradigm in protein engineering that replaces the burdensome iterative procedures with straightforward faster geometric tools that can design binding proteins. In this dissertation, I have made an attempt at advancing the state-of-the-art in binding protein engineering. First, energetic and conformational features of binding interfaces were studied and this knowledge was then used to develop geometric tools for the design of antibodies and fixed-backbone binding proteins for any target epitope. In Chapter 2, the development of a non-redundant database of antibody-antigen complexes and the energy contributions to binding on a per-residue basis are discussed. In Chapter 3, the findings from Chapter 2 are built upon to develop a novel computational approach for the *de novo* design of antibodies. Although the designed antibodies are predicted to be statistically superior to naturally occurring ones ($p < 0.0001$ for numerous metrics), they failed to bind when tested experimentally. Motivated by this discrepancy, in Chapter 4, features built on quantitative metrics regarding conformational stability of binding interface residues are developed and analyzed. Chapter 5 describes improvements to the approach from Chapter 3 and the subsequent case study results regarding its effectiveness in generating binding antibodies are described. In Chapter 6, a novel variable sequence docking approach for fixed-backbone protein scaffolds is developed. Results regarding its performance in comparison with other docking tools are described. Finally, in Chapter 7 the learnings, contributions, and future work regarding the projects completed in this dissertation are discussed. An illustrative description of the dissertation is provided in Figure 1.5.

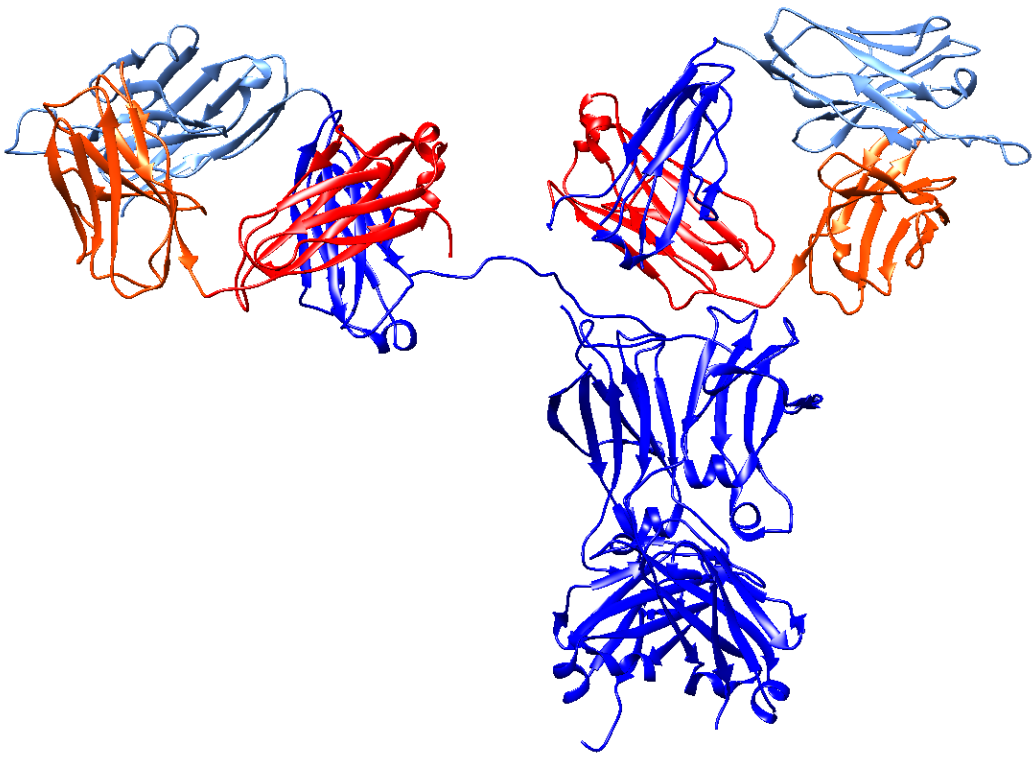

Figure 1.1 The structure of Protein Data Bank (PDB) [74] file 1HZH [75]. This is a human IgG and its heavy chains are depicted in blue and its light chains in red, with the variable domains shown in lighter shades.

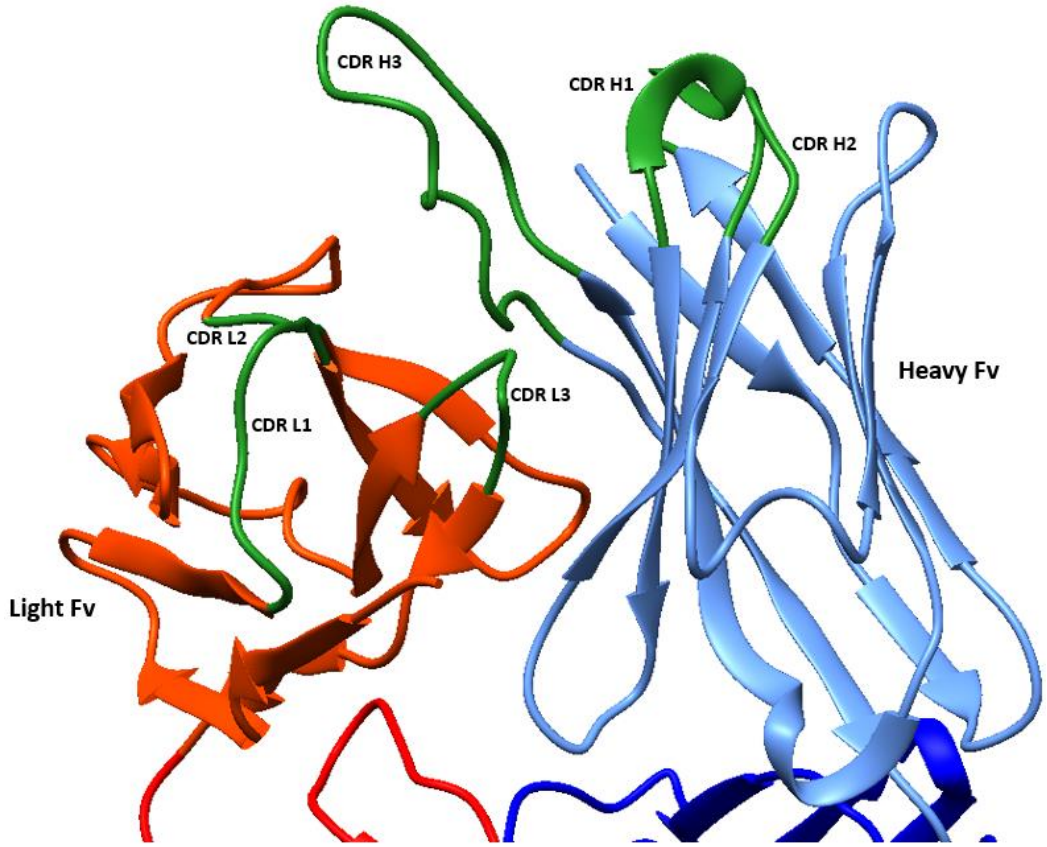

Figure 1.2 Heavy and light CDR loops (in green) of a human IgG. PDB 1HZH [75]

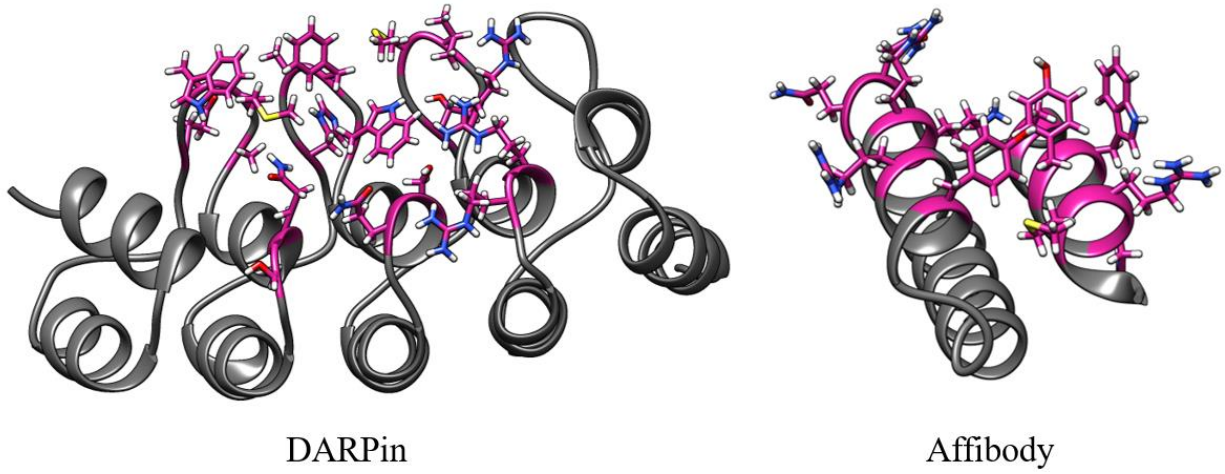

Figure 1.3 Examples of alternative binding scaffolds. Depicted are Affibody (PDB: 3MZW) and DARPin (PDB: 6FPA) structures. Their variable residues that mutate to bind target proteins are colored in pink.

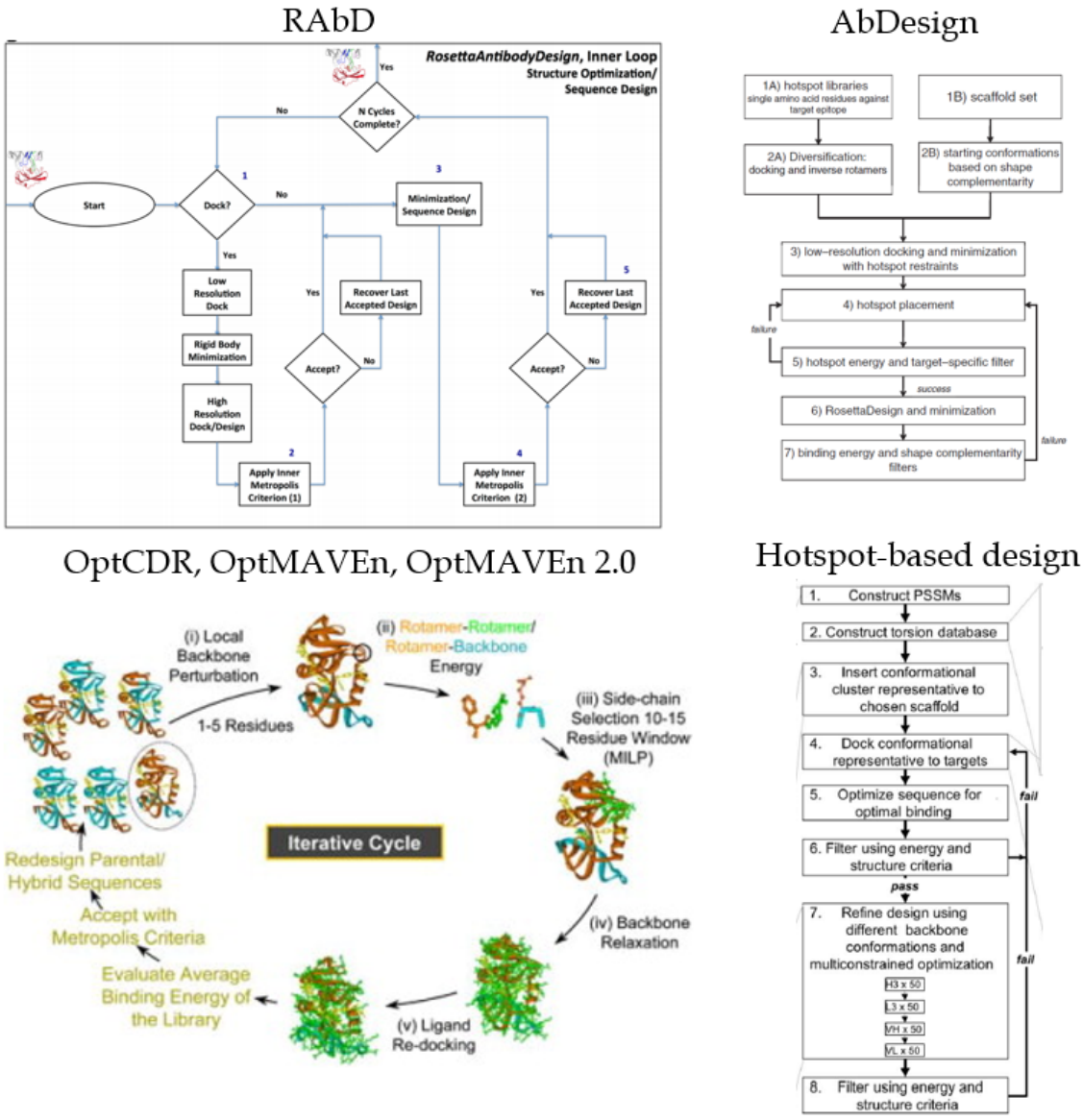

Figure 1.4 Design workflows of currently available antibody design approaches: RAbD, AbDesign, OptCDR, OptMAVEn and Hotspot-based design. Though OptCDR, OptMAVEn and OptMAVEn 2.0 follow different approaches, they all employ the above shown iterative cycle for affinity maturation.

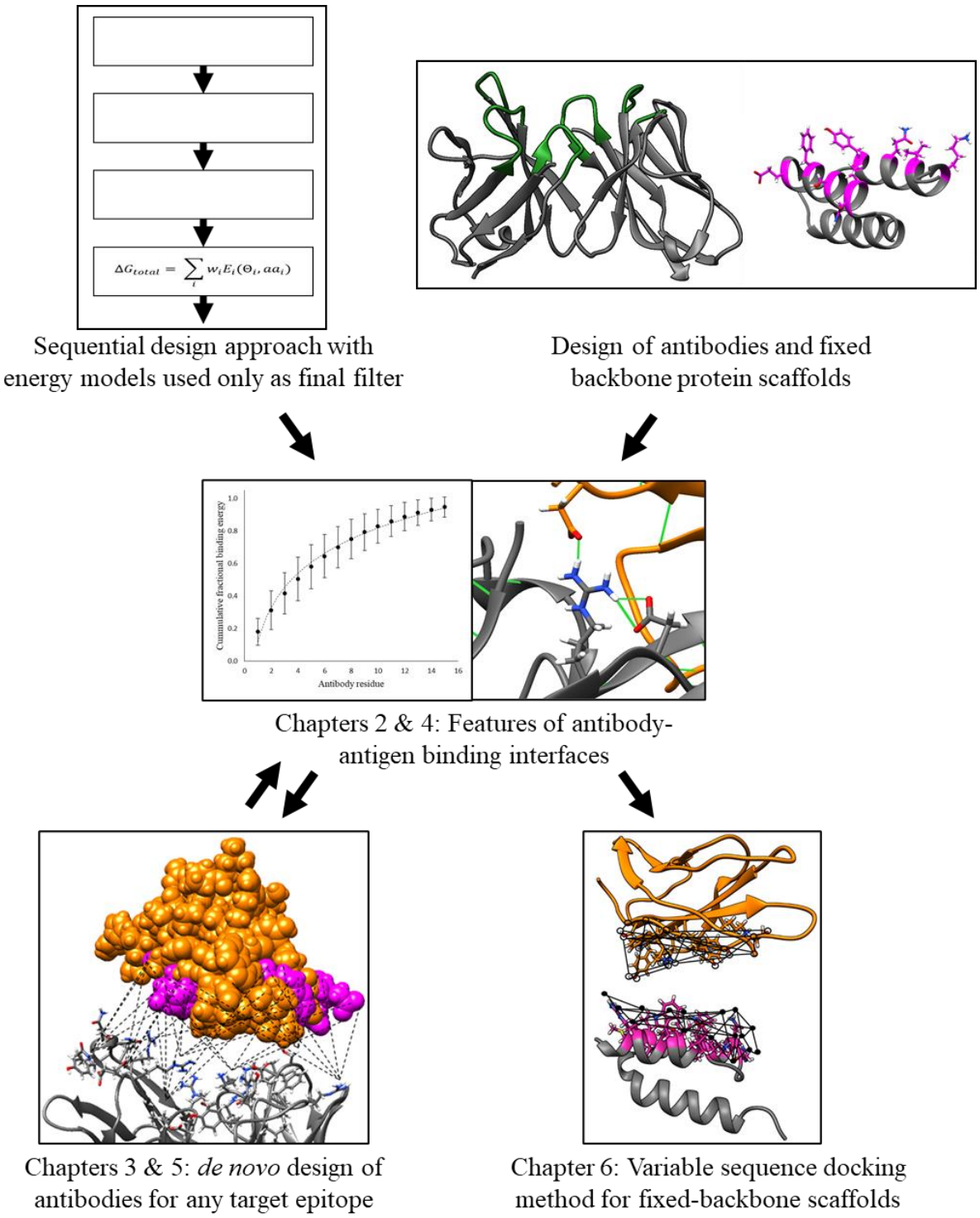

Figure 1.5 Illustrative description of dissertation.

References

[1] G. M. Cooper, “The Central Role of Enzymes as Biological Catalysts.,” in *The Cell: A Molecular Approach*, Sunderland (MA): Sinauer Associates, 2000.

[2] S. Nussey and S. Whitehead, “Chapter 1, Principles of endocrinology,” in *Endocrinology: An Integrated Approach*, Oxford: BIOS Scientific Publishers, 2001.

[3] J. Berg, J. Tymoczko, and L. Stryer, “Chapter 3, Protein Structure and Function.,” in *Biochemistry. 5th edition*, New York: W H Freeman, 2002.

[4] L. L. Hamm, N. Nakhoul, and K. S. Hering-Smith, “Acid-Base Homeostasis,” *Clin J Am Soc Nephrol*, vol. 10, no. 12, pp. 2232–2242, Dec. 2015, doi: 10.2215/CJN.07400715.

[5] E. L. MacKenzie, K. Iwasaki, and Y. Tsuji, “Intracellular iron transport and storage: from molecular mechanisms to health implications,” *Antioxid Redox Signal*, vol. 10, no. 6, pp. 997–1030, Jun. 2008, doi: 10.1089/ars.2007.1893.

[6] D. D. Chaplin, “Overview of the immune response,” *J Allergy Clin Immunol*, vol. 125, no. 2 Suppl 2, pp. S3–S23, Feb. 2010, doi: 10.1016/j.jaci.2009.12.980.

[7] Lumen, “Antibodies.” https://courses.lumenlearning.com/boundless-biology/chapter/antibodies/

[8] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik, “Immunobiology: The Immune System In Health And Disease,” *Immuno Biology 5*, p. 892, 2001, doi: 10.1111/j.1467-2494.1995.tb00120.x.

[9] G. Vidarsson, G. Dekkers, and T. Rispens, “IgG subclasses and allotypes: from structure to effector functions,” *Front Immunol*, vol. 5, p. 520, Oct. 2014, doi: 10.3389/fimmu.2014.00520.

[10] R. J. Pantazes and C. D. Maranas, “OptCDR: A general computational method for the design of antibody complementarity determining regions for targeted epitope binding,” *Protein Engineering, Design and Selection*, vol. 23, no. 11, pp. 849–858, 2010, doi: 10.1093/protein/gzq061.

[11] C. J. Janeway, P. Travers, M. Walport, and M. Shlomchik, “The generation of diversity in immunoglobulins,” in *Immunobiology: The Immune System in Health and Disease*, 5th ed., New York: Garland Science, 2001.

[12] T. Li, R. J. Pantazes, and C. D. Maranas, “OptMAVEn – A New Framework for the de novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes,” *PLoS ONE*, vol. 9, no. 8, 2014, doi: 10.1371/journal.pone.0105954.

[13] S. J. Moore and J. R. Cochran, “Chapter nine - Engineering Knottins as Novel Binding Agents,” in *Methods in Enzymology*, vol. 503, K. D. Wittrup and G. L. Verdine, Eds. Academic Press, 2012, pp. 223–251. doi: https://doi.org/10.1016/B978-0-12-396962-0.00009-4.

[14] R. J. Hosse, A. Rothe, and B. E. Power, “A new generation of protein display scaffolds for molecular recognition,” *Protein Sci*, vol. 15, no. 1, pp. 14–27, Jan. 2006, doi: 10.1110/ps.051817606.

[15] J. Bertschinger, D. Grabulovski, and D. Neri, “Selection of single domain binding proteins by covalent DNA display,” *Protein Engineering, Design and Selection*, vol. 20, no. 2, pp. 57–68, Jan. 2007, doi: 10.1093/protein/gzl055.

[16] A. Koide, C. W. Bailey, X. Huang, and S. Koide, “The fibronectin type III domain as a scaffold for novel binding proteins11Edited by J. Wells,” *Journal of Molecular Biology*, vol. 284, no. 4, pp. 1141–1151, 1998, doi: https://doi.org/10.1006/jmbi.1998.2238.

[17] D. A. Richards, “Exploring alternative antibody scaffolds: Antibody fragments and antibody mimics for targeted drug delivery,” *Drug Discovery Today: Technologies*, vol. 30, pp. 35–46, 2018, doi: https://doi.org/10.1016/j.ddtec.2018.10.005.

[18] L. A. Stern, B. A. Case, and B. J. Hackel, “Alternative non-antibody protein scaffolds for molecular imaging of cancer,” *Current Opinion in Chemical Engineering*, vol. 2, no. 4, pp. 425–432, 2013, doi: https://doi.org/10.1016/j.coche.2013.08.009.

[19] R. N. Gilbreth and S. Koide, “Structural insights for engineering binding proteins based on non-antibody scaffolds,” *Current Opinion in Structural Biology*, vol. 22, no. 4, pp. 413–420, 2012, doi: https://doi.org/10.1016/j.sbi.2012.06.001.

[20] R. Simeon and Z. Chen, “In vitro-engineered non-antibody protein therapeutics,” *Protein & Cell*, vol. 9, no. 1, pp. 3–14, 2018, doi: 10.1007/s13238-017-0386-6.

[21] A. M. Alsultan, D. Y. Chin, C. B. Howard, C. J. de Bakker, M. L. Jones, and S. M. Mahler, “Beyond Antibodies: Development of a Novel Protein Scaffold Based on Human Chaperonin 10,” *Scientific Reports*, vol. 6, no. 1, p. 37348, 2016, doi: 10.1038/srep37348.

[22] S. Ståhl, T. Gräslund, A. Eriksson Karlström, F. Y. Frejd, P.-Å. Nygren, and J. Löfblom, “Affibody Molecules in Biotechnological and Medical Applications,” *Trends in Biotechnology*, vol. 35, no. 8, pp. 691–712, 2017, doi: https://doi.org/10.1016/j.tibtech.2017.04.007.

[23] B. Altunay *et al.*, “HER2-directed antibodies, affibodies and nanobodies as drug-delivery vehicles in breast cancer with a specific focus on radioimmunotherapy and radioimmunoimaging,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, no. 5, pp. 1371–1389, 2021, doi: 10.1007/s00259-020-05094-1.

[24] O. N. Shilova and S. M. Deyev, “DARPins: Promising Scaffolds for Theranostics,” *Acta Naturae*, vol. 11, no. 4, pp. 42–53, 2019, doi: 10.32607/20758251-2019-11-4-42-53.

[25] A. M. Edwards, R. Isserlin, G. D. Bader, S. v Frye, T. M. Willson, and F. H. Yu, “Too many roads not taken,” *Nature*, vol. 470, no. 7333, pp. 163–165, 2011, doi: 10.1038/470163a.

[26] G. KÖHLER and C. MILSTEIN, “Continuous cultures of fused cells secreting antibody of predefined specificity,” *Nature*, vol. 256, no. 5517, pp. 495–497, 1975, doi: 10.1038/256495a0.

[27] K. Chapman *et al.*, “Preclinical development of monoclonal antibodies: considerations for the use of non-human primates,” *MAbs*, vol. 1, no. 5, pp. 505–516, 2009, doi: 10.4161/mabs.1.5.9676.

[28] F. B. Arslan, K. Ozturk, and S. Calis, “Antibody-mediated drug delivery,” *International Journal of Pharmaceutics*, vol. 596, p. 120268, 2021, doi: https://doi.org/10.1016/j.ijpharm.2021.120268.

[29] M. Yamashita, Y. Katakura, and S. Shirahata, “Recent advances in the generation of human monoclonal antibody,” *Cytotechnology*, vol. 55, no. 2–3, pp. 55–60, Dec. 2007, doi: 10.1007/s10616-007-9072-5.

[30] V. G. Poosarla, T. Li, B. C. Goh, K. Schulten, T. K. Wood, and C. D. Maranas, “Computational de novo design of antibodies binding to a peptide with high affinity,” *Biotechnol Bioeng*, vol. 114, no. 6, pp. 1331–1342, Jun. 2017, doi: 10.1002/bit.26244.

[31] I. Samish, C. M. MacDermaid, J. M. Perez-Aguilar, and J. G. Saven, “Theoretical and Computational Protein Design,” *Annual Review of Physical Chemistry*, vol. 62, no. 1, pp. 129–149, Mar. 2011, doi: 10.1146/annurev-physchem-032210-103509.

[32] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, 2017, doi: 10.1021/acs.jctc.7b00125.

[33] K. Vanommeslaeghe *et al.*, “CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields,” *Journal of Computational Chemistry*, vol. 31, no. 4, pp. 671–690, Mar. 2010, doi: 10.1002/jcc.21367.

[34] D. A. Case *et al.*, “AMBER 2017.” University of California, San Francisco., 2017.

[35] W. R. P. Scott *et al.*, “The GROMOS Biomolecular Simulation Program Package,” *The Journal of Physical Chemistry A*, vol. 103, no. 19, pp. 3596–3607, May 1999, doi: 10.1021/jp984217f.

[36] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, “The FoldX web server: an online force field,” *Nucleic Acids Research*, vol. 33, no. suppl_2, pp. W382–W388, Jul. 2005, doi: 10.1093/nar/gki387.

[37] J. Adolf-Bryfogle *et al.*, *RosettaAntibodyDesign (RAbD): A general framework for computational antibody design*, vol. 14, no. 4. 2018. doi: 10.1371/journal.pcbi.1006112.

[38] J. K. X. Maier and P. Labute, “Assessment of fully automated antibody homology modeling protocols in molecular operating environment,” *Proteins*, vol. 82, no. 8, pp. 1599–1610, Aug. 2014, doi: 10.1002/prot.24576.

[39] S. J. Weiner *et al.*, “A new force field for molecular mechanical simulation of nucleic acids and proteins,” *J Am Chem Soc*, vol. 106, no. 3, pp. 765–784, 1984.

[40] J. Weiser, P. S. Shenkin, and W. C. Still, “Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO),” *Journal of Computational Chemistry*, vol. 20, no. 2, pp. 217–230, Jan. 1999, doi: 10.1002/(SICI)1096-987X(19990130)20:2<217::AID-JCC4>3.0.CO;2-A.

[41] U. Haberthür and A. Caflisch, “FACTS: Fast analytical continuum treatment of solvation,” *Journal of Computational Chemistry*, vol. 29, no. 5, pp. 701–715, Apr. 2008, doi: 10.1002/jcc.20832.

[42] J. Kleinjung and F. Fraternali, “Design and application of implicit solvent models in biomolecular simulations,” *Curr Opin Struct Biol*, vol. 25, no. 100, pp. 126–134, Apr. 2014, doi: 10.1016/j.sbi.2014.04.003.

[43] T. Lazaridis and M. Karplus, “Discrimination of the native from misfolded protein models with an energy function including implicit solvation 11Edited by A. R. Fersht,” *Journal of Molecular Biology*, vol. 288, no. 3, pp. 477–487, 1999, doi: https://doi.org/10.1006/jmbi.1999.2685.

[44] S. J. Fleishman, J. E. Corn, E. M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker, “Hotspot-centric de novo design of protein binders,” *Journal of Molecular Biology*, vol. 413, no. 5, pp. 1047–1062, 2011, doi: 10.1016/j.jmb.2011.09.001.

[45] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, 2017, doi: 10.1021/acs.jctc.7b00125.

[46] R. Chowdhury, M. F. Allan, and C. D. Maranas, “OptMAVEn-2.0: De novo Design of Variable Antibody Regions Against Targeted Antigen Epitopes,” *Antibodies*, vol. 7, no. 3, p. 23, 2018, doi: 10.3390/antib7030023.

[47] G. Moroncini *et al.*, “Motif-grafted antibodies containing the replicative interface of cellular PrP are specific for PrP^{Sc},” *Proc Natl Acad Sci U S A*, vol. 101, no. 28, pp. 10404 LP – 10409, Jul. 2004, doi: 10.1073/pnas.0403522101.

[48] J. M. Perchiacca, A. R. A. Ladiwala, M. Bhattacharya, and P. M. Tessier, “Structure-based design of conformation- and sequence-specific antibodies against amyloid β,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 1, pp. 84 LP – 89, Jan. 2012, doi: 10.1073/pnas.1111232108.

[49] C. F. Barbas, L. R. Languino, and J. W. Smith, “High-affinity self-reactive human antibodies by design and selection: targeting the integrin ligand binding site,” *Proceedings of the National Academy of Sciences*, vol. 90, no. 21, pp. 10003 LP – 10007, Nov. 1993, doi: 10.1073/pnas.90.21.10003.

[50] O. Stoevesandt and M. J. Taussig, “Phospho-specific antibodies by design,” *Nature Biotechnology*, vol. 31, p. 889, Oct. 2013.

[51] S. M. Lippow, K. D. Wittrup, and B. Tidor, “Computational design of antibody-affinity improvement beyond in vivo maturation,” *Nature Biotechnology*, vol. 25, p. 1171, Sep. 2007.

[52] K. Tharakaraman *et al.*, “Redesign of a cross-reactive antibody to dengue virus with broad-spectrum activity and increased in vivo potency,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 17, p. E1555 LP-E1564, Apr. 2013, doi: 10.1073/pnas.1303645110.

[53] K. E. Tiller and P. M. Tessier, “Advances in Antibody Design,” *Annual Review of Biomedical Engineering*, vol. 17, no. 1, pp. 191–216, Dec. 2015, doi: 10.1146/annurev-bioeng-071114-040733.

[54] S. Fischman and Y. Ofran, “Computational design of antibodies,” *Current Opinion in Structural Biology*, vol. 51, pp. 156–162, 2018, doi: https://doi.org/10.1016/j.sbi.2018.04.007.

[55] D. Kuroda, H. Shirai, M. P. Jacobson, and H. Nakamura, “Computer-aided antibody design,” *Protein Engineering, Design and Selection*, vol. 25, no. 10, pp. 507–522, Jun. 2012, doi: 10.1093/protein/gzs024.

[56] G. D. Lapidoth *et al.*, “AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences,” *Proteins: Structure, Function and Bioinformatics*, vol. 83, no. 8, pp. 1385–1406, 2015, doi: 10.1002/prot.24779.

[57] M. C. Saraf, G. L. Moore, N. M. Goodey, V. Y. Cao, S. J. Benkovic, and C. D. Maranas, “IPRO: An iterative computational protein library redesign and optimization procedure,” *Biophysical Journal*, vol. 90, no. 11, pp. 4167–4180, 2006, doi: 10.1529/biophysj.105.079277.

[58] K. C. Entzminger *et al.*, “De novo design of antibody complementarity determining regions binding a FLAG tetra-peptide,” *Scientific Reports*, vol. 7, no. 1, p. 10295, 2017, doi: 10.1038/s41598-017-10737-9.

[59] R. J. Pantazes and C. D. Maranas, “MAPs: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies,” *BMC Bioinformatics*, vol. 14, no. 1, p. 168, 2013, doi: 10.1186/1471-2105-14-168.

[60] V. G. Poosarla, T. Li, B. C. Goh, K. Schulten, T. K. Wood, and C. D. Maranas, “Computational de novo design of antibodies binding to a peptide with high affinity,” *Biotechnology and Bioengineering*, vol. 114, no. 6, pp. 1331–1342, Jun. 2017, doi: 10.1002/bit.26244.

[61] S. J. Fleishman *et al.*, “Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin,” *Science (1979)*, vol. 332, no. 6031, pp. 816 LP – 821, May 2011, doi: 10.1126/science.1202617.

[62] X. Liu *et al.*, “Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping,” *Scientific Reports*, vol. 7, p. 41306, Jan. 2017.

[63] R. Chen, L. Li, and Z. Weng, “ZDOCK: An initial-stage protein-docking algorithm,” *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 1, pp. 80–87, Jul. 2003, doi: https://doi.org/10.1002/prot.10389.

[64] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, “HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information,” *J Am Chem Soc*, vol. 125, no. 7, pp. 1731–1737, Feb. 2003, doi: 10.1021/ja026939x.

[65] S. Lyskov and J. J. Gray, “The RosettaDock server for local protein-protein docking,” *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. W233–W238, Jul. 2008, doi: 10.1093/nar/gkn216.

[66] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, “ClusPro: a fully automated algorithm for protein–protein docking,” *Nucleic Acids Research*, vol. 32, no. suppl_2, pp. W96–W99, Jul. 2004, doi: 10.1093/nar/gkh354.

[67] M. Torchala, I. H. Moal, R. A. G. Chaleil, J. Fernandez-Recio, and P. A. Bates, “SwarmDock: a server for flexible protein–protein docking,” *Bioinformatics*, vol. 29, no. 6, pp. 807–809, Mar. 2013, doi: 10.1093/bioinformatics/btt038.

[68] G. Nimrod *et al.*, “Computational Design of Epitope-Specific Functional Antibodies,” *Cell Reports*, vol. 25, no. 8, pp. 2121-2131.e5, 2018, doi: https://doi.org/10.1016/j.celrep.2018.10.081.

[69] R. J. Pantazes, M. J. Grisewood, T. Li, N. P. Gifford, and C. D. Maranas, “The Iterative Protein Redesign and Optimization (IPRO) suite of programs,” *Journal of Computational Chemistry*, vol. 36, no. 4, pp. 251–263, Feb. 2015, doi: https://doi.org/10.1002/jcc.23796.

[70] G. Teng and F. N. Papavasiliou, “Immunoglobulin Somatic Hypermutation,” *Annual Review of Genetics*, vol. 41, no. 1, pp. 107–120, Dec. 2007, doi: 10.1146/annurev.genet.41.110306.130340.

[71] M. Arslan, D. Karadağ, and S. Kalyoncu, “Protein engineering approaches for antibody fragments: directed evolution and rational design approaches,” *Turkish journal of biology = Turk biyoloji dergisi*, vol. 43, no. 1, pp. 1–12, Feb. 2019, doi: 10.3906/biy-1809-28.

[72] J. Dou *et al.*, “De novo design of a fluorescence-activating β-barrel,” *Nature*, vol. 561, no. 7724, pp. 485–491, 2018, doi: 10.1038/s41586-018-0509-0.

[73] C. Longxing *et al.*, “De novo design of picomolar SARS-CoV-2 miniprotein inhibitors,” *Science (1979)*, vol. 370, no. 6515, pp. 426–431, Oct. 2020, doi: 10.1126/science.abd9909.

[74] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[75] E. O. Saphire *et al.*, “Crystal Structure of an Intact Human IgG: Antibody Asymmetry, Flexibility, and a Guide for HIV-1 Vaccine Design BT - Glycobiology and Medicine,” 2003, pp. 55–66.

2. Chapter 2 - Antibody-antigen database

As mentioned in Chapter 1, antibodies recognize epitopes on antigens and bind specifically and strongly at these locations. This specificity is achieved through strong structural fitting of the interacting protein surfaces and the presence of hydrophobic, electrostatic and vdW interactions which stabilize the binding interface. The knowledge of antibody binding interface features can be used in computational protein engineering tasks such as designing novel antibodies that bind to their target molecules and predicting target binding regions of given antibody and antigen structures. This Chapter will provide a background on the work done in this field of research and present the methodology and results of a multi-force field analysis of a large database of antibody-antigen complexes. The contents in this Chapter have been adapted from our publication [1].

Background

For over two decades, features of antibody-antigen binding interfaces have been studied in detail using the crystal structures of antibody-antigen complexes. Such studies have been made possible by the advent of high resolution imaging tools such as X-ray diffraction and NMR spectroscopy. Initially, Mian et al. studied six antibody structures which included only one antigen [2]. Work by Lo Conte et al. and Sundberg and Mariuzza studied 19 and 30 complexes respectively [3] [4]. The former analyses studied features of the CDR residues that bind to the antigen (i.e., paratope) in more detail when compared with epitope properties. W. Chen et al. studied interaction methods and molecular surface properties of both paratopes and epitopes in 155 protein-antibody complexes, with 53 unique structures [5]. The work of Fellouse et al. constructed paratope libraries with a diversity of only four amino acids and studied the features of two complexes for the generation of synthetic antibody libraries [6]. Shape complementarity (Sc) is the widely used

binding metric that accounts for the conformational fitting of the two interacting surfaces [7]. Kuroda and Gray studied the features of hydrogen bonds and Sc in 6637 non-redundant protein-protein interfaces which included 191 antibody-antigen complexes [8]. An average Sc score of 0.70 +/- 0.06 was reported for antibody-antigen complexes. Nguyen et al. analyzed 403 antibody-antigen complexes using computational methods to find hydrogen bonds and vdW, hydrophobic and ionic interactions. The authors reported that the binding interfaces are enriched in short chain hydrophilic residues while TYR is the most frequent amino acid making the highest number of hydrogen bonds. Having observed waters ranging from 5 to 25 molecules in binding interfaces, the authors believe that the interfaces are stabilized by these molecules through hydrogen binding and by negating any poor shape complementarity between the epitope and paratope. Martineau et al. analyzed 227 antibody antigen complexes using the FoldX force field [9]. After calculating the change in free energy due to ALA mutations at each residue position in the variable domains, the authors identified that 8 residues in 30 key positions made an average contribution to 80% of the binding energy. Moreover, TYR was found in abundant frequency at the key paratope positions followed by GLY, SER, TRP, ASP and ASN. The importance of TYR in antibody binding interfaces was also reported in the work by Mumey et al. in their analysis of a database of 53 non-redundant antibody-antigen structures [10]. The authors also studied spatial and geometrical features of the interface such as the gap volume index, dimensions and shape.

The previous studies have identified numerous features about paratopes, including which amino acids are most commonly used and the types of interactions that can occur. Similar to the work of Martineau et al., which used only one force field, antibody-antigen interfaces were analyzed using more force fields without any ALA mutations. CHARMM36, AMBER and Rosetta

molecular mechanics force fields were used to quantify the percent contribution to binding of every paratope residue and residue type in a database of 498 antibody – protein complexes.

Methodology

Initially, 2498 antibody structures were gathered from the International Immunogenetics Information System 3D Structure Database [11]. 1344 antibodies with at least five CDR mutations from one another were identified, of which 492 were in complex with protein/peptide antigens. Those structures were selected for this study.

Missing atoms were added and the overall energies of each of the 492 complexes were minimized in each force field. Next, the energies of interaction between every antibody residue and every antigen residue were tabulated. Due to the pairwise additive nature of the energy functions and known inaccuracies in the approximations inherent in their formulations, the total energy can vary significantly from complex to complex and generally increases in magnitude as the total number of amino acids increases. Therefore, the energies were normalized to the percent of total energy for each complex to facilitate comparisons between complexes.

Results

The percent contribution to binding energy of residues was determined to follow a logarithmic growth for all the force fields, as shown in Figures 2.1, 2.2 and 2.3. The R^2 of the logarithmic fits for antibodies were 0.986, 0.993 and 0.989 with the AMBER, CHARMM36 and Rosetta force fields, respectively. Using CHARMM36, AMBER and Rosetta, the first five residues contribute to an average of 59.2%, 52.4% and 58.0% of the total binding energy respectively. Similar observations have been made computationally by Martineau et al. [9]. Also, the percent

contributions to binding energy are marginally greater using CHARMM36 and Rosetta than with AMBER. Without experimental data, it is not clear which is the most accurate prediction, but all three force fields show the same logarithmic growth trend. Similar to this finding, experimental studies have demonstrated that in most complexes, five antigen residues are sufficient for specific and high affinity binding [12], [13]. There are, on average, > 40 CDR residues and >200 residues in the variable domains. Comparing these numbers to the results in Figures 2.1, 2.2 and 2.3 show us that a minute fraction of antibody residues contribute to a majority of the binding affinity. Thus, the top five paratope residues in each complex were defined as “significant residues”.

Figure 2.4 depicts the fractional frequency of amino acid usage in the significant residues versus the CDRs as a whole for all three force fields. The CDR residue trends do not follow any specific trend with respect to hydrophobicity of the amino acids types. Hydrophobic residues PHE and ILE have higher frequencies than hydrophilic residues such as LYS, GLU and ARG. This observation hints at the importance of the structural contribution of each residue to the CDR loop shape over its hydrophobicity even though CDR loops are surface exposed structures. SER is the most frequent CDR amino acid since it can play dual roles: structural through its small side chain which offers high conformational flexibility [14] and functional through its hydrophilic hydroxyl group. TYR is the second most frequent CDR residue due to its high propensity to engage in binding interactions as mentioned later. GLY is the third most frequent CDR amino acid since it plays a structural role similar to SER along with offering its backbone atoms for hydrogen bonding. The hydrophilic residues with small side chains, THR and ASP, are the fourth and fifth most abundant CDR residues.

The trends regarding significant residues are consistent with expected behaviours: charged and polar amino acids which can participate in binding interactions are overrepresented in the

significant residues relative to hydrophobic residue types when compared with CDR residue preferences. HIS and GLN are underrepresented in significant residues in spite of their polar side chains. This could partly be due to the presence of large hydrophobic hydrocarbon groups in the side chains along with the polar groups. Furthermore, significant residues are also depleted in SER due to the structural role it plays in CDR loops as mentioned before. The top five amino acid frequencies in significant residues for CHARMM36, AMBER and Rosetta are ASP>TYR>ARG>SER>GLU, ASP>TYR>ARG>GLU>SER and TYR>ASP>ARG>TRP>PHE respectively. Amongst the hydrophilic residues, TYR is the most significant residue type by a large margin using Rosetta and ranks second behind ASP by a small margin with the AMBER and CHARMM36 force fields. TYR has 0.4% and 6% lower frequency than ASP according to AMBER and CHARMM36 respectively but 81% higher frequency in Rosetta calculations. Similar observations regarding TYR's importance in paratopes have been made earlier in several publications [6] [15]. The significance of TYR for binding can be attributed to the presence of an aromatic ring and hydroxyl group in its side chain, which lends it the ability to engage in both π stacking and hydrogen bonds with epitope atoms. ASP is the second most significant residue due to its ability to engage in hydrogen and ionic bonds through its anionic oxygen atoms. Furthermore, the presence of only a single methylene group does not offer any major hydrophobic effects. ARG is the third most significant residue types according to all the force fields. Its positively charged side chain can participate in both hydrogen bonds and ionic interactions. There were major differences between the force fields for the fourth and fifth most abundant significant residue type selections. While CHARMM36 and AMBER selected hydrophilic residue types SER and GLU, Rosetta selected aromatic residue types TRP and PHE. This difference might be a result of CHARMM36 and AMBER's inability to adequately capture the types of interactions aromatic

rings can engage in: cation-π, π-π and hydrophobic interactions. Alternatively, Rosetta's parameterization to effectively simulate the hydrophobic cores of protein structures may have also led to this contradictory behaviour between force fields.

In order to compare the performance of the force fields with regards to the different amino acids, the similarities between the amino acid fractional frequencies for each force field were studied. To do this, the absolute difference of the percent deviations from the average between each pair of force field results for each amino acid and the average of these percent deviations for each force field comparison were calculated. The results are detailed in Table 2.1, which shows that AMBER and CHARMM36 give similar results when compared with the results from Rosetta. This was expected since both AMBER and CHARMM36 follow similar energy formulations to calculate the enthalpic and solvation energies while Rosetta is heavily dependent on statistical parameters obtained from large protein structure databases.

Discussion

In the context of protein engineering, the key lesson that was taken from these analyses is that a handful of residues, either hydrophilic or aromatic, contribute the majority of the binding energy. Thus, a straightforward antibody design approach could be developed that prioritizes the presence of these few residues in its CDR loops making strong interactions with the epitope in the form of either hydrogen bonds, cation-π stacking or π-π stacking interactions.

I developed a non-redundant database of 498 antibody-protein complexes and used the CHARMM36, AMBER and Rosetta force fields to energetically minimize and analyze the binding interactions in these complexes. This allowed for the quantification of the contributions to binding of single residues in antibodies and antigens. The primary finding is that the cumulative

contributions follow a logarithmic growth, with only a few residues contributing most of the binding energy. This data should aid in the development of computational methods to design binding proteins and in the development of diagnostics, vaccines, and therapeutics.

Table 2.1 Differences between percent deviations from average calculated using CHARMM, AMBER and Rosetta force fields for each amino acid.

	CHARMM vs AMBER	CHARMM vs Rosetta	AMBER vs Rosetta
TYR	13.18	15.83	29.00
ASP	9.34	42.77	33.42
ARG	0.66	45.90	45.24
SER	15.13	60.29	45.15
TRP	33.54	113.34	79.80
ASN	16.54	7.22	23.75
GLU	3.35	62.10	58.75
PHE	31.99	121.97	89.98
THR	2.44	10.35	12.80
LYS	5.07	53.70	58.77
GLY	1.13	86.04	84.92
GLN	34.01	45.21	11.20
LEU	4.60	147.96	113.36
ILE	9.67	132.45	122.78
VAL	27.62	167.46	139.84
HIS	44.30	9.88	54.18
ALA	4.13	121.43	125.56
PRO	51.73	91.28	39.55
MET	0.43	74.88	75.31
CYS	43.19	85.53	128.72
Average	19.10	74.78	68.60

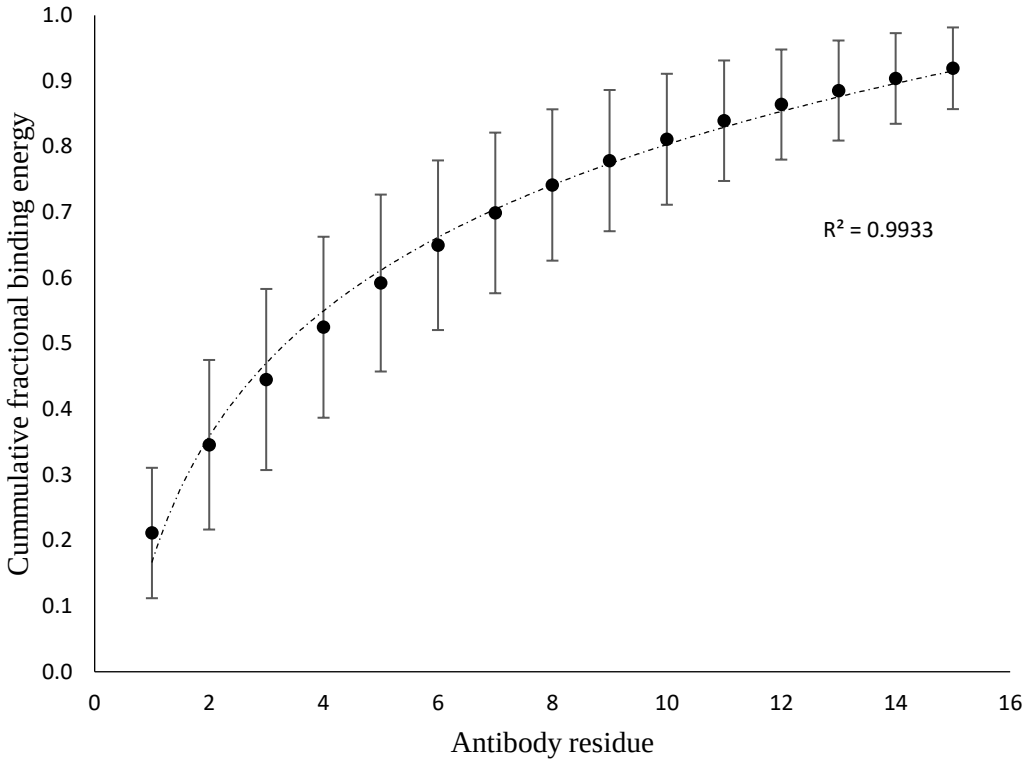

Figure 2.1 Average cumulative fractional binding energy calculated using CHARMM36 force field.

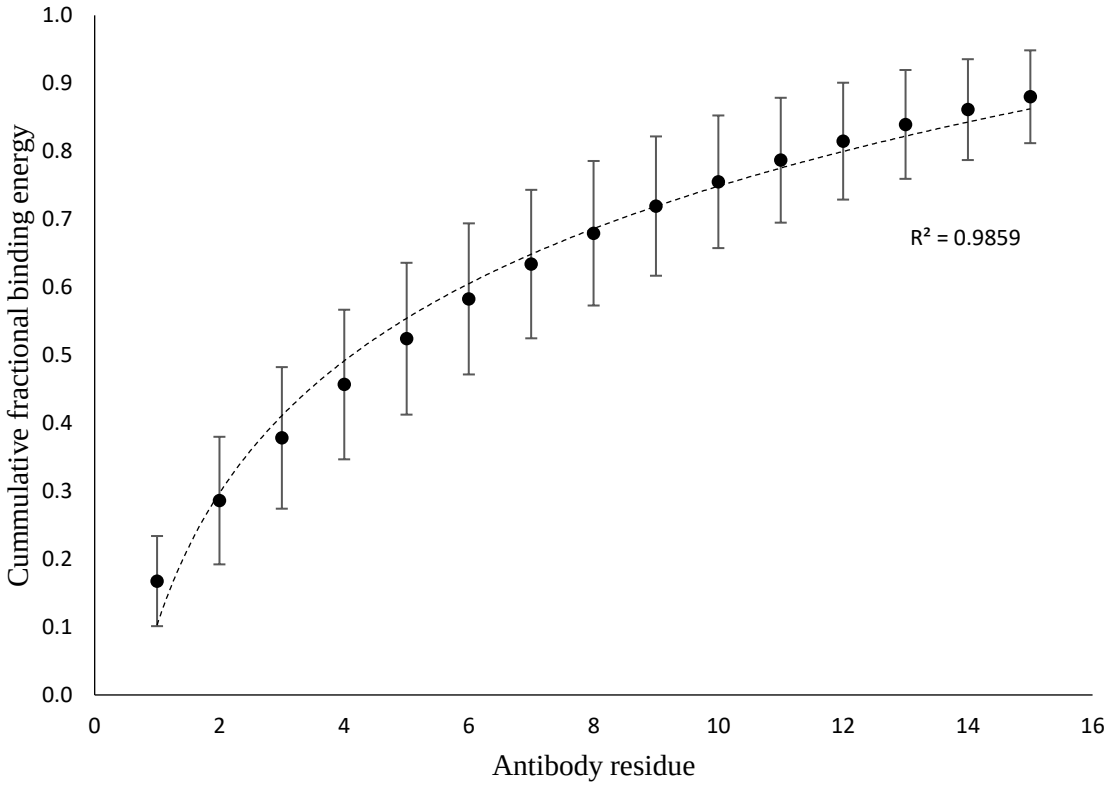

Figure 2.2 Average cumulative fractional binding energy calculated using AMBER force field.

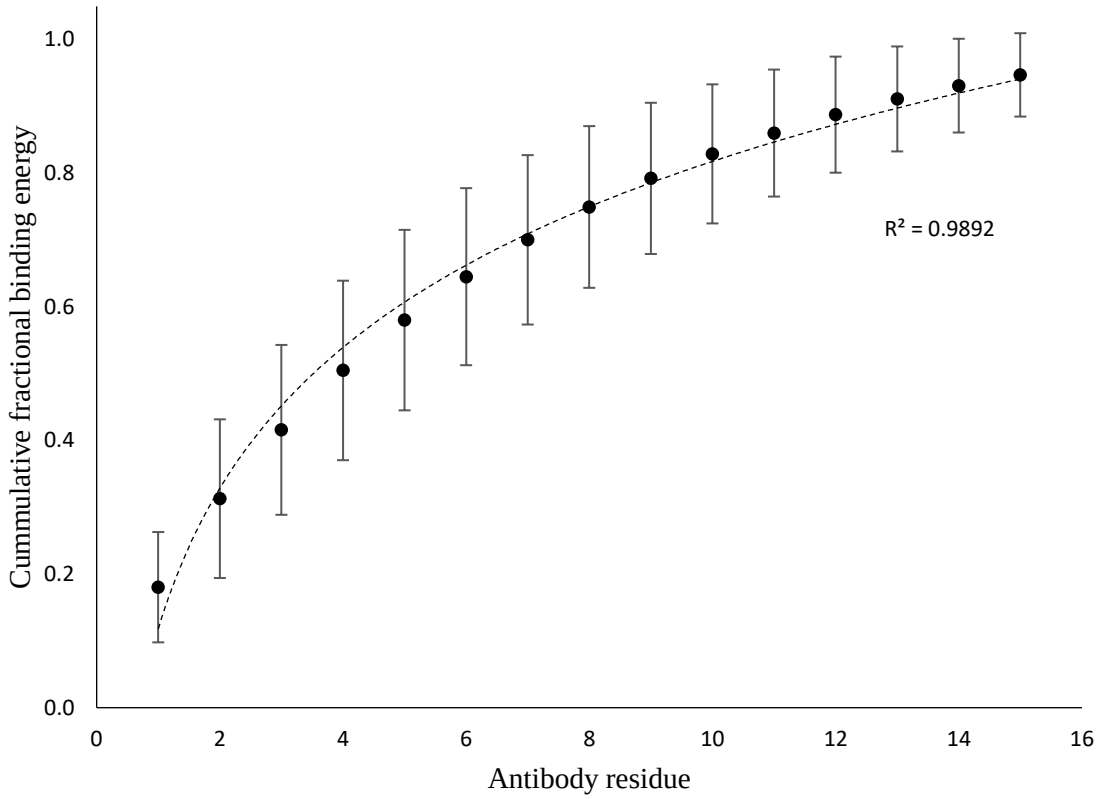

Figure 2.3 Average cumulative fractional binding energy calculated using Rosetta force field.

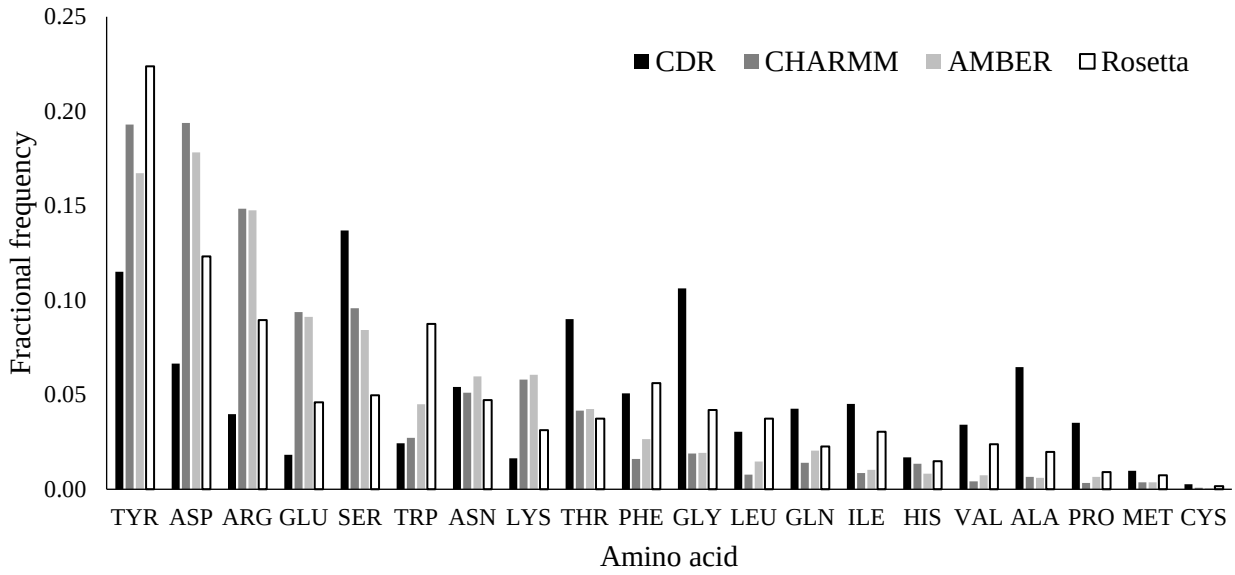

Figure 2.4 Fractional frequency of each amino acid obtained using CHARMM36, AMBER and Rosetta force fields.

References

[1] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” *Computer Aided Chemical Engineering*, vol. 44, pp. 2113–2118, 2018.

[2] I. S. Mian, A. R. Bradwell, and A. J. Olson, “Structure, function and properties of antibody binding sites,” *Journal of Molecular Biology*, vol. 217, no. 1, pp. 133–151, 1991, doi: https://doi.org/10.1016/0022-2836(91)90617-F.

[3] L. Lo Conte, C. Chothia, and J. Janin, “The atomic structure of protein-protein recognition sites11Edited by A. R. Fersht,” *Journal of Molecular Biology*, vol. 285, no. 5, pp. 2177–2198, 1999, doi: https://doi.org/10.1006/jmbi.1998.2439.

[4] E. J. Sundberg and R. A. B. T.-A. in P. C. Mariuzza, “Molecular recognition in antibody-antigen complexes,” in *Protein Modules and Protein-Protein Interaction*, vol. 61, Academic Press, 2002, pp. 119–160. doi: https://doi.org/10.1016/S0065-3233(02)61004-6.

[5] S. W. Chen and M. H. V. V. R. and J.-L. Pellequer, “Structure-Activity Relationships in Peptide-Antibody Complexes: Implications for Epitope Prediction and Development of Synthetic Peptide Vaccines,” *Current Medicinal Chemistry*, vol. 16, no. 8. pp. 953–964, 2009. doi: http://dx.doi.org/10.2174/092986709787581914.

[6] F. A. Fellouse, C. Wiesmann, and S. S. Sidhu, “Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12467–12472, Aug. 2004, doi: 10.1073/pnas.0401786101.

[7] M. C. Lawrence and P. M. Colman, “Shape Complementarity at Protein/Protein Interfaces,” *Journal of Molecular Biology*, vol. 234, no. 4, pp. 946–950, 1993, doi: https://doi.org/10.1006/jmbi.1993.1648.

[8] D. Kuroda and J. J. Gray, “Shape complementarity and hydrogen bond preferences in protein-protein interfaces: implications for antibody modeling and protein-protein docking,” *Bioinformatics (Oxford, England)*, vol. 32, no. 16, pp. 2451–2456, Aug. 2016, doi: 10.1093/bioinformatics/btw197.

[9] G. Robin, Y. Sato, D. Desplancq, N. Rochel, E. Weiss, and P. Martineau, “Restricted Diversity of Antigen Binding Residues of Antibodies Revealed by Computational Alanine Scanning of 227 Antibody–Antigen Complexes,” *Journal of Molecular Biology*, vol. 426, no. 22, pp. 3729–3743, 2014, doi: https://doi.org/10.1016/j.jmb.2014.08.013.

[10] T. Ramaraj, T. Angel, E. A. Dratz, A. J. Jesaitis, and B. Mumey, “Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures,” *Biochimica et biophysica acta*, vol. 1824, no. 3, pp. 520–532, Mar. 2012, doi: 10.1016/j.bbapap.2011.12.007.

[11] F. Ehrenmann, Q. Kaas, and M. P. Lefranc, “IMGT/3dstructure-DB and IMGT/domaingapalign: A database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MHcSF,” *Nucleic Acids Research*, vol. 38, no. SUPPL.1, 2009, doi: 10.1093/nar/gkp946.

[12] K. F. Sykes, J. B. Legutki, and P. Stafford, “Immunosignaturing : a critical review,” *Trends in Biotechnology*, vol. 31, no. 1, pp. 45–51, 2013.

[13] J. P. Pellois, X. Zhou, O. Srivannavit, T. Zhou, E. Gulari, and X. Gao, “Individually addressable parallel peptide synthesis on microchips,” *Nature Biotechnology*, vol. 20, no. 9, pp. 922–926, 2002, doi: 10.1038/nbt723.

[14] M. van Rosmalen, M. Krom, and M. Merkx, “Tuning the Flexibility of Glycine-Serine Linkers To Allow Rational Design of Multidomain Proteins,” *Biochemistry*, vol. 56, no. 50, pp. 6565–6574, Dec. 2017, doi: 10.1021/acs.biochem.7b00902.

[15] F. A. Fellouse, P. A. Barthelemy, R. F. Kelley, and S. S. Sidhu, “Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code,” *Journal of Molecular Biology*, vol. 357, no. 1, pp. 100–114, 2006, doi: 10.1016/j.jmb.2005.11.092.

3. Chapter 3 - AUBIE

Building on the observation from Chapter 2 that more than half of the binding energy in antibodies is contributed by five paratope residues, I hypothesized that a design strategy focusing on engineering several key, strong interactions between the target epitope and paratope would simplify and reduce the computational load. To this end, I have developed and tested the Algorithm for Ultra-rapid Binding Interaction Engineering (AUBIE) for the *de novo* design of binding loops for antibodies and alternative scaffolds that bind to a specific antigen epitope. AUBIE quickly designs libraries of PDB structure files that are expected to have strong binding interactions with the target epitope. AUBIE was used to generate libraries of antibody structures that bind to the human epidermal growth factor receptor 2 (HER2) and the resulting predictions were experimentally tested. The Chapter will be provide a description of the AUBIE methodology and case study results. This work will be submitted for publication in 2022.

Methodology

AUBIE is divided into two main steps: database generation and protein design. In the first step, the Protein Data Bank is scanned for loops that can fit into the binding regions of the chosen framework and strong interaction regions around these binding loops are identified. The second step includes the positioning of the antigen around the framework and geometrically identifying which compatible binding loops from the database align their strong interaction regions with relevant epitope atoms.

Binding loops database generation

The required input information for this step is the PDB file of the binding protein scaffold, the first and last residue information of each binding region, the maximum number of binding loop

structures and the minimum and maximum residue lengths of the binding loops searched. Here, a binding region is defined as the segment of the scaffold structure that will be replaced by loops, referred to as binding loops, obtained from the PDB (Figure 3.1, Step 1). In the first step of database generation, AUBIE searches for binding loops that fit into the specified scaffold binding regions. AUBIE uses the entire PDB available as of 3/13/2018 as its source of binding loops. The PDB was refined by purging it of any incomplete protein chains, heteroatoms and chains that contained non-natural amino acids. All the structures in the PDB are searched for protein loops for each binding region that meet a geometric grafting criteria and are compatible with the scaffold. The geometric criteria entails that the distances between backbone atom pairs in the first and last residues of the loop must be within a small threshold distance to the corresponding atom pair distances in the specified attachment point residues in the scaffold. AUBIE uses the following backbone atom pairs: $C\alpha_i$-$C\alpha_f$, $C\beta_i$-$C\beta_f$, $C\alpha_i$-$C\beta_f$, $C\beta_i$-$C\alpha_f$, N_i-N_f, O_i-O_f and C_i-C_f where ‘i’ and ‘f’ refer to the initial and final attachment point residues (Figure 3.1, Step 2). This approach ensures continuity in the backbone dihedral angles from the scaffold into the binding loop and hence the stability of the grafted loop. There is a variability in distance between the binding loop attachment points since biomolecules are in constant motion. Hence, a deviation cutoff is utilized while comparing the pairwise distances. AUBIE uses a default deviation of 0.225 Å. The variability of protein structures can be captured through MD simulations. Users have the option of providing a set of scaffold structures obtained from a MD simulation, which will be used to determine the starting scaffold structure and attachment residue distance variability. In this case, the scaffold structure is then determined by first calculating the average positions for all the atoms from all the MD structures provided and then selecting one of the input structures that is most similar to the

average structure. The standard deviations are then used as the backbone atom pairwise distance deviations.

The selected binding loops are checked with steric clash and energy filters to ensure their compatibility with other loops and the constant part of the scaffold i.e., the framework (Figure 3.1, Step 3). The loops that pass the following two criteria are eventually stored in the binding loop database.

1. The binding loop must be compatible with the framework structure. The antibody-antigen database from Chapter 2 had revealed that 10% of CDRs had more than 16 steric clashes (i.e., at least one pair of non-Hydrogen atoms, also known as heavy atoms, are less than the sum of their vdW radii apart) and more than 237.5 kcal/mol of interaction energy with the framework. Binding loops are classified as framework compatible if they have lower steric clashes and interaction energy than these metrics. The interaction energy was calculated as the sum of the electrostatic and vdW energy calculated using the CHARMM22 force field [1].
2. The binding loop must be compatible with at least one binding loop from all the other binding regions. This way it was ensured that every binding loop can be part of a final solution. Using the antibody-antigen database, it was determined that 10% of CDRs have more than three steric clashes and 55.6 kcal/mol energy interaction with the other CDRs. Hence loops that do not pass these steric and interaction energy criteria are deleted. The compatibility information of each binding loop is stored for future use during the later protein design portion of AUBIE.

The framework is positioned at the origin with the binding loops oriented in the positive z direction. AUBIE does this by positioning the centroid of the attachment point backbone atoms at

the origin and ensuring that the best fit plane that passes through these backbone atoms has zero x and y axes slopes. Each binding loop identified from the previous step is positioned so that the backbone atoms of the first and last residues match the position of their respective backbone atoms in the attachment point residues in the framework.

As mentioned before, the AUBIE approach attempts to engineer a few strong interactions between the designed protein and target epitope. It does so by identifying specific regions around binding loop residues for epitope atoms to occupy to make strong binding interactions. In this step, knowledge of the geometries of optimal attractive interactions is used to generate a library of strong binding regions (SBRs) for each of the binding loops identified in the previous step (Figure 3.1, Step 4). The library will contain SBRs for all rotamers of all the residues in the binding loops since residue side chains are expected to repack to better interact with the epitope. The strong interaction types covered in the library are multiple H-bonds between charged residues, single hydrogen bonds and various orientations of π-π and cation-π interactions. Previous work has shown that the lowest energy conformations for π-π stacking follow the parallel displaced and T shaped arrangements [2]. Hence these interaction geometries were incorporated into AUBIE. Regarding cation-π interactions, the lowest energy conformations incorporated into AUBIE include the parallel and T-shaped arrangements shown in Figure 3.2. Table 3.1 gives information regarding the paratope and epitope atoms and paratope residues for each interaction type.

For each interaction in Table 3.1, positional information about the epitope atoms that are important for binding and bond orientation is obtained and stored. This is done using geometrical constraints obtained from literature and the antibody-antigen database as shown in Table 3.2. The distance information for hydrogen bonds was obtained from the antibody-antigen database, while the remaining information was obtained from literature as detailed in Table 3.2. The geometric

constraints of all SBRs consist of a fixed position for an atom and secondary constraints such as a fixed position for another atom, fixed planar presence for an atom or fixed planar presence for a group of atoms. The secondary constraints for each SBR depend on the interaction type and flexibility of the orientation and the epitope side chain involved. These constraints need to ensure that all possible conformations of the strong interactions are covered. SBRs with secondary constraints as fixed positions for two atoms (SBRs 9, 11 and 15) can be occupied by only a single epitope residue conformation. On the contrary, several epitope residue conformations can meet the requirements of other SBRs. SBRs 1, 2, 5 and 6 simulate single H-bond between paratope Hydrogen and epitope Oxygen. A fixed position for bond acceptor, oxygen, and a plane for the acceptor antecedent, carbon, are determined. A plane is required to capture the two optimal positions of the carbon atom due to the presence of two electron lone pairs of the oxygen atom, both of which could participate in the hydrogen bond. Similarly, SBR 10 simulates H-strong H-bond complex with paratope LYS and epitope acids. A plane is required to capture the two oxygen atom positions along with a fixed position of the carbon atom that is bonded to the oxygen atoms. For SBRs 7 and 8, which simulate H-bond between paratope Oxygen and epitope Hydrogen, the fixed positions of the hydrogen and donor atoms are determined for both the electron lone pairs on the acceptor oxygen atoms. For SBRs 11 and 15, which need to simulate epitope ARG positions, a fixed conformation of the guanidinium group is determined. Similarly, in SBR 9, fixed conformation of the carboxylic acid group is determined in order to simulate strong H-bond complex with paratope ARG. For SBRs 12 and 16, the point coordinates of the nitrogen and carbon atoms in the epitope LYS are determined. For SBRs to be occupied by aromatic residues for either cation-π or π-π interactions, the constraints include a fixed coordinate for either an atom (SBRs 17 and 23) or the center of the aromatic ring (SBRs 13, 14, 19, 20, 21 and 22) along with a planar

constraint for all the six ring atoms. SBR 18, which represents a particular conformation of T-shaped π-π stacking, is the only exception since it consists of two point constraints: a ring atom and the ring center. For π-π or parallel cation-π interactions such as SBRs 13, 17, 18, 20, 22, and 23, SBRs are computed twice: above and below the ring. SBRs 3 and 4 are currently not part of the algorithm and are incorporated in future versions of AUBIE in Chapter 5. The database of compatible binding loops and optimal binding positions needs to be generated only once for each scaffold. The same database can be used for multiple design runs with that scaffold.

Protein design

In the AUBIE protein design workflow, the binding loop database is systematically searched for loop combinations that make strong interactions with the epitope without having any steric or charge clashes with each other and the antigen. The input information required for this step is the PDB file of the antigen, the epitope residue numbers, the minimum number of strong interactions required in each solution and the maximum number of solutions needed to be identified. The first step of the workflow orients the epitope to point at the binding pocket (i.e., the epitope points along the negative z-axis, Figure 3.3, Step 1). The binding protein framework and binding loops, which are already positioned around the origin and oriented towards the positive z direction facing the epitope, are loaded from the database. If the binding protein, which is comprised of the framework and binding loops, is fixed, the antigen has six degrees of freedom, three translational and three rotational, to search for binding conformations around the binding protein. The three translational degrees of freedom are movements along the three Cartesian axes and the three rotational degrees of freedom comprise the three Euler rotations around each of the three axes. Large rotations around the x and y axes would yield incompatible conformations since AUBIE positions the epitope to face in the negative z direction, which in turn would lead to steric

clashes between the antigen and binding protein. AUBIE rotates the antigen around the z axis in one degree steps and for each rotation, simultaneously searches for the three translational movements of the antigen that would yield strong affinity binding protein-antigen complexes (Figure 3.e, Step 2).

For each rotation, AUBIE employs a two-step workflow to identify sets of compatible binding loops. First, AUBIE identifies combinations of epitope residues that can all simultaneously align with the minimum required SBRs. Subsequently, the second step checks if the binding loops of the paratope residues are compatible with each other, the framework and the antigen.

AUBIE searches for congruent pairings of SBR and epitope residues (Figure 3.3, Step 3) by checking if the secondary SBR coordinates are less than 0.33 Å away from their respective epitope atoms after aligning the SBR primary coordinate with the respective epitope atom. For example, in SBR type 1, after matching the epitope oxygen atom with the primary SBR coordinate, it is ensured that the carbon atom of that epitope residue is less than 0.33 Å away from the SBR plane. This step applies to all SBRs since all SBR types consist of a primary atom coordinate along with other secondary coordinates which could either belong in a plane or point(s). A congruent movement signifies a strong interaction between the SBR paratope residue and epitope residue. For a group of congruent movements to be classified as a probable solution (Figure 3.5, Step 4), it must satisfy the following criteria.

1. The translational movements of each pair of congruent movements in the group must not differ by more than 0.33 Å. This ensures that all the epitope residues in the group simultaneously occupy their respective SBRs and, at the same time, allows for some wiggle room for the placement of the epitope atoms in the SBRs.

2. No two congruent movements must be harbored by binding loops that have steric clashes. The compatibility information between the binding loops is pre-calculated during the database generation step. Moreover, if two congruent movements are from SBRs belonging to the same residue, then one of them has to be backbone atom interaction. Congruent movements from the same binding loop are also allowed.
3. If both the congruent movements include a point coordinate and a plane in their SBRs (SBRs 1, 2, 5, 6 and 10), additional verification is performed that the primary points of the SBRs are more than 2 Å apart to avoid possible clashes of the secondary points.

Figure 3.4 shows three movements, two of which are congruent, to illustrate this concept. Following the identification of groups of compatible congruent movements, groups that have binding loop similarity over 60% compared to other solution groups are purged to ensure a rich diversity in the final library of the solutions.

Each solution group identified in the previous step is further analyzed for compatibility (Figure 3.3, Step 5).

1. Initially, the antigen is positioned so that the epitope atoms of the congruent movements match with their respective SBRs. This is done by translationally moving the antigen by the average x, y and z movements of all the congruent movements in the group.
2. After the antigen is positioned, two compatibility checks are performed: between the binding loops and the antigen and between the framework and the antigen. The compatibility checks are performed by computing steric clashes and electrostatic clashes. A steric clash is defined when two protein backbone atoms from different structures are closer than 80% of their vdW radii. AUBIE allows steric clashes involving side chain atoms under the assumption that the side chains would repack in order to avoid the clash but retain

the binding affinity. Solution conformations that have more than two steric clashes are rejected. The shortest distance between two atoms with the same charge (i.e., between the positively charged nitrogen atoms of LYS and ARG and the negatively oxygen atoms of ASP and GLU) in the antibody-antigen database is 4.2 Å. Therefore, an electrostatic clash is defined when two atoms with the same charge are closer than 4.2 Å. Histidine is not accounted as a charged amino acid since AUBIE assumes that the binding takes place at normal pH.

3. In the scenario that a solution group does not have loops for a binding region in the framework, AUBIE scans the database for a binding loop belonging to the incomplete binding region that would pass the above mentioned compatibility checks. The PDB format file of the now complete binding protein and antigen is then generated and output to the user.

Designs that pass these compatibility checks are immediately output as a final successful solution in the PDB format (Figure 3.3, Step 6). In other words, AUBIE designs binding proteins that have a sufficient number of favorable interactions (i.e., strong binding interactions) and no obvious detrimental interactions, without doing energy calculations.

Results

In order to test AUBIE's performance, HER-2 binding antibodies were designed and experimentally tested. The HER2 antigen is overexpressed in breast cancer cells and is the target antigen for various therapeutics [3]. Trantuzumab (sold as Herceptin) is a HER2 binding therapeutic monoclonal antibody. The general epitope region selection for AUBIE and the experimental testing was done by Dr. Jamie Spangler's group at Johns Hopkins University. They

wished to create bispecific antibodies made of the known HER2 binding Herceptin Fab domain and a novel Fab domain that bound an adjacent epitope on HER2. The antigen structure and selected epitope are shown in Figure 3.5. For the database generation step, the CDRs and framework region were obtained from the Herceptin-SCFV (Single Chain Fragment Variable) structure found in PDB 4X4X. Herceptin was selected as the framework source owing to its known stability, solubility and experience of the Spangler group with it's experimental generation. The six CDRs were selected as the binding regions. The scaffold structure and CDR loop regions are shown in Figure 3.6.

For the database generation step, a maximum limit of 500 loops per binding region was set. The database generation step identified 494 loops for H1, 498 loops for H2, 489 loops for H3, 472 loops for L1, 498 loops for L2 and 481 loops for L3 in approximately 11.5 hours. The calculations were performed using an Intel Xeon E5-2660 processor with a speed of 2.60 GHz. The large number of candidate loops for each CDR ensures a large combinatorial solution space for the design step to search through. This antibody database is generated only once and can be used for any antibody design step. For the generation of AUBIE designs, two design runs were performed, each with unique epitopes. Multiple epitope selections allowed us to explore more final design orientations as the epitope dictates the antigen positioning before rotations. The two epitopes consisted of 23 and 18 residues respectively. The HER2 antigen structure was obtained from PDB 1N8Z. A minimum limit of six interactions and a maximum limit of 300 total solutions was set for all the design runs. The two AUBIE design runs generated 300 and 69 designs in 35.5 hours and three days respectively using the previously mentioned Intel Xeon processor. An initial test of AUBIE against the same 12 amino acid epitope bound by Herceptin generated 300 designs in less than 6 hours using the same processor. This demonstrates that AUBIE runtimes scale rapidly

as the size of the target epitope increases. Target epitope sizes of 23 and 18 residues are relatively large. Smaller epitopes sizes are enough to generate large solution sets, AUBIE is able to identify numerous solutions in a short time frame (i.e., < 10 hours) for reasonably sized (i.e., < 13 residues) epitopes.

All the AUBIE generated designs were run through all-atom energy minimizations using the CHARMM22 force field with the FACTS implicit solvation model [1]. CHARMM22 was used to initially minimize design energies since prior experience with the force field has shown that it is better at resolving side chain clashes than AMBER and Rosetta. The CHARMM22 minimized designs were subsequently minimized with AMBER's ff14SB [4] and Rosetta's REF2015 [5] force fields. Upon energy minimizations, the designs were analyzed for binding metrics such as buried surface area and binding energies. The AMBER binding energies were obtained by summing up pairwise residue energies between the antibody and antigen. The buried surface area and Rosetta binding energies were obtained by using the InterfaceAnalyzer module from Rosetta [6]. Analysis of such metrics for binding affinity prediction has been done in previous antibody design work [7]. Table 3.4 lists the average values for these metrics from the AUBIE generated designs and the antibody-antigen database, as well as the number of hydrogen bonds and salt bridges in the structures. 10 designs were selected for experimental testing based on these metrics.

The selected designs have statistically better properties than those in the antibody database for five metrics: Amber-calculated binding energies ($p < 0.0001$), Rosetta-calculated binding energies ($p = 0.02$), buried surface area ($p < 0.0001$), salt bridges ($p = 0.0004$), and number of hydrogen bonds ($p < 0.0001$). These results demonstrate that not only is AUBIE able to rapidly design antibodies using limited computational resources, but that it can also design ones that are computationally predicted to be superior to natural antibodies.

Experimental testing and discussion

The selected AUBIE-designed antibodies were evaluated for binding by the Spangler Lab at Johns Hopkins University. The antibodies were expressed using yeast surface display. Full expression and proper folding of the antibodies was confirmed with protein L. Unfortunately, while a control HER2-binding antibody, Trastuzumab, demonstrated low nanomolar affinity under these conditions, the AUBIE designs displayed no detectable binding up to 2 μM. As these experiments were not performed by the author of this dissertation, they are not being described in further detail in this document.

Upon learning that the top AUBIE designs did not bind to HER2, an in-depth analysis of the AUBIE binding interfaces was performed. It was observed that AUBIE designed paratopes consisted of a large number of ARG, LYS and GLN residues. The side chains of these residues have multiple polar atoms, which allows them to form multiple H-bonds. On the other hand, the side chains of these residues have many degrees of freedoms and hence are referred to as highly flexible residues from here onwards. The total number of CDR ARG, LYS and GLN residues per complex in the antibody-antigen database and AUBIE designs were calculated. Structures from the antibody-antigen database and the AUBIE designed set consisted of 4.87 ± 1.72 and 8.63 ± 2.25 highly flexible CDR residues per complex respectively. Table 3.4 lists the number of highly flexible CDR residues in the top AUBIE designs. The average AUBIE design has around four more highly flexible residues than the average complex from the antibody-antigen database. From the top 10 AUBIE designs, six designs had more highly flexible CDR residues than 84% (one standard deviation above mean) of the complexes from the antibody antigen database. ARG, LYS and GLN had increased odds of forming H-bonds and being part of an AUBIE solution since they

have multiple polar atoms. Interactions with these highly flexible side chains are associated with large entropic cost. Antibody paratopes consist of relatively larger number of low entropy amino acids such as SER and TYR, which have lower entropic cost upon binding. Considering the highly beneficial binding energies obtained for the AUBIE designs, forcefields were not able to adequately capture the large entropic cost associated with making interactions with ARG, LYS and GLN side chains. This inability of forcefields to capture entropy costs efficiently has also been reported in literature [8]. In Chapters 4 and 5, I looked more in depth on conformational stability of paratopes and epitopes and made modifications to the AUBIE approach. More detailed computational benchmarking and comparisons with the modified AUBIE design approach is described then.

Table 3.1 Interaction types, paratope and epitope chemical groups involved and applicable paratope residue types for SBRs. Interacting atom shown in bold.

SBR	Interaction type	Paratope	Epitope	Paratope Residue
1	H-bond	N-**H**	**O**=C	All
2	H-bond	N-**H**	H-**O**-C	All
3	H-bond	C=**O**	**H**-N	All
4	H-bond	C=**O**	**H**-O-C	All
5	H-bond	C-O-**H**	**O**=C	SER, THR, TYR
6	H-bond	C-O-**H**	H-**O**-C	SER, THR, TYR
7	H-bond	C-O-**H**	**H**-N	SER, THR, TYR
8	H-bond	C-O-**H**	**H**-O-C	SER, THR, TYR
9	H-bond complex	ARG	**O**=C-**O**	ARG
10	H-bond complex	LYS	**O**=C-**O**	LYS
11	H-bond complex	**O**=C-**O**	ARG	ASP GLU
12	H-bond complex	**O**=C-**O**	LYS	ASP GLU
13	Cation-pi	ARG	Aromatic	ARG
14	Cation-pi	LYS	Aromatic	LYS
15	Cation-pi	Aromatic	ARG	PHE, TRP, TYR
16	Cation-pi	Aromatic	LYS	PHE, TRP, TYR
17	π-π (parallel displaced)	Aromatic	Aromatic	PHE, TRP, TYR
18	π-π (T-shaped)	Aromatic	Aromatic (T horizontal)	PHE, TRP, TYR
19	π-π (T-shaped)	Aromatic	Aromatic (T-vertical)	PHE, TRP
20	Cation-π (parallel displaced)	HIS	Aromatic	HIS
21	Cation-π (T-shaped)	HIS	Aromatic	HIS
22	Cation-π (parallel displaced)	Aromatic	HIS	PHE, TRP, TYR
23	Cation-π (T-shaped)	Aromatic	HIS	PHE, TRP, TYR

Table 3.2 SBR distance constraints

	Distance (Å)	Source
A hydrogen bond between an N-H (donor) and O=C (acceptor)	2.89	[9]
A hydrogen bond between an N-H (donor) and a HO-C (acceptor)	3.00	[9]
A hydrogen bond between an O-H (donor) and a O=C (acceptor	2.79	[9]
A hydrogen bond between an O-H (donor) and a HO-C (acceptor)	2.86	[9]
The distances from C to C in an ARG to ASN/GLN interaction	3.96	[9]
The distance from N to C in a LYS to ASN/GLN interaction	3.18	[9]
The distance from the central C in ARG to the center of an aromatic ring in an ARG-π interaction	3.75	[10]
The distance from the N in LYS to the center of an aromatic ring in a LYS – π interaction	4.20	[10]
The vertical displacements from center of ring to center of ring in a parallel displaced π-π stacking interaction	3.75	[11]
The horizontal displacements from center of ring to center of ring in a parallel displaced π-π stacking interaction	1.50	[12]
The distance from center of ring to center of ring in a T-shaped π-π interaction	5.00	[11]
The distance from the center of histidine ring to the center of benzene ring in a parallel histidine-π interaction	3.50	[10]
The distance from the center of histidine ring to the center of benzene ring in a T-shaped histidine-π interaction	4.25	[10]

Table 3.3 Binding metrics such as AMBER and Rosetta forcefield calculated binding energies, buried solvent accessible surface area (SASA), number of salt bridges and H-bonds and number of highly flexible CDR amino acids in antibody-antigen database, AUBIE designed solution set and best 10 tested AUBIE designs.

Name	AMBER binding energy (kcal/mol)	Rosetta binding energy (kcal/mol)	Buried SASA (Å^2)	Number of salt bridges	Number of H-bonds	Number of ARG, LYS and GLN
Antibody-antigen database	-139.4 ± 43.46	-54.66 ± 16.04	1903.77 ± 489.20	4.19 ± 3.40	9.41 ± 4.02	4.87 ± 1.72
AUBIE database	-170.66 ± 31.83	-40.38 ± 9.90	2153.70 ± 386.35	6.88 ± 2.39	8.22 ± 3.30	8.63 ± 2.25
Top 10 AUBIE designs	-225.32 ± 20.88	-59.46 ± 5.64	2705.05 ± 200.29	7.10 ± 1.91	13.50 ± 1.78	7.40 ± 2.01
P-values	$2.06\ 10^{-8}$	$1.60\ 10^{-2}$	$1.36\ 10^{-8}$	$3.94\ 10^{-4}$	$8.70\ 10^{-6}$	$1.77\ 10^{-3}$
Design_40	-245.81	-65.41	2829.76	8	14	6
Design_41	-240.92	-63.47	3001.11	9	14	5
Design_42	-216.28	-52.10	2582.06	7	13	7
Design_43	-249.09	-67.54	2965.49	6	16	5
Design_47	-202.02	-55.74	2396.38	7	11	9
Design_51	-189.09	-62.39	2863.84	3	16	7
Design_119	-250.12	-57.47	2650.95	10	14	11
Design_139	-225.13	-53.23	2630.45	8	14	9
Design_227	-220.11	-53.72	2623.53	7	11	6
Design_289	-214.57	-63.55	2506.90	6	12	9

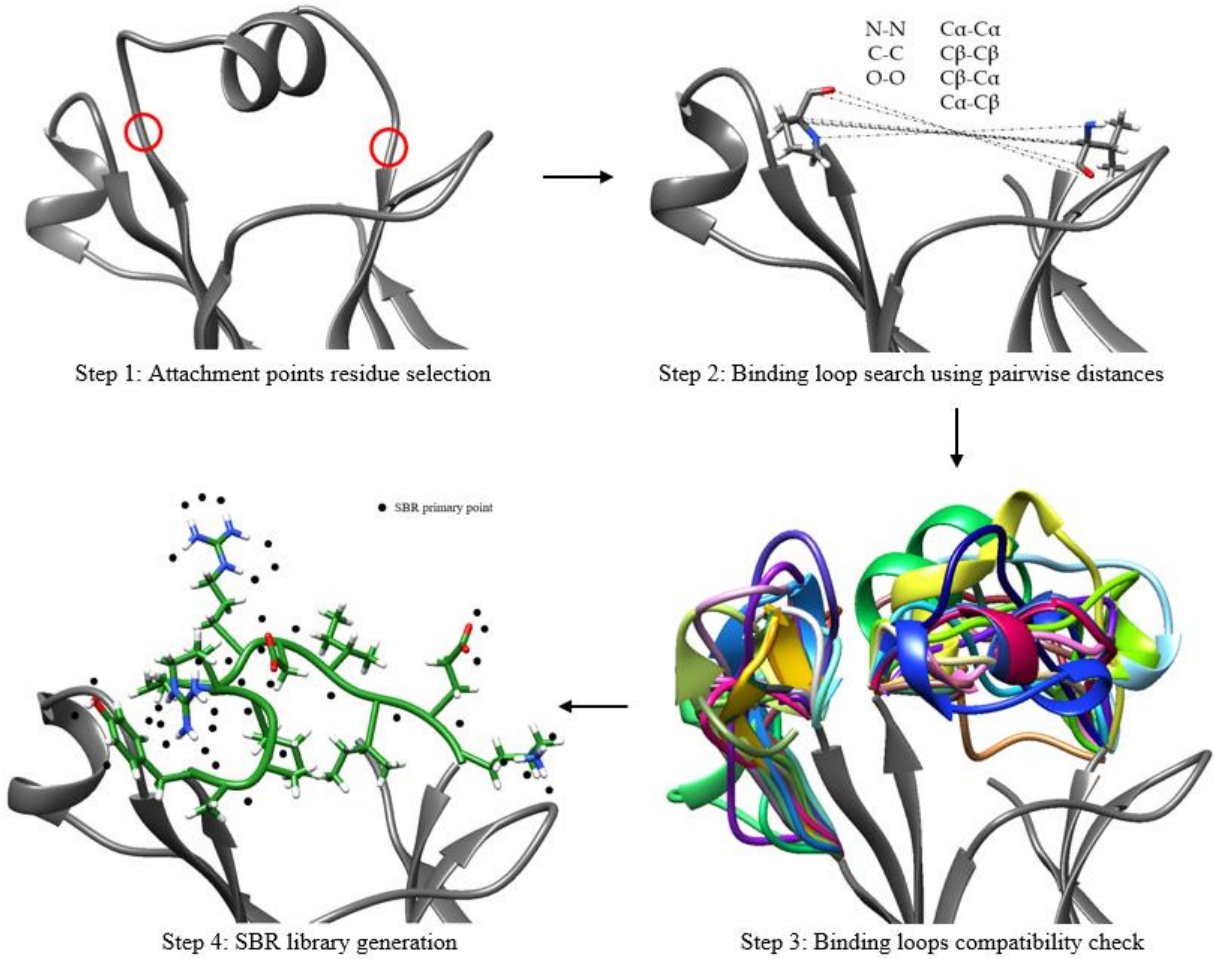

Figure 3.1 Database generation step workflow

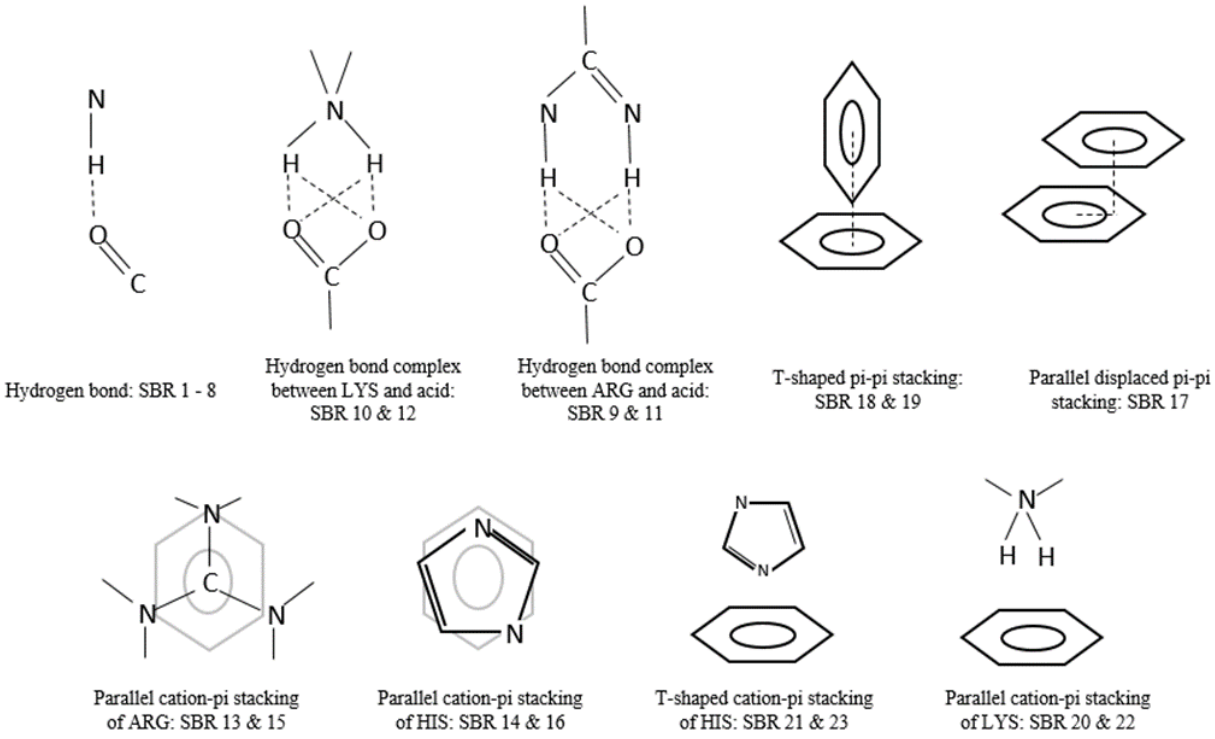

Figure 3.2 Interaction types simulated by SBRs in AUBIE

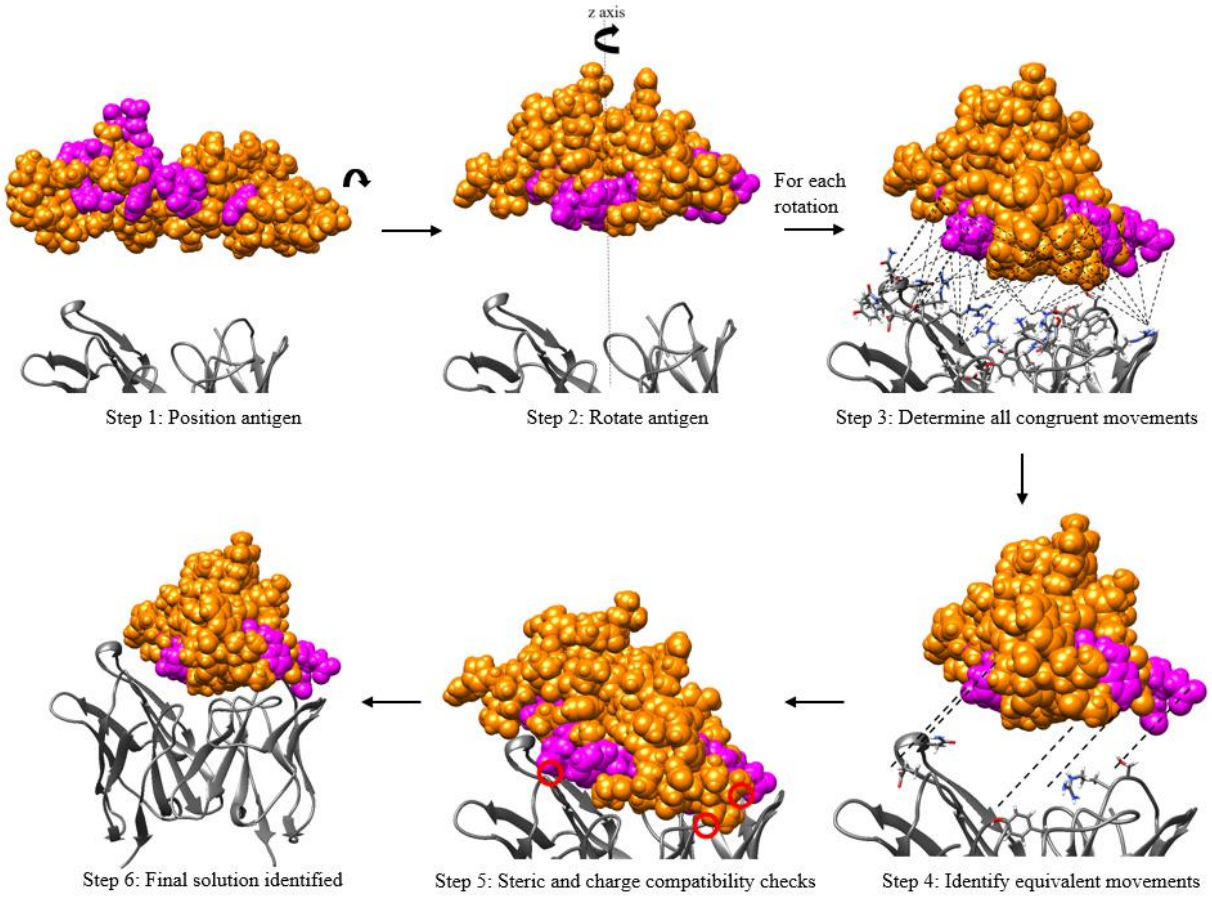

Figure 3.3 AUBIE binding protein design workflow

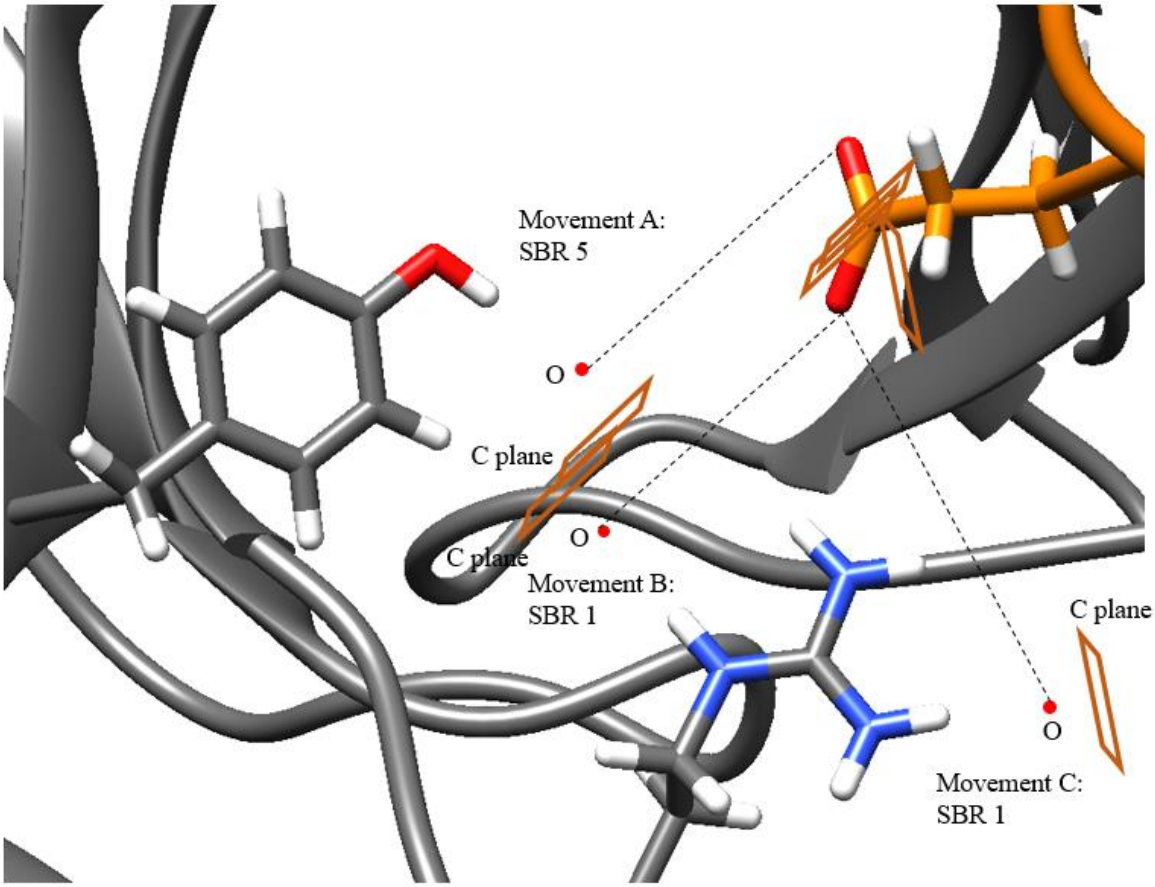

Figure 3.4 Movements of three SBRs between two paratope residues (gray) and one epitope residue (orange) are shown. Movements A and B are congruent while movement C is not congruent with any other movement.

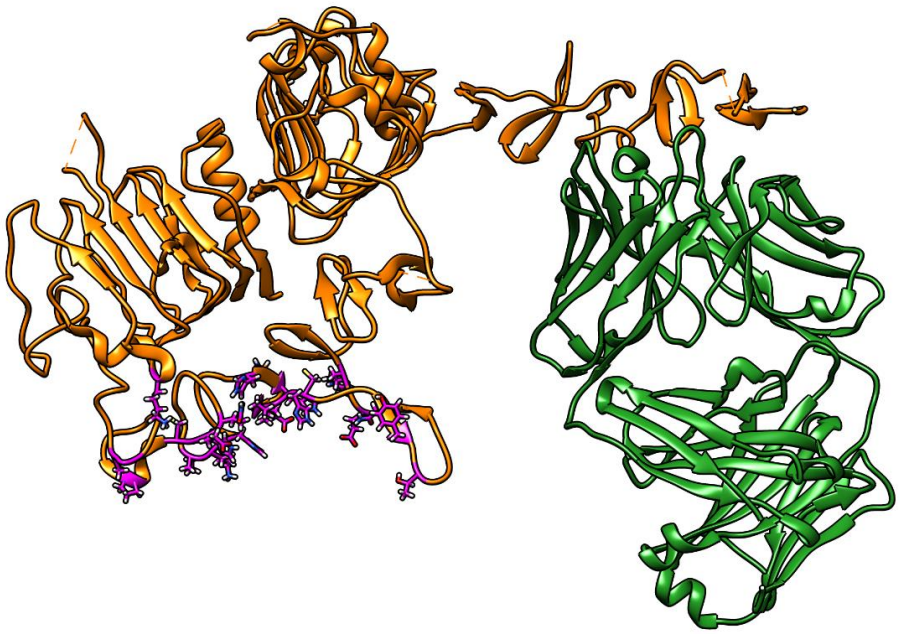

Figure 3.5 HER2 antigen (orange) in complex with Herceptin (green) from PDB 1N8Z. Epitope selected for AUBIE design run shown in magenta.

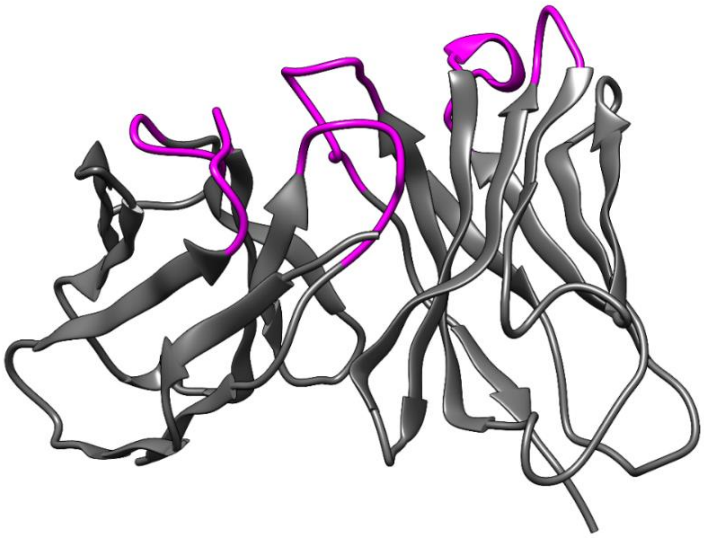

Figure 3.6 Selected antibody scaffold and binding regions for AUBIE database generation. Framework and CDR binding regions shown in gray and magenta respectively. Structure obtained from PDB 4X4X.

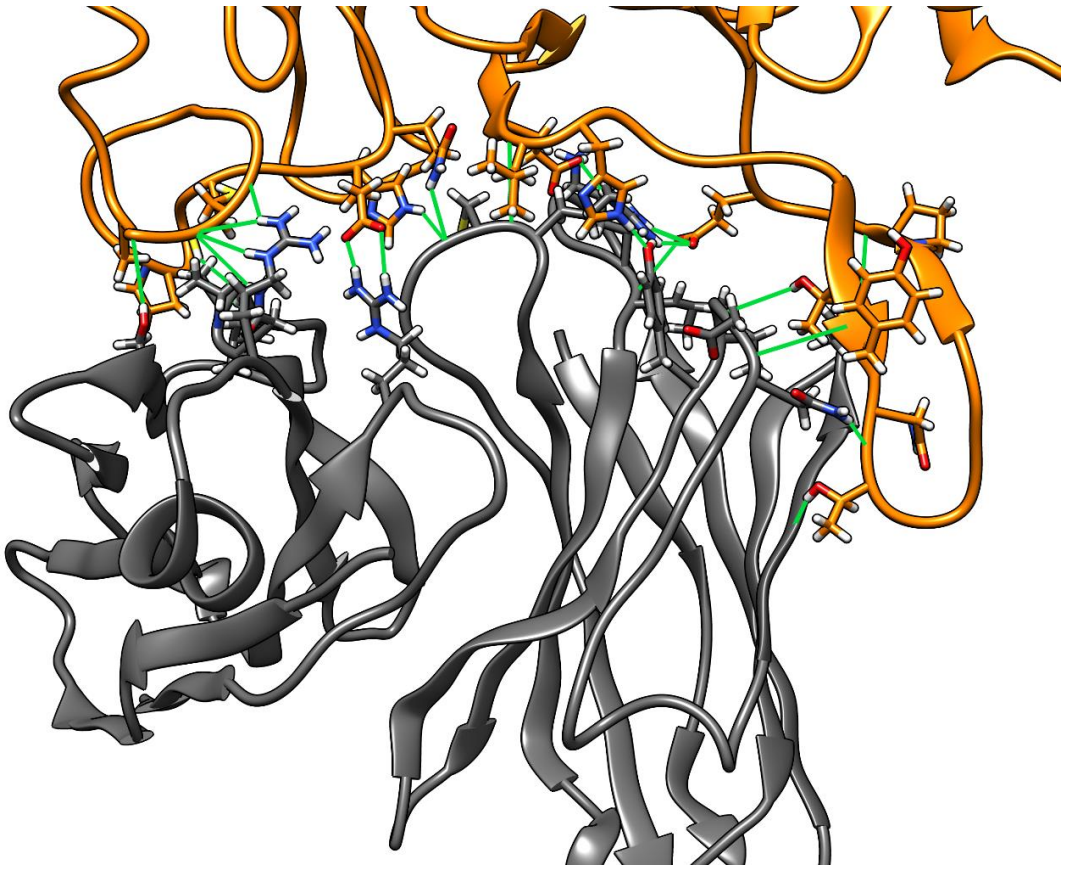

Figure 3.7 Binding interface of AUBIE design 119. AUBIE antibody and HER2 antigen shown in gray and orange respectively. H-bonds shown in green.

References

[1] B. R. Brooks *et al.*, “CHARMM: the biomolecular simulation program,” *J Comput Chem*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[2] G. B. McGaughey, M. Gagné, and A. K. Rappé, “π-Stacking Interactions: ALIVE AND WELL IN PROTEINS*,” *Journal of Biological Chemistry*, vol. 273, no. 25, pp. 15458–15463, 1998, doi: https://doi.org/10.1074/jbc.273.25.15458.

[3] D. Gajria and S. Chandarlapaty, “HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies,” *Expert Rev Anticancer Ther*, vol. 11, no. 2, pp. 263–275, Feb. 2011, doi: 10.1586/era.10.226.

[4] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, “ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB,” *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3696–3713, Aug. 2015, doi: 10.1021/acs.jctc.5b00255.

[5] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, Jun. 2017, doi: 10.1021/acs.jctc.7b00125.

[6] P. B. Stranges and B. Kuhlman, “A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds,” *Protein Sci*, vol. 22, no. 1, pp. 74–82, Jan. 2013, doi: 10.1002/pro.2187.

[7] J. Adolf-Bryfogle *et al.*, “RosettaAntibodyDesign (RAbD): A general framework for computational antibody design,” *PLOS Computational Biology*, vol. 14, no. 4, pp. e1006112-, Apr. 2018, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006112

[8] S. J. Fleishman, S. D. Khare, N. Koga, and D. Baker, “Restricted side chain plasticity in the structures of native proteins and complexes,” *Protein Science*, vol. 20, no. 4, pp. 753–757, Apr. 2011, doi: https://doi.org/10.1002/pro.604.

[9] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” in *Computer Aided Chemical Engineering*, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds. Elsevier, 2018, pp. 2113–2118. doi: https://doi.org/10.1016/B978-0-444-64241-7.50347-5.

[10] K. Kumar, S. M. Woo, T. Siu, W. A. Cortopassi, F. Duarte, and R. S. Paton, “Cation–π interactions in protein–ligand binding: theory and data-mining reveal different roles for lysine and arginine,” *Chemical Science*, vol. 9, no. 10, pp. 2655–2665, 2018, doi: 10.1039/C7SC04905F.

[11] C. Chipot *et al.*, “Benzene Dimer : A Good Model for π - π Interactions in Proteins ? A Comparison between the Benzene and the Toluene Dimers in the Gas Phase and in an Aqueous Solution,” vol. 7863, no. 19, pp. 11217–11224, 1996, doi: 10.1021/ja961379l.

[12] M. O. Sinnokrot, and C. D. Sherrill, “Highly Accurate Coupled Cluster Potential Energy Curves for the Benzene Dimer :,” pp. 10200–10207, 2004, doi: 10.1021/jp0469517.

4. Chapter 4 - Pre-binding conformational stability

In Chapter 3, I learned that while AUBIE can design antibodies computationally predicted to be superior to natural ones, they are ineffective experimentally, possibly due to a heavy reliance on long chain amino acids such as ARG, LYS and GLN. The total number of such side chains in AUBIE designs was significantly higher than the average complex from the antibody-antigen database. Long chain amino acids like ARG, LYS and GLN have many degrees of freedom and hence have large entropic energies if not stabilized by steric or electrostatic interactions. Chapter 2 reported that antibody paratopes have more aromatic residues like TYR and TRP, which offer more conformational stability due to their large size and low degrees of freedom. This Chapter will provide an in-depth analysis of features related to conformational stability of binding interface residues in antibody-antigen and non-antibody protein-protein complexes. This work will be submitted for publication in 2022.

Introduction

Other work has also studied the various aspects of conformational stability in protein binding interfaces. Qiao et al. performed 100 ns long MD simulations of 10 antibody-antigen complexes before and after complex formation [1]. They found that epitope contact residues displayed lower B-factor values than non-epitope antigen surface residues before binding. This observation was true even for epitopes made of non-secondary structural elements such as helices and β-strands. Uversky and Regenmortel reported that both antibody and antigen interfaces consist of some degree of pre-binding flexibility rather than highly rigid or flexible conformations [2]. They recommend a “flexible lock – adjustable key” mode of interaction for antibody-antigen complexes. Fernandez-Quintero et al. analyzed 10 pairs of antibody fragment domains before and

after affinity maturation through metadynamics and MD simulations [3]. They reported that mutations introduced during affinity maturation reduced HCDR3 conformational diversity, hence making the paratope more rigid. Mishra and Mariuzza have also reported similar findings that increased rigidity is part of the modifications brought upon via affinity maturation for increased specificity and affinity [4]. An analysis of 178 enzyme active sites by Bartlett et al. revealed that active site residues have lower B-factor values than other enzyme surface residues [5]. Side chains of large amino acids such as ARG or LYS are tethered via hydrogen bonds with one of their polar atoms while the other(s) react with the substrate. Wang et al. analyzed databases of antibody-antigen and non-antibody protein-protein complexes for interface residue side chain entropy along with other properties [6]. They reported that paratopes select low entropy amino acids and hence their side chain entropies are lower than epitope side chain entropies. Additionally, they learned that side chain entropies of non-antibody receptor and ligand proteins are higher than those of antibody paratope residues. Fleishman et al. computed rotamer probabilities of interface aromatic side chains in designed and native protein complexes that had comparable computational binding energy [7]. They reported that native side chains displayed higher rotamer probabilities when compared to designed side chains. These observations coincide with the ones with AUBIE designs, which also had high binding energies but poor side chain stability.

The commonly used hotspot-based design approach is based on designing binding proteins that target few strong interactions [8], [9]. No database wide quantitative analysis of pre-stabilization in terms of interactions has been done even though literature is available underscoring the significance of paratope pre-binding stability. Current literature has focused on pre-stabilization on a residue basis, rather than on an interaction basis. In this work, databases of protein-antigen, protein-peptide and non-antibody protein-protein databases were analyzed to

identify pre-stabilization related frequencies and features of various interaction types, binding loops and surfaces. Knowledge from this analysis can help develop constraints or rules that can guide antibody or binding protein design.

Methodology

Two sets of databases, one made of antibody-antigen complexes and the other made of non-antibody protein-protein complexes, were primarily analyzed in this work. The antibody-antigen set was obtained by extracting Chapter 1 database complexes that had an antigen size larger than 15 residues and resolutions better than 3 Å. A total of 268 complexes were obtained, out of which 57 were nanobody-antigen complexes. Nanobodies are antibodies that consist of only a single heavy chain variable domain instead of the heavy and light chain Fab complex [10]. The non-antibody protein-protein database was generated by collecting complexes from the PDB4Bind database [11] that had binding affinity strong than 100 nM, resolutions better than 3 Å and were made of single chain binding proteins. Furthermore, complexes with proteins larger than 400 residues or sequence similarity 90% with any other database structure were deleted. Complexes with large proteins were not used for computational runtimes purposes, while those with high similarity were removed to ensure non-redundancy. A total of 231 complexes were obtained. Table 4.1 lists the PDB ids of all the complexes in the two databases, as well as those of a third database composed of 75 antibody-peptide (i.e., complexes < 15 amino acids) used in a subsequent study. All the complexes in the three databases were initially run through vacuum energy minimization using CHARMM36 [12] with fixed backbone atoms, followed by energy minimizations using REF2015 [13]. This minimization protocol was used for all designed protein complexes in this

work hereafter and was used for database complexes for consistency purposes. Reasons for the use of this routine will be described later in Chapter 5.

The protein-protein interactions analyzed in this work include H-bonds, salt bridges and hydrophobic interactions. Interactions were identified using structural definitions rather than with force field dependent energy calculations since force fields are known to have biases towards certain interaction types or structural elements [14]–[16]. H-bonds were determined by identifying polar hydrogen atoms and acceptor atoms that were less than 2.5 Å apart, possessed donor-hydrogen-acceptor angles less than 120° [17], and had donor-acceptor-acceptor antecedent angles less than 90°. Salt bridges were determined by identifying negatively charged oxygen and nitrogen atoms from ARG, LYS side chains and N-terminals that were less than 4 Å apart [18]. All the amino acid rotamers for this work were obtained from the Dunbrack rotamer library [19]. Buried SASA was calculated using principles from work by Shrake and Rupley [20]. For each atom, solvent exposed spheres were generated and the atom specific SASA was calculated by determining what regions of these spheres were not buried by other solvent spheres. The radius of a solvent sphere was the sum of the atom's vdW radius and the probe radius of 1.4 Å. Hydrophobic interactions between residues were estimated by determining the loss of SASA of nonpolar atoms upon binding. The buried nonpolar SASA was calculated by considering the SASA loss of only C and S atoms.

The nonpolar SASA lost between all pairs of residue parts in the antibody-antigen database were collected. Nonpolar contacts were defined as significant if the nonpolar SASA lost was more than one standard deviation larger than the mean value. The 84.1st percentile value from the entire nonpolar contact set was 24.38 Å^2. Hence, two residue elements were defined to be making a hydrophobic interaction/nonpolar contact if the buried nonpolar SASA between them was more

than 24.38 $Å^2$. To ensure that the threshold was not excluding residues that made significant hydrophobic interactions, nonpolar contacts made by known significant nonpolar residues were analyzed. The SKEMPI 2.0 database was searched for ALA mutations made to nonpolar amino acids in antibody-antigen complexes that resulted in loss in binding affinity by at least an order of magnitude [21]. Through visual analysis, mutations vital for structural stability of the antibody rather than to binding were filtered out. Table 4.2 lists the mutations, initial and final affinities and largest nonpolar contact SASA. The limit of 24.38 $Å^2$ misses 5 out of 35 (~15%) of the mutations.

In order to develop and analyze features related to conformational flexibility before complex formation, residue elements were classified as either stable or unstable. A residue was broken down into two component elements: backbone and side chain. Stability definitions were based on probabilities of finding residue elements in their native conformational state. For backbone elements, it was assumed that the covalent bonds of the backbone provide inherent stability so all backbone elements were classified as stable. For side chain elements, rotamers of the native amino acid were used to represent the alternative states that side chain can take. A side chain element was defined to be stable if it met either of two conditions: 1) it made two salt bridge interactions or 2) more than 80% of all rotamers had at least one heavy atom (i.e., non-hydrogen) vdW clash with surrounding atoms. Two heavy atoms were defined to have a vdW clash if they were closer than 90% of the sum of their vdW radii [12]. Salt bridge interactions are the most attractive electrostatic interaction and are known to be highly stable [22]. Hence charged side chains making two or more salt bridges were classified as stable side chains. The 80% limit is not applied to SER and THR since they can have only three rotamer states and side chains with no more than one non-clashing rotamer were classified as stable. Side chains of ALA, PRO and GLY were classified as stable because they do not have alternative rotamer states.

Results

After classifying all interface residues as either stable or unstable and identifying the H-bonds, salt bridges and nonpolar contacts these residues were making, the data was analyzed to quantify several binding interface related features. These features included stability-related binding frequencies at different structural levels: complex, CDR, antibody position and amino acid type. A metric to quantify paratope and epitope conformational stability before binding was also defined and analyzed.

First, the frequencies of different interaction types based on the pre-binding stability of the two interacting residue parts were analyzed. Figure 4.1 plots the different percentage frequencies of H-bonds, salt bridges and nonpolar contacts between stable-stable, stable-unstable and unstable-unstable residue parts. Figure 4.2 does the same for interactions from the protein-protein database. Nonpolar contacts were found to be the most frequent interaction type in antibody-antigen complexes followed by H-bonds and salt bridges. This result was expected since all residues can make nonpolar contacts while only polar and charged amino acids can make H-bonds and salt bridges respectively. Similar trends were seen in the protein-protein database but with more frequencies of nonpolar contacts. Interactions made by one stable and one unstable residue parts were more than twice as frequent as other interaction types in both the databases. Stable-stable salt bridges were rare and unstable-stable and unstable-unstable salt bridges were approximately equally frequent in both datasets. Similar frequencies of stable-stable and stable-unstable H-bonds were observed in both datasets. An explanation for this is that stable backbone parts form a large fraction of these H-bonds. Abundance of backbone atoms in atom contact pairs in binding interfaces has been reported in literature [23].

Upon learning that a majority of interactions in binding interfaces consisted of at least one stable residue part, it was analyzed if these higher frequencies of stable parts is an evolutionarily selected feature or consistent with random chance given the frequency of stable residues in binding surfaces. To answer this question, the probabilities of the occurrence of the four types of interactions were calculated and compared to the actual frequencies. The probabilities were obtained by calculating what fractions of the epitope and paratope SASA were composed of stable and unstable parts. After the probabilities and actual frequencies of the four interaction types were calculated for each interface, averages and standard deviations were computed. The results are shown in Figures 4.3 and 4.4. On average, stable-stable interactions were made more frequently than expected through random chance. The opposite was true for unstable-unstable interactions, as they occurred less frequently than expected. These patterns were true for protein-protein binding interfaces too. Antibody-antigen interfaces disfavor interactions with unstable antigen side chains while favoring interactions with stable antigen residue parts. Antibody binding interfaces select interaction pairs that consist of unstable low entropy side chains from the antibody rather than higher entropy side chains from the antigens since antibody paratopes consist of more lower entropy side chains than their epitopes. Further quantification of entropy loss from paratopes and epitopes is required. In protein-protein interfaces, both the binding surfaces do not have any distinguishing features unlike antibodies. Thus, they were not classified as epitopes or paratopes and all stable-unstable interactions were lumped into one group as shown in Figure 4.4. It was observed that the frequencies of stable-unstable interactions in protein-protein interfaces were similar to those expected from random chance. Hence, the selection of stability was only reflected in stable-stable and unstable-unstable interactions.

After the interaction-level analysis, binding patterns on a residue level were examined. The different types of interactions made by individual unstable antigen amino acids were studied since epitope surfaces have more unstable side chains that can be targeted for antibody design. The results are shown in Figures 4.5, 4.6 and 4.7. For charged amino acids, more than 50% of unstable side chains make single or multiple interactions with stable residue parts. A considerable fraction of the remaining side chains made multiple strong salt bridge interactions with other unstable side chains to compensate for the loss in entropy of the two binding side chains. The long chain ARG and LYS side chains make more nonpolar contacts than shorter chained ASP and GLU side chains, which depend more on salt bridge formation. For amino acids like ASP, GLU, ASN and GLN, that have multiple polar groups, approximately half of all unstable side chains make H-bond(s) with stable residue parts. ASP and GLU form more multiple H-bonds since they can form salt bridges too. Aromatic side chains TYR, HIS and TRP depend more on nonpolar contacts with stable residue parts. Interestingly, SER and THR also form predominantly more nonpolar contacts than H-bonds with stable residue parts. The most abundant interaction type for nonpolar amino acids was multiple nonpolar contacts. For all the amino acids, a large fraction of these multiple nonpolar contacts consisted of either a single or all stable residue parts. This fraction was lower for PRO due to its smaller side chain and for HIS since it has two Nitrogen atoms in side chain ring.

Next, paratope and epitope conformational stability before binding were quantified and this property was analyzed between databases and the AUBIE designed interfaces from Chapter 3. To quantify conformational flexibility loss upon binding, the total non-clashing rotamers lost upon binding were counted and divided by the total SASA loss upon binding. This metric is termed Rotamers Lost per Solvent Accessible Surface Area lost (RL-SASA). The RL-SASAs, averages,

and standard deviations were calculated for all epitopes and paratopes from the databases and AUBIE solutions. The results are shown in Figure 4.8.

The RL-SASAs for paratopes were significantly lower than those from the epitopes from the antibody-antigen complexes. This result was expected since it is known that antibody paratopes consist primarily of low entropy amino acids such as SER and TYR. Similar results have been obtained in other work [6]. Meanwhile, epitope surfaces consist of relatively higher number of high entropy amino acids. Similarly, protein-protein binding surfaces also consist of several unstable high entropy side chains and hence have larger RL-SASA values than antibody paratopes. One possible reason for the lack of high entropy side chains in antibody paratopes could be the minor instabilities of CDR loops. While this work has considered all backbone atoms to be stable, backbone atoms of secondary structures like helices and β-strands are less flexible than backbone atoms of regular loops such as CDRs. Epitope surfaces can consist of the stable secondary structural elements while CDRs cannot, and hence may compensate for this weaker backbone stability by selecting low entropy amino acids. Peptides do not consist of a fixed backbone conformation and are more flexible than regular protein structures. Whether antibody paratopes for peptide antigens compensate for this extra instability through the rigidification of their paratopes was investigated. Upon comparing the RL-SASAs of paratopes from antibody-antigen and antibody-peptide complexes, it was determined that the average RL-SASA from peptide binding paratopes is statistically lower than the average from antigen-binding paratopes (p-value from one-sided t-test assuming unequal variances was 0.0004). Of note, the paratope RL-SASA values from the top AUBIE designs were significantly larger than those from the antibody-antigen complexes. This was most probably caused by the abundance of high entropy ARG and LYS CDR

residues. Hence it is likely that the AUBIE designs did not bind since their paratopes were highly unstable and not adequately stabilized during complex formation.

A common design strategy for the de novo design of antibodies for target epitopes is to dock or design a CDR loop towards a hotspot residue(s) and to build the remaining antibody structure based on the position of the docked CDR loop. The docking or designing of CDR loops becomes easier if it makes multiple interactions with the target epitope residue since it reduces the degrees of freedom in the problem. The number of unstable epitope side chains that make interactions with multiple residues from the same CDR was analyzed. Figure 4.9 displays the percentage frequencies of epitope side chains that make different combinations of electrostatic and nonpolar interactions with residues from the same CDR. As expected, for large amino acids such as ARG, LYS, PHE, TRP, LEU and MET, more than half of unstable side chains make interactions with two or more residues from the same CDR. ARG, LYS, GLN and GLU are able to make more electrostatic and nonpolar interactions than the other amino acids because have both hydrophobic and polar parts. PHE side chains, while not being the largest amino acid, seem to be the easiest to target since they had the largest fraction of multiple residue binding interactions. Thus docking CDR loops to have two or more residues make interactions with a single unstable large side chain is a viable first step towards antibody design.

The final feature analyzed in this work was the likelihood of specific antibody positions to host interaction forming stable side chains of individual amino acids. This was done by calculating the percentage frequencies of interactions at each antibody position. Moreover, this was done for each amino acid type. All antibody residue positions were numbered using the IMGT numbering scheme. The results are shown in Figures 4.10, 4.11 and 4.12. CDR positions that were most likely to host stable side chains included the end residues of the CDR1 loops, positions 36 to 38, and the

beginning and end residues of HCDR2 loop, positions 55 to 57 and 64 to 66. Initial CDR3 residue positions 107-109 were also observed to be common hosts of stable side chains of several amino acids such as VAL, PRO, GLN and GLU. Amino acid side chains at CDR2 end regions and CDR3 stem regions end up making several interactions with the epitope and have restricted flexibility since these regions form the central binding region of all antibodies. Unlike other aromatic amino acids, which were most common in the previously mentioned CDR3 and CDR2 positions, stable HIS side chains were common at positions L31, H40 and H113. Similarly, unlike other amino acids, negatively charged ASP and GLU side chains were commonly found at CDR3 end position 116 of heavy and light chains respectively. Some framework positions were more likely to host stable side chains of certain amino acids than other CDR positions. For example, stable ARG and LYS side chains were common at framework position L80. Stable ARG side chains were most prevalent at H106. PRO and backbone interactions were unique since they were most frequent only in CDR3 positions. CDR3 loops are relatively longer than other CDR loops and hence are less likely to have surrounding residues to help stabilize its side chains. Therefore, interactions made by low flexibility elements like PRO and backbone parts were frequent in CDR3 positions.

Discussion

In this chapter, databases of antibody-antigen and protein-protein complexes were analyzed to learn more about the prestabilization features of binding interface residues. Initially, residue side chains were classified as either stable or unstable and the various H-bonds, salt bridges and nonpolar contacts they made were analyzed. The main learnings from the various features studied can be summarized in the following points:

1. Most interactions in binding interfaces consist of at least one stable residue part.

2. Stability in interface is selected rather than being a natural result of pre-binding stabilities.
3. Unstable charged and nonpolar side chains make salt bridges and multiple nonpolar contacts respectively, most with stable residue parts to compensate for entropy loss.
4. Antibody paratopes are more rigid than their epitopes. Peptide-binding paratopes are more rigid than antigen-binding paratopes.
5. More than half of all unstable large chain epitope side chains make interactions with two or more residues from a single CDR.
6. Specific antibody residue positions are more likely to host stable side chains than other residue position

Non-antibody protein-protein interfaces consisted of more unstable-unstable interactions than antibody-antigen interfaces. One likely cause for this is that protein-protein binding interfaces have marginally different interaction mechanisms due to the amino acid compositions of the binding surfaces. Antibody paratopes consist of more aromatic residues while other binding surfaces consist of more nonpolar residues such as ILE, LEU, VAL, PRO, and MET [6].

The results from this Chapter can be used to guide binding protein design algorithms. One of the benefits of computational antibody design is targeted epitope binding, and epitope selection is a crucial step in this process. The selected epitope should preferably minimize or reduce the energetic cost of binding. One approach would be to use the learning from this work and select epitopes that are conformationally stable. Since the entropic cost of binding to such an epitope would be low, it gives the antibody designer more leeway in using more flexible CDR amino acids. Force fields are widely used in protein design work, though their limitations in assessing computationally designed binding interfaces are known [7]. One of the features that they fail to accurately capture is entropic energy [24]. Other metrics are often used along with binding energy,

such as shape complementarity or buried surface area, to assist in the analysis of binding interfaces. Results from this work can be further built on to develop novel metrics such as RL-SASA can be used to further examine binding interfaces. Such metrics can also be used in other steps of binding protein design such as docking or affinity maturation.

Table 4.1 PDB file IDs of complexes in the antibody-antigen database, non-antibody protein-protein database and antibody-peptide database.

Antibody-antigen database PDB IDs
1A14, 1A2Y, 1AR1, 1BJ1, 1BQL, 1C08, 1CZ8, 1DEE, 1DQJ, 1EZV, 1FNS, 1FSK, 1G9M 1IQD, 1JHL, 1JPS, 1JRH, 1LK3, 1NCA, 1NDM, 1NSN, 1OB1, 1ORS, 1OSP, 1OTS, 1P2C, 1QFU, 1RJL, 1TPX, 1TZH, 1UJ3, 1V7M, 1ZTX, 2ADF, 2B0S, 2BDN, 2CMR, 2DD8, 2EIZ, 2H9G, 2J4W, 2J88, 2OSL, 2OZ4, 2Q8A, 2QQK, 2QQN, 2R0L, 2R56, 2VXQ, 2VXS, 2VXT 2WUB, 2XQB, 2XRA, 2XWT, 2YPV, 2ZCH, 3BDY, 3BKY, 3BN9, 3D85, 3DVG, 3EFD, 3FFD 3G04, 3G5Y, 3GBM, 3GBN, 3GI8, 3H0T, 3H42, 3K2U, 3KR3, 3L5W, 3L5Y, 3LD8, 3LHP 3LZF, 3MA9, 3MAC, 3MLW, 3NFP, 3NH7, 3NPS, 3PGF, 3R1G, 3RKD, 3SKJ, 3SOB 3SQO 3TJE, 3U30, 3U9P, 3UJJ, 3WIH, 3WKM, 3X3F, 3ZKM, 4AG4, 4AL8, 4BZ2, 4CAD, 4CMH 4D3C, 4D9Q, 4DGI, 4DKF, 4DTG, 4EDW, 4EDX, 4F15, 4FFV, 4FFY, 4FQJ, 4GMS, 4HT1, 4I77, 4I9W, 4IJ3, 4J6R, 4JB9, 4JR9, 4K2U, 4K3J, 4K94, 4KI5, 4KRP, 4KUC, 4L5F 4LEO, 4LIQ, 4LMQ, 4LQF, 4LSR, 4LSU, 4LU5, 4LVH, 4M48, 4M62, 4M8Q, 4N9G 4NNP 4NP4, 4O58, 4O9H, 4OGX, 4OKV, 4OLU, 4PLJ, 4QCI, 4QHU, 4QNP, 4QWW, 4R8W 4RDQ 4RRP, 4RWY, 4S1Q, 4TSB, 4U0R, 4UU9, 4UV7, 4XAK, 4XNY, 4XVT, 4XZU, 4YDK, 4YDL 4YE4, 4YK4, 4YUE, 4ZFF, 4ZFO, 4ZPT, 5A3I, 5ANM, 5BO1, 5BV7, 5C7X, 5CZX, 5D70 5D71, 5D72, 5D93, 5D96, 5DFV, 5DO2, 5DUR, 5E94, 5EPM, 5F96, 5F9O, 5F9W, 5FB8 5IES, 5IKC, 5K59, 5KN5, 5KVD, 5KVE, 5KVF, 5KVG, 1DZB, 2YBR, 3UX9, 3UYP, 3UZQ 4XGZ, 4YJZ, 5DFW, 1BZQ, 1JTO, 1KXQ, 1KXT, 1KXV, 1MVF, 1OP9, 1RI8, 1RJC, 1SQ2 1ZV5, 1ZVH, 2BSE, 2I26, 2P4A, 2VYR, 2XT1, 3CFI, 3EZJ, 3K3Q, 3K74, 3K7U, 3K80 3P9W, 3QSK, 3STB, 3V0A, 3ZKQ, 4AQ1, 4DK3, 4EIG, 4EIZ, 4FHB, 4GRW, 4HEM, 4HEP 4I0C, 4IOS, 4KML, 4LDE, 4LGP, 4LHJ, 4NBX, 4NBZ, 4NC2, 4ORZ, 4P2C, 4PGJ, 4QO1 4WEM, 4WEN, 4WEU, 5C2U, 5C3L, 5E0Q, 5FV1, 5HVF
Non-antibody protein-protein database PDB IDs
1GPW, 6SAK, 6F0G, 2CJS, 5EOA, 1TM1, 1TM7, 3TNF, 1TAW, 3ZRZ, 3E1Z, 3GMW, 1EZU 1G0V, 2GWW, 3FPU, 1D6R, 6JB8, 5BNQ, 3F4Y, 6NDZ, 6DWH, 5J7C, 4QLP, 4HFK, 4ML7, 1L2W, 6J7W, 1PXV, 4ZGM, 2J0T, 5AYS, 3C5T, 4GI3, 1P27, 2GOX, 1DP5, 2NYZ, 1EER, 1Y6K, 2VOH, 3DGC, 3K1R, 4C7N, 2QJA, 2VOI, 3N4I, 3LMS, 3WDG, 1ZJD, 6OG4,

6NE2 6DWF, 1FLT, 4JO6, 3FJU, 4W6Y, 2D10, 2OUL, 3QT2, 5AYR, 2HWN, 3K9M, 3KJ0, 3KJ1, 4AYI, 3ONW, 3EBA, 2QJ9, 4AFQ, 4EQA, 6FTO, 3VPJ, 4GLV, 1DPJ, 1XX9, 2BO9, 5JDS, 6F0H, 6E3I, 1TE1, 5IMM, 1EWY, 4JW2, 5YIP, 1LQS, 4UDM, 3BX1, 4XWJ, 6FU9, 5UUL 4ZW2, 1OPH, 1XG2, 5N48, 4W6X, 2UUY, 4GH7, 4ILW, 1XU1, 3P92, 1FLE, 4HX3, 1YVB, 3ZU7, 5MTM, 5N47, 3WN7, 5XSQ, 3C9A, 2J8X, 2VAY, 5VZ4, 2V9T, 6JJW, 3EGG, 4MP0 1TA3, 2HRK, 4C2B, 4L0P, 5XCO, 5E7F, 6FE4, 4K5A, 1SV0, 3UL4, 5DJU, 1TDQ, 3RBB 1VET, 2Z7F, 2G81, 4JE4, 5NVM, 6BX5, 6E3J, 1KAC, 1SYQ, 2Q0O, 2OOR, 2Z58, 3HG0 3P9W, 4ZQU, 5DJT, 2NZ1, 6JHW, 2J12, 2YGG, 3FXD, 3HUG, 3H8K, 4G6V, 4IU3, 3BK3 4EKD, 5F5S, 1FC2, 1KGY, 5K22, 5UA4, 5Y4R, 6FUB, 1YCS, 5TZP, 6OBN, 6H46, 4BD9 4LYL, 4DH2, 4N7V, 5M72, 2YQ7, 5DC4, 3BRT, 1RKC, 1T01, 5YWR, 1GUA, 1I5K, 6J9H 3BRV, 3IM4, 6FHP, 1NF3, 2JBY, 4BRW, 3T04, 2A78, 4AOR, 2HEV, 2F4M, 6D4P, 3QBR 5B64, 5J4A, 1PVH, 2WWX, 3H8D, 5FW5, 5MR5, 2UYZ, 3QHT, 4YN0, 3IXS, 4KFZ, 3BH6 1MCV, 1AZZ, 1JTD, 1MXE, 1PPE, 1R0R, 1UGH, 2NQD, 2QC1, 2SIC, 2SNI, 2VLN, 2XTT 3FP6, 3G7Z, 3IT8, 3ME2, 3QHY, 3SGB, 3VV2, 4K5B, 4Q5U, 5O0W
Antibody-peptide database PDB IDs
1KCR, 3GJG, 4J8R, 4JO3, 1I8I, 2GSI, 5FGB, 2ZPK, 2OR9, 3QG6, 4HS6, 2H1P, 1P4B 1SM3, 2HKF, 1ACY, 3D0L, 2QHR, 3E8U, 5CIN, 4TUL, 1UWX, 3GO1, 3HR5, 2QSC, 5EOR 1BOG, 1CU4, 1E4W, 1KCS, 1TJG, 1Q1J, 4WHY, 5DLM, 3UO1, 4JZN, 1FPT, 5FGC, 3FN0 5A2I, 1SVZ, 4N8C, 4JO1, 3MLT, 1KTR, 5AUM, 4DGV, 4TUJ, 1KC5, 4LKX, 2Y36, 4HPY, 2V17, 4JZO, 3MLX, 5I8C, 4HS8, 4ZTO, 4RAV, 4G6A, 1NAK, 4O51, 4GAJ, 2HH0, 1XGY 5EOC, 4WHT, 4Q0X, 4HZL, 4XVJ, 3GHE, 3MLR, 2A6D, 5EOQ, 4Z0X

Table 4.2 Initial and final experimental binding affinities of ALA mutations to nonpolar residues. The largest buried nonpolar SASAs lost due to the mutations are also reported.

PDB	Mutation	Initial Kd (nM)	Final Kd (nM)	Largest nonpolar contact ($Å^2$)
1DQJ	YC20A	2.78E-09	7.14E-07	20.21
1JRH	HH100bA	1.20E-08	2.11E-07	20.22
3SE8	FH100dA	6.30E-08	3.71E-07	20.51
3SE8	FL97A	6.30E-08	2.51E-07	23.23
1VFB	YA32A	2.00E-08	1.92E-07	23.84
1DQJ	YA50A	2.78E-09	2.56E-07	26.56
1VFB	YB101A	2.00E-08	No binding	28.07
3NGB	WH50A	2.20E-08	1.88E-07	29.86
3NGB	YL91A	2.20E-08	6.33E-07	30.19
3SE9	VH56A	2.50E-08	1.28E-07	31.08
3SE8	YH59A	6.30E-08	2.61E-07	31.38
3NGB	VH57A	2.20E-08	2.10E-07	31.39
1AHW	YC156A	3.40E-09	>2.1E-06	34.10
1VFB	IC124A	1.25E-08	1.00E-07	38.32
3SE9	LL91A	2.50E-08	1.24E-07	38.92
2NYY	FA953A	1.25E-10	1.24E-07	43.46
1JRH	WH53A	1.20E-08	7.16E-07	44.95
1JRH	YH32A	1.20E-08	1.35E-07	45.25
1JRH	VI51A	1.44E-08	3.46E-07	45.25
1NMB	YH99A	4.55E-08	1.69E-06	45.58
2NYY	HA1064A	1.25E-10	3.17E-05	47.99
3BN9	FB97A	1.23E-11	>1E-06	49.20
1JRH	YI49A	1.44E-08	4.48E-06	51.00
3HFM	YH53A	2.38E-09	5.26E-07	52.22
3NGB	VH73A	2.20E-08	1.34E-07	53.10
1DQJ	YB33A	2.78E-09	3.13E-05	53.11
1VFB	WA92A	2.00E-08	2.00E-06	54.62
1JRH	WL92A	1.20E-08	1.40E-06	56.13
4ZS6	WA535A	3.12E-07	No binding	59.15
1JRH	WI82A	1.44E-08	3.01E-05	60.07
3SE8	WH54A	6.30E-08	2.23E-07	63.67
3G6D	FH103A	1.84E-11	No binding	64.87
1DQJ	WB98A	2.78E-09	1.15E-05	76.93

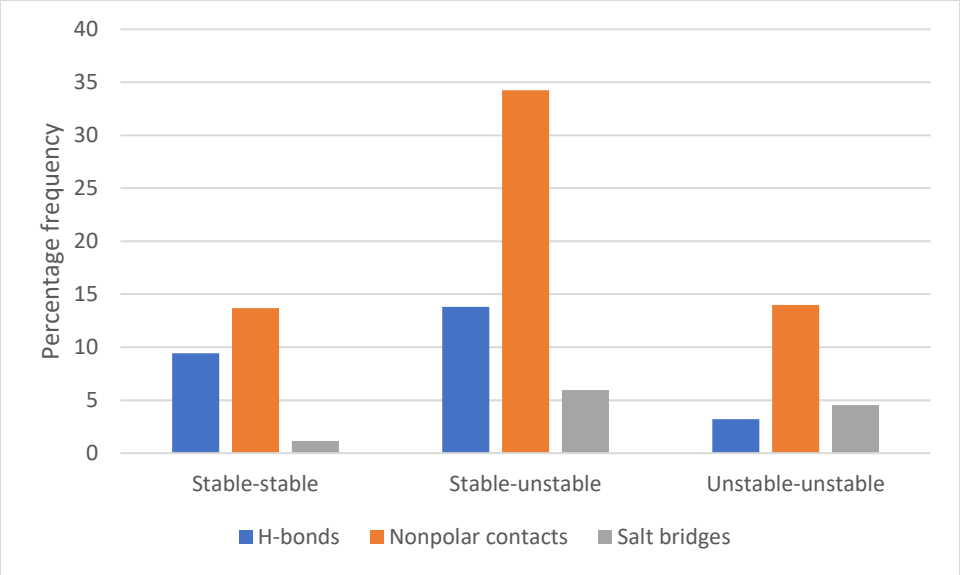

Figure 4.1 Percentage frequencies of H-bonds, salt bridges and nonpolar contacts made by two stable residue parts, one stable and one unstable residue parts and two unstable residue parts in the antibody-antigen database.

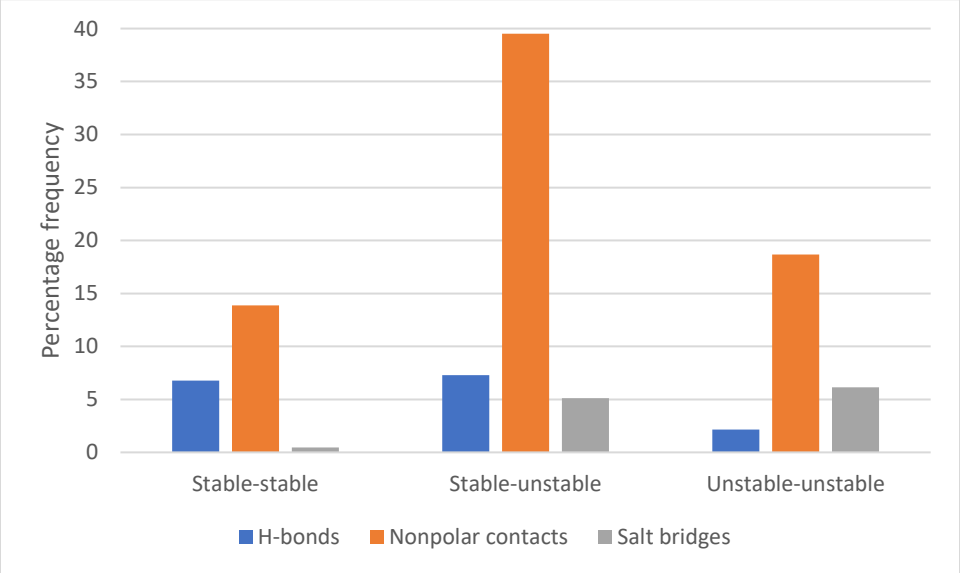

Figure 4.2 Percentage frequencies of H-bonds, salt bridges and nonpolar contacts made by two stable residue parts, one stable and one unstable residue parts and two unstable residue parts in the antibody-antigen database.

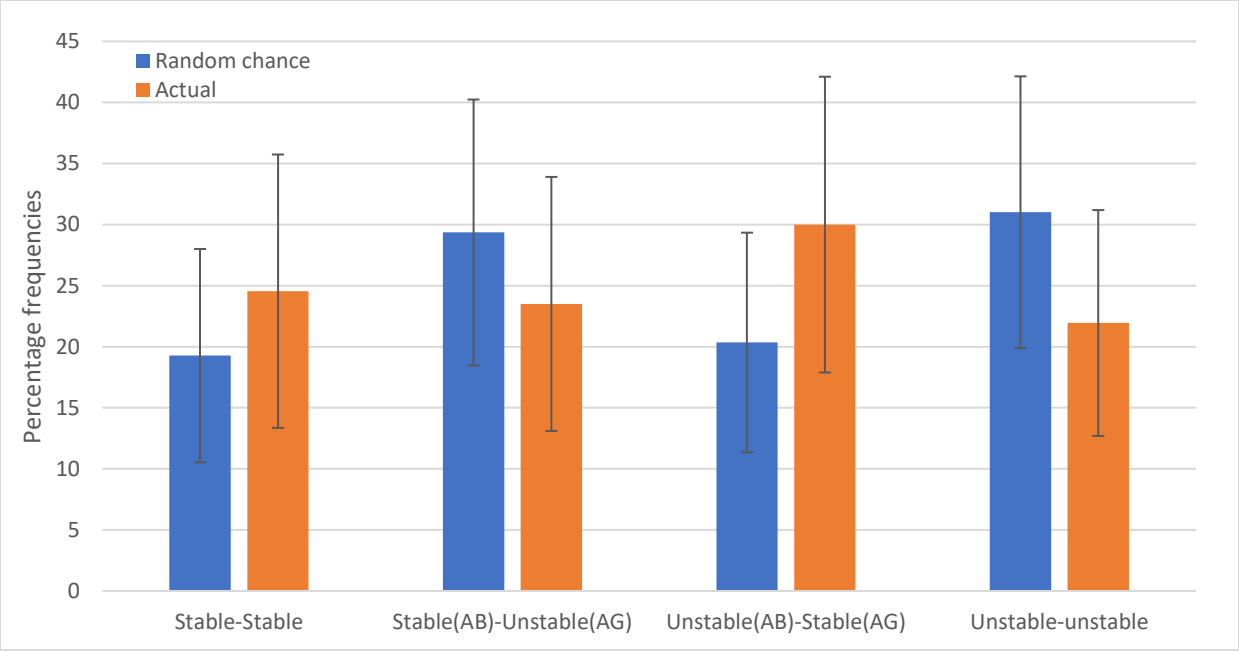

Figure 4.3 Average and standard deviations of percentage frequencies of different interaction types predicted by random chance and actual calculations in the antibody-antigen database.

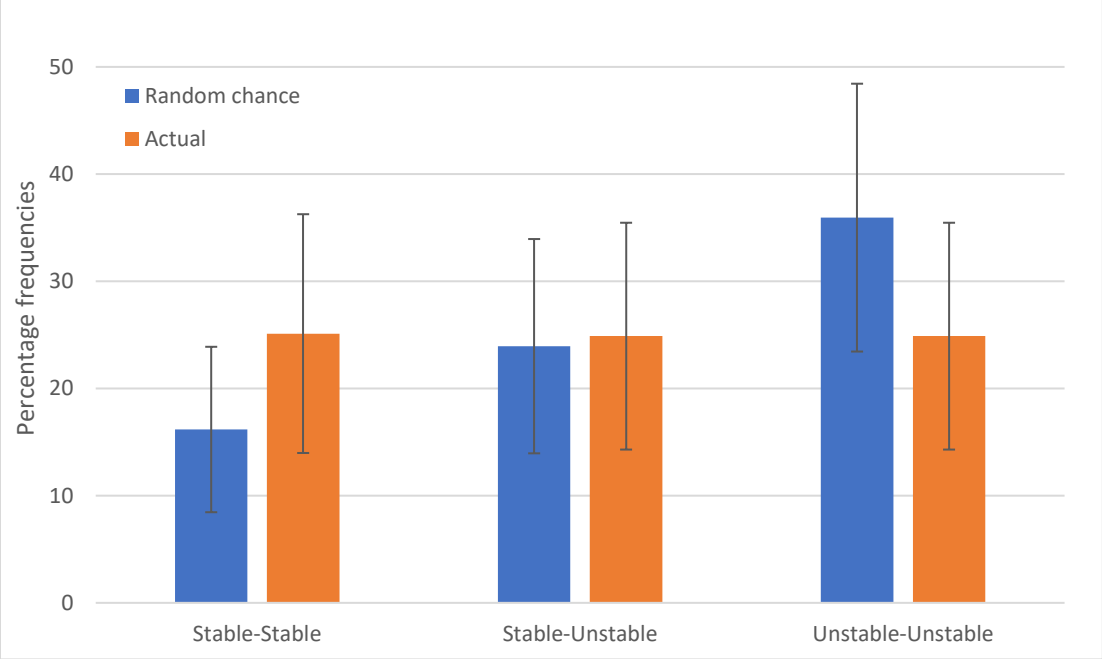

Figure 4.4 Average and standard deviations of percentage frequencies of different interaction types predicted by random chance and actual calculations in the protein-protein database.

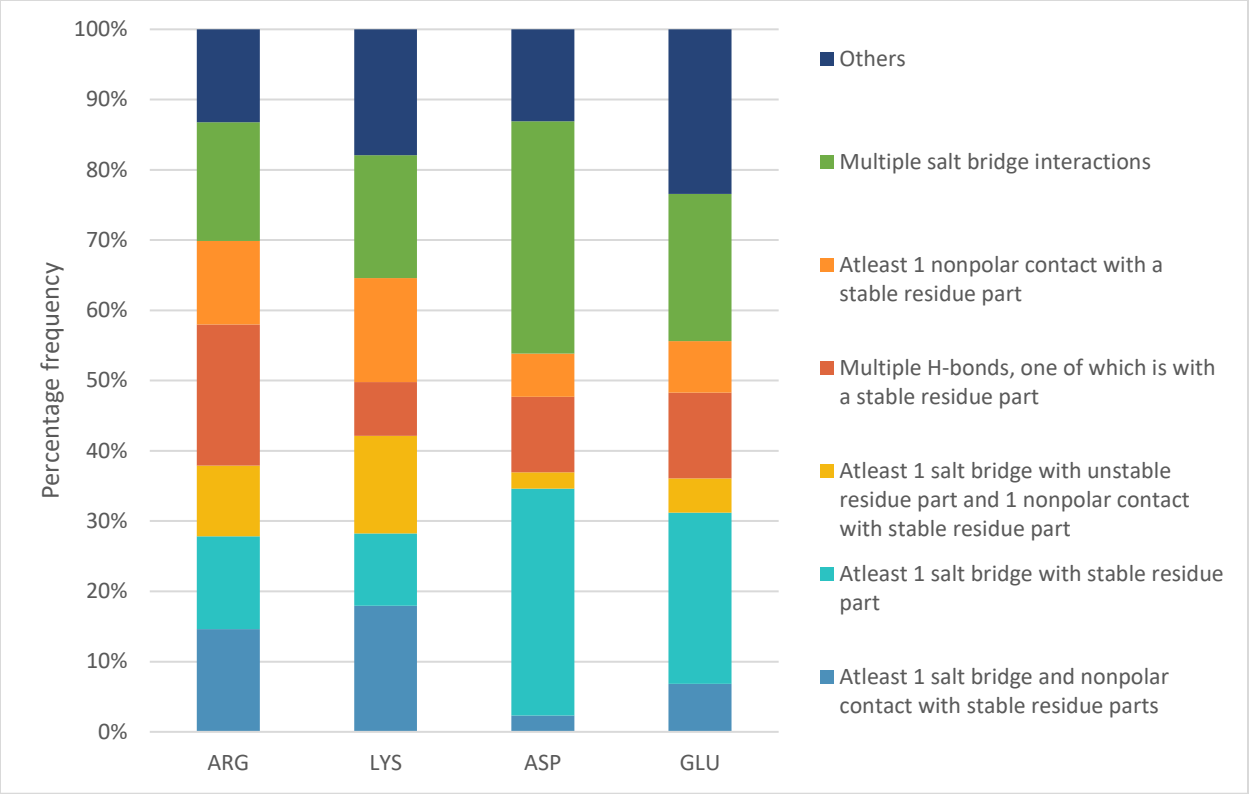

Figure 4.5 Percentage frequencies of unstable charged side chains forming different types of interactions in antibody-antigen complexes.

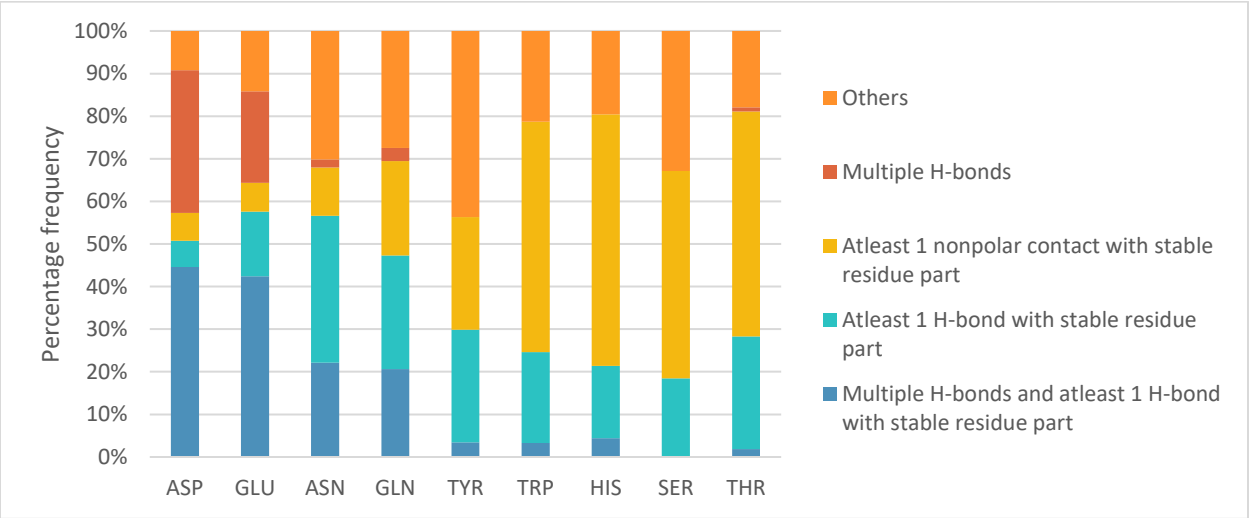

Figure 4.6 Percentage frequencies of polar charged side chains forming different types of interactions in antibody-antigen complexes.

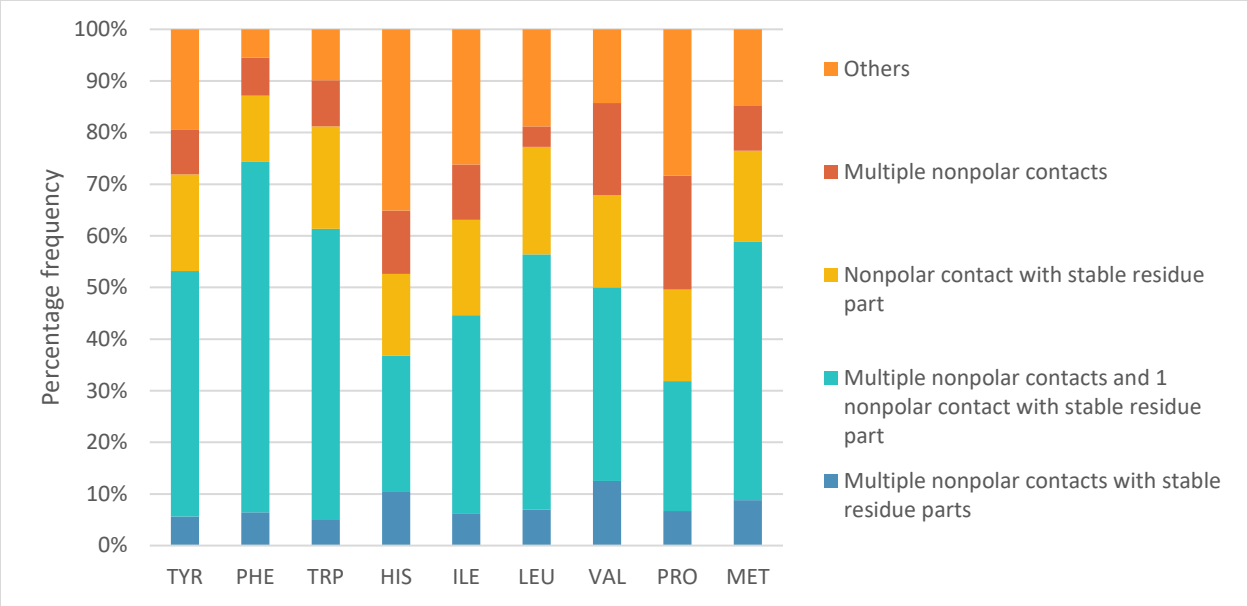

Figure 4.7 Percentage frequencies of unstable nonpolar side chains forming different types of interactions in antibody-antigen complexes.

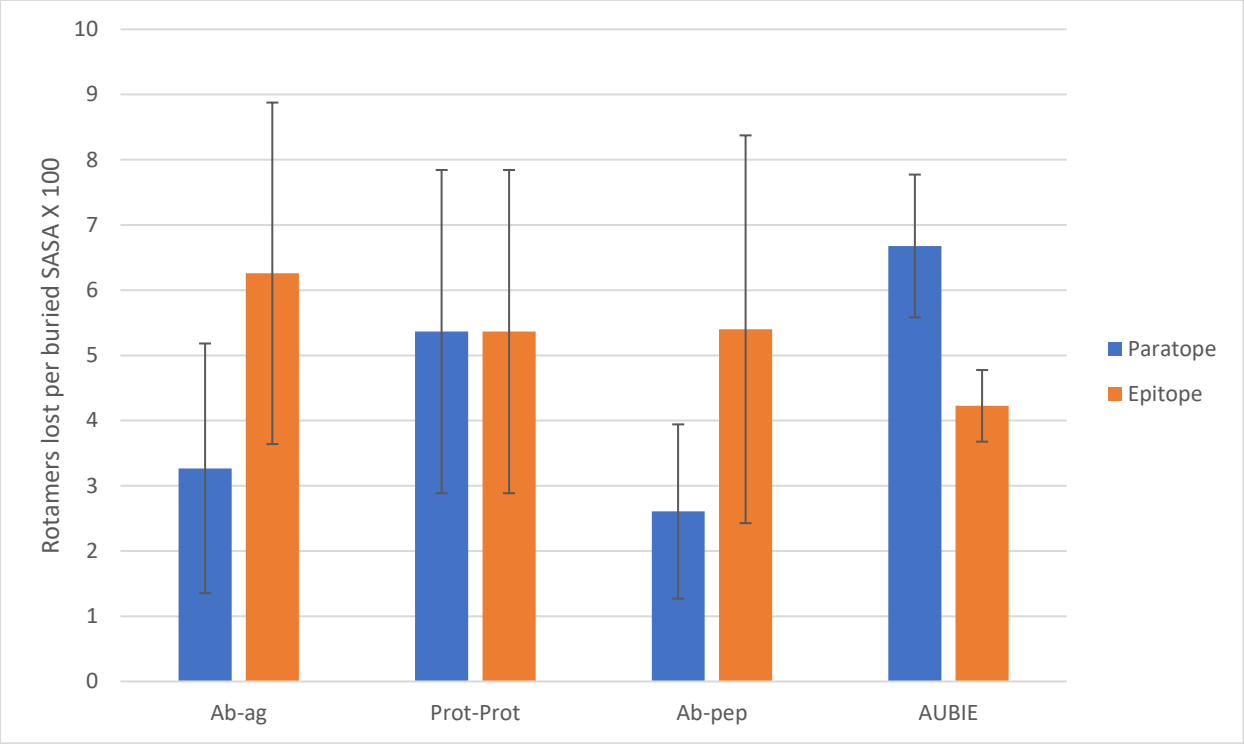

Figure 4.8 Average and standard deviations of RL-SASA values for epitopes and paratopes from different databases.

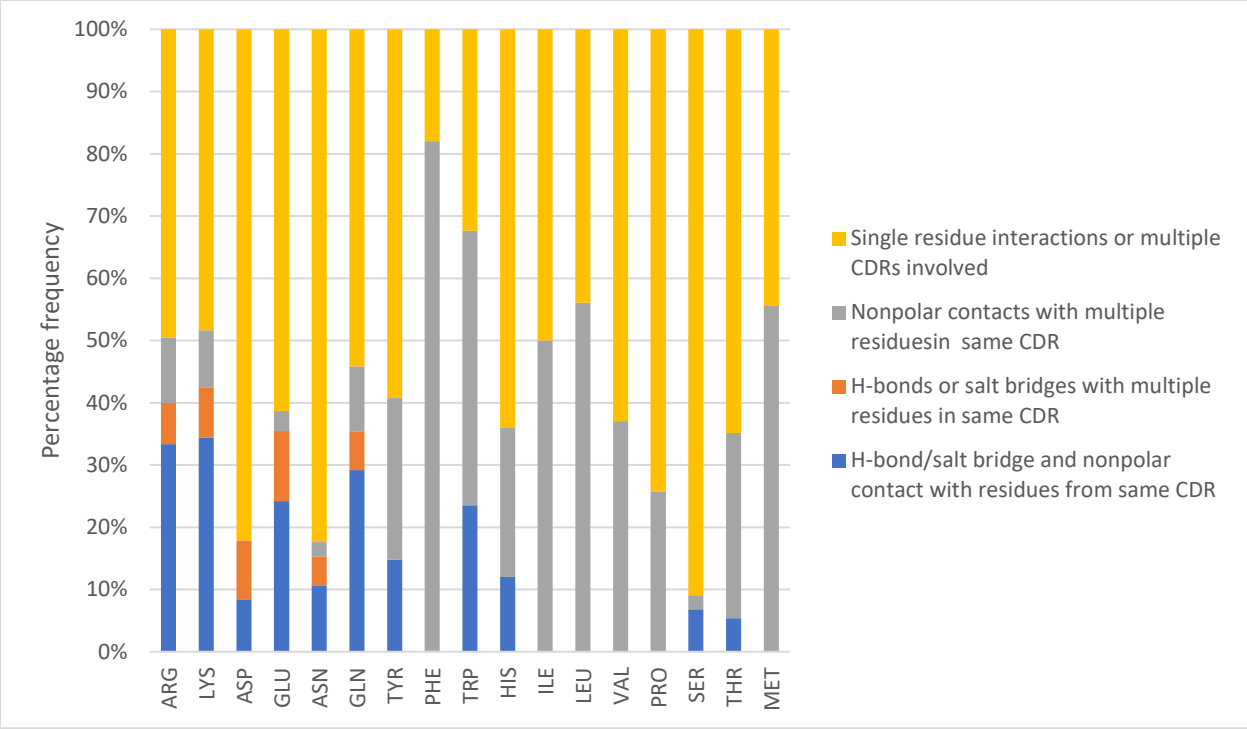

Figure 4.9 Percentage frequencies of unstable side chains forming multiple electrostatic interactions, multiple nonpolar contacts and both electrostatic and nonpolar interactions with multiple residues in the same CDR.

		ARG	LYS	ASP	GLU	GLN	ASN	THR	SER	TYR	TRP	PHE	HIS	MET	LEU	ILE	VAL	PRO	ALA	BB
FR1	H_1	0.0	0.0	0.0	4.3	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4
	H_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	0.0	0.1
	H_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0
	H_16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_24	0.0	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CDR1	H_27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	H_28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	H_29	0.0	0.0	0.0	0.0	0.0	0.0	6.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	1.3
	H_30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	H_34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_35	1.4	6.5	0.0	0.0	0.0	1.4	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	0.0	2.0
	H_36	0.0	0.0	1.6	0.0	0.0	1.4	10.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.8
	H_37	0.0	0.0	5.5	0.0	0.0	0.0	2.4	4.1	9.6	0.8	2.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	2.3
	H_38	1.9	3.2	3.1	0.0	0.0	4.1	4.0	2.0	5.9	14.3	4.6	3.1	0.0	4.5	0.0	0.0	1.6	12.0	1.2
FR2	H_39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_40	0.0	0.0	0.0	2.1	0.0	4.1	0.0	2.0	0.3	0.0	0.0	15.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_46C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_51	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_52	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_55	3.8	0.0	0.0	13.8	0.0	0.0	2.4	2.0	2.0	13.5	1.3	0.0	53.8	1.5	0.0	2.9	0.0	2.0	0.5
CDR2	H_56	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6
	H_57	0.0	3.2	10.2	2.1	0.0	14.9	3.2	10.2	7.4	7.4	4.6	3.1	15.4	10.4	37.5	2.9	0.0	4.0	0.7
	H_58	0.9	0.0	0.0	0.0	0.0	2.7	4.0	4.1	1.7	2.4	0.7	0.0	0.0	0.0	0.0	0.0	1.6	0.0	1.8
	H_59	1.9	3.2	1.6	0.0	5.0	2.7	3.2	0.0	0.0	0.0	2.0	0.0	15.4	4.5	37.5	0.0	0.0	5.0	3.4
	H_60	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.0	0.0	0.0
	H_61	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	H_62	0.0	0.0	3.9	0.0	0.0	0.0	1.6	4.1	0.5	0.0	10.5	0.0	0.0	0.0	0.0	0.0	1.6	0.0	3.6
	H_63	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1
	H_64	0.0	0.0	5.5	1.1	0.0	1.4	6.4	2.0	2.4	0.0	1.3	6.3	0.0	0.0	6.3	14.3	0.0	9.0	1.9
	H_65	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3	3.2	0.0	1.5
FR3	H_66	8.0	3.2	6.3	0.0	0.0	12.2	2.4	0.0	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	H_67	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
	H_68	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0
	H_69	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	0.0	0.9
	H_70	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_72	0.9	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_74	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_75	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_80	6.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_82	0.5	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	H_83	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	H_84	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.4	0.0	0.0	0.0	0.0	0.0	0.4
	H_84A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_84B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	0.0	0.0
	H_85	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_86	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_88	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_90	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 4.10 Percentage frequencies of interactions made at by all amino acid side chains and backbone atoms at antibody positions H1 to H90.

		ARG	LYS	ASP	GLU	GLN	ASN	THR	SER	TYR	TRP	PHE	HIS	MET	LEU	ILE	VAL	PRO	ALA	BB
CDR3	H_105	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	H_106	13.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	1.4	0.0	0.0	0.0
	H_107	4.7	6.5	7.9	8.5	15.0	0.0	1.6	4.1	0.7	5.0	3.3	6.3	0.0	4.5	0.0	0.0	0.0	2.0	0.9
	H_108	4.2	3.2	3.9	10.6	5.0	1.4	0.0	0.0	1.0	2.4	5.3	0.0	0.0	9.0	0.0	11.4	3.2	2.0	3.7
	H_109	1.9	0.0	3.1	2.1	5.0	2.7	0.0	0.0	3.4	1.6	1.3	9.4	0.0	0.0	0.0	7.1	19.0	6.0	4.9
	H_110	2.3	0.0	3.1	0.0	0.0	1.4	5.6	0.0	2.9	3.4	2.6	0.0	0.0	0.0	0.0	0.0	3.2	0.0	4.6
	H_111	0.0	0.0	0.8	1.1	0.0	0.0	0.0	0.0	2.4	0.0	3.9	0.0	0.0	4.5	0.0	5.7	0.0	7.0	3.6
	H_111A	0.0	0.0	0.8	1.1	0.0	0.0	2.4	2.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	4.3	3.2	1.0	2.7
	H_111B	2.8	0.0	0.0	0.0	0.0	0.0	0.8	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.9
	H_111C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	H_111D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	3.2	3.0	0.5
	H_111E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_112	1.9	0.0	3.9	4.3	0.0	4.1	0.0	0.0	1.4	3.2	0.7	0.0	0.0	3.0	0.0	0.0	7.1	0.0	1.7
	H_112A	0.9	0.0	0.0	0.0	0.0	2.7	2.4	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.5	0.0	2.3
	H_112B	0.5	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.2
	H_112C	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.1	0.0	0.0	0.0	0.0	3.2	3.0	1.2
	H_112D	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.4
	H_112F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	H_113	1.9	0.0	6.3	3.2	0.0	4.1	0.8	0.0	4.5	7.4	2.6	12.5	0.0	10.4	0.0	1.4	3.2	4.0	1.7
	H_114	1.4	0.0	0.0	1.1	0.0	0.0	0.0	2.0	1.2	1.9	0.7	0.0	0.0	6.0	0.0	2.9	1.6	1.0	1.0
	H_115	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	0.0	0.0	1.5	0.0	1.4	0.0	1.0	0.2
	H_116	1.4	0.0	10.2	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	3.0	0.0	1.4	0.0	1.0	0.0
	H_117	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FR1	L_1	0.0	0.0	2.4	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
	L_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.3	1.4	0.0	0.0	0.1
	L_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1
CDR1	L_27	0.0	0.0	0.0	2.1	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	L_28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	L_29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.8	0.0	0.4
	L_30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_31	0.9	0.0	1.6	2.1	0.0	0.0	4.0	0.0	1.4	0.0	0.0	15.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_32	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.3
	L_33	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.2
	L_34	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	L_35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.4
	L_36	3.8	6.5	3.1	1.1	0.0	2.7	0.0	28.6	1.5	0.0	3.9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.3
	L_37	0.0	6.5	0.0	0.0	0.0	8.1	4.0	2.0	1.7	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5
	L_38	4.7	9.7	1.6	0.0	0.0	2.7	0.0	0.0	9.3	1.6	7.2	0.0	0.0	0.0	0.0	1.4	0.8	6.0	0.9
FR2	L_40	0.5	0.0	0.0	1.1	0.0	5.4	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	L_42	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_52	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.5	0.0	0.0	0.0	0.0	0.0
	L_55	0.0	3.2	0.0	1.1	0.0	0.0	0.0	0.0	7.9	0.0	2.6	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
CDR2	L_56	5.6	19.4	0.8	4.3	0.0	0.0	1.6	0.0	3.5	2.4	0.0	0.0	0.0	9.0	0.0	0.0	0.0	5.0	0.5
	L_57	0.0	3.2	2.4	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	0.0	1.0	0.1
	L_65	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4

Figure 4.11 Percentage frequencies of interactions made at by all amino acid side chains and backbone atoms at antibody positions H105 to L65.

		ARG	LYS	ASP	GLU	GLN	ASN	THR	SER	TYR	TRP	PHE	HIS	MET	LEU	ILE	VAL	PRO	ALA	BB
FR3	L_66	3.3	3.2	0.0	0.0	20.0	2.7	7.2	2.0	0.1	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	0.0	0.0
	L_67	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
	L_68	0.5	0.0	0.8	6.4	0.0	0.0	0.0	0.0	1.2	0.0	0.7	6.3	0.0	0.0	0.0	0.0	2.4	0.0	0.4
	L_69	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	3.0	0.6
	L_70	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	L_71	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	L_72	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.1
	L_74	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.1
	L_78	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	L_79	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_80	1.9	9.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	L_83	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_84	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
CDR3	L_105	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0
	L_106	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	L_107	4.2	0.0	0.8	0.0	5.0	1.4	0.8	0.0	4.4	6.3	7.2	6.3	0.0	3.0	0.0	0.0	0.0	0.0	3.5
	L_108	2.3	0.0	2.4	1.1	15.0	6.8	0.0	4.1	0.8	11.1	1.3	3.1	0.0	1.5	0.0	0.0	0.0	2.0	6.8
	L_109	0.0	0.0	0.0	3.2	0.0	5.4	3.2	6.1	1.5	1.6	0.0	0.0	0.0	0.0	0.0	8.6	0.8	1.0	3.1
	L_110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	L_113	0.0	0.0	0.0	0.0	0.0	0.0	4.0	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.9
	L_114	0.0	0.0	2.4	0.0	0.0	1.4	5.6	0.0	0.1	0.0	1.3	0.0	0.0	1.5	0.0	2.9	10.3	4.0	4.5
	L_115	0.0	0.0	0.8	0.0	0.0	0.0	0.8	0.0	0.0	1.9	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.9
	L_116	4.2	0.0	0.0	16.0	5.0	0.0	0.0	0.0	3.0	7.1	5.9	0.0	0.0	6.0	0.0	0.0	4.0	2.0	0.4
	L_117	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	5.9	0.0	0.0	0.0	0.0	1.4	0.0	0.0	0.1

Figure 4.12 Percentage frequencies of interactions made at by all amino acid side chains and backbone atoms at antibody positions L66 to L117.

References

[1] X. Qiao, L. Qu, Y. Guo, and T. Hoshino, “Secondary Structure and Conformational Stability of the Antigen Residues Making Contact with Antibodies,” *The Journal of Physical Chemistry B*, vol. 125, no. 41, pp. 11374–11385, Oct. 2021, doi: 10.1021/acs.jpcb.1c05997.

[2] V. N. Uversky and M. H. v van Regenmortel, “Mobility and disorder in antibody and antigen binding sites do not prevent immunochemical recognition,” *Critical Reviews in Biochemistry and Molecular Biology*, vol. 56, no. 2, pp. 149–156, Mar. 2021, doi: 10.1080/10409238.2020.1869683.

[3] M. L. Fernández-Quintero, J. R. Loeffler, L. M. Bacher, F. Waibl, C. A. Seidler, and K. R. Liedl, “Local and Global Rigidification Upon Antibody Affinity Maturation ,” *Frontiers in Molecular Biosciences* , vol. 7. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmolb.2020.00182

[4] A. K. Mishra and R. A. Mariuzza, “Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing ,” *Frontiers in Immunology* , vol. 9. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2018.00117

[5] G. J. Bartlett, C. T. Porter, N. Borkakoti, and J. M. Thornton, “Analysis of Catalytic Residues in Enzyme Active Sites,” *Journal of Molecular Biology*, vol. 324, no. 1, pp. 105–121, 2002, doi: https://doi.org/10.1016/S0022-2836(02)01036-7.

[6] M. Wang, D. Zhu, J. Zhu, R. Nussinov, and B. Ma, “Local and global anatomy of antibody-protein antigen recognition,” *Journal of Molecular Recognition*, vol. 31, no. 5, p. e2693, May 2018, doi: https://doi.org/10.1002/jmr.2693.

[7] S. J. Fleishman, S. D. Khare, N. Koga, and D. Baker, “Restricted side chain plasticity in the structures of native proteins and complexes,” *Protein Science*, vol. 20, no. 4, pp. 753–757, Apr. 2011, doi: https://doi.org/10.1002/pro.604.

[8] S. J. Fleishman, J. E. Corn, E.-M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker, “Hotspot-centric de novo design of protein binders,” *J Mol Biol*, vol. 413, no. 5, pp. 1047–1062, Nov. 2011, doi: 10.1016/j.jmb.2011.09.001.

[9] X. Liu *et al.*, “Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping,” *Scientific Reports*, vol. 7, no. 1, p. 41306, 2017, doi: 10.1038/srep41306.

[10] T. de Meyer, S. Muyldermans, and A. Depicker, “Nanobody-based products as research and diagnostic tools,” *Trends in Biotechnology*, vol. 32, no. 5, pp. 263–270, 2014, doi: https://doi.org/10.1016/j.tibtech.2014.03.001.

[11] Z. Liu *et al.*, “PDB-wide collection of binding data: current status of the PDBbind database,” *Bioinformatics*, vol. 31, no. 3, pp. 405–412, Feb. 2015, doi: 10.1093/bioinformatics/btu626.

[12] B. R. Brooks *et al.*, “CHARMM: the biomolecular simulation program,” *J Comput Chem*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[13] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, Jun. 2017, doi: 10.1021/acs.jctc.7b00125.

[14] A. B. Rubenstein, K. Blacklock, H. Nguyen, D. A. Case, and S. D. Khare, “Systematic Comparison of Amber and Rosetta Energy Functions for Protein Structure Evaluation,” *Journal*

of Chemical Theory and Computation, vol. 14, no. 11, pp. 6015–6025, Nov. 2018, doi: 10.1021/acs.jctc.8b00303.

[15] M. D. Smith, J. S. Rao, E. Segelken, and L. Cruz, "Force-Field Induced Bias in the Structure of Aβ21–30: A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields," *Journal of Chemical Information and Modeling*, vol. 55, no. 12, pp. 2587–2595, Dec. 2015, doi: 10.1021/acs.jcim.5b00308.

[16] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, "Development and Analyses of a Database of Antibody – Antigen Complexes," in *Computer Aided Chemical Engineering*, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds. Elsevier, 2018, pp. 2113–2118. doi: https://doi.org/10.1016/B978-0-444-64241-7.50347-5.

[17] P. A. Wood, F. H. Allen, and E. Pidcock, "Hydrogen-bond directionality at the donor H atom—analysis of interaction energies and database statistics," *CrystEngComm*, vol. 11, no. 8, pp. 1563–1571, 2009, doi: 10.1039/B902330E.

[18] C. Roy and S. Datta, "ASBAAC: Automated Salt-Bridge and Aromatic-Aromatic Calculator," *Bioinformation*, vol. 14, no. 4, pp. 164–166, Apr. 2018, doi: 10.6026/97320630014164.

[19] R. L. Dunbrack Jr and F. E. Cohen, "Bayesian statistical analysis of protein side-chain rotamer preferences," *Protein Sci*, vol. 6, no. 8, pp. 1661–1681, Aug. 1997, doi: 10.1002/pro.5560060807.

[20] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin," *Journal of Molecular Biology*, vol. 79, no. 2, pp. 351–371, 1973, doi: https://doi.org/10.1016/0022-2836(73)90011-9.

[21] J. Jankauskaitė, B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio, and I. H. Moal, "SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation," *Bioinformatics*, vol. 35, no. 3, pp. 462–469, Feb. 2019, doi: 10.1093/bioinformatics/bty635.

[22] C.-W. Lee, H.-J. Wang, J.-K. Hwang, and C.-P. Tseng, "Protein thermal stability enhancement by designing salt bridges: a combined computational and experimental study," *PLoS One*, vol. 9, no. 11, pp. e112751–e112751, Nov. 2014, doi: 10.1371/journal.pone.0112751.

[23] P. Hung-Pin, L. K. Hao, J. Jhih-Wei, and Y. An-Suei, "Origins of specificity and affinity in antibody–protein interactions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 26, pp. E2656–E2665, Jul. 2014, doi: 10.1073/pnas.1401131111.

[24] T. A. Whitehead, D. Baker, and S. J. Fleishman, "Chapter One - Computational Design of Novel Protein Binders and Experimental Affinity Maturation," in *Methods in Protein Design*, vol. 523, A. E. B. T.-M. in E. Keating, Ed. Academic Press, 2013, pp. 1–19. doi: https://doi.org/10.1016/B978-0-12-394292-0.00001-1.

5. Chapter 5 - AUBIE Modifications

From Chapters 3 and 4, I learned that AUBIE designs consisted of relatively large numbers of highly flexible CDR residues which led to them having highly unstable paratopes before binding. Using the knowledge from Chapter 4 and better experience with force field performance, modifications were made to certain parameters, interaction selection criteria, minimization routines and residue steric clash filters to generate antibodies with more stable paratopes. The modified AUBIE approach was used to design antibodies for 25 randomly selected antigens and analyzed for binding metrics. This Chapter will describe the modifications and new results. This work will be submitted for publication in 2022.

AUBIE changes

Types of interactions

The prestabilization features studied in Chapter 4 demonstrated that interactions between unstable side chains are rare and interactions are likely to consist of at least one stable side chain or backbone element. Furthermore, paratopes are filled with low entropy, hydrophilic amino acids. Thus, the AUBIE approach was modified to search for H-bonds involving at least one backbone atom. All interactions made by antibody ARG, LYS and GLN residues were omitted. H-bonds between antibody SER, THR and TYR and antigen ASP were allowed since H-bonds between these residues were relatively more frequent than other types of H-bonds in the antibody-antigen database. To ensure the search for salt bridges, only negatively charged paratope residues were allowed to form them. From the analysis of the types of interactions that were being selected in AUBIE solutions, it was observed that almost all the interactions were H-bonds. This may have been caused due to the distance thresholds being too tight for the other cation-π and π-π

interactions. To reduce computational runtimes, only H-bonds were searched for in the current implementation. All the permitted types of H-bonds are listed in Table 5.1.

Distance threshold

The distance matching threshold was increased from 0.33 Å to 0.50 Å to increase the solution space searched after the types of interactions searched were reduced. The new threshold resulted in large enough solution sets as discussed in the results.

Residue clashing

In the previous implementation of AUBIE, residue side chain atoms were allowed to clash under the assumption that energy minimizations would fix the clashes. A more conservative approach was used here to lower the odds of prestabilized interactions in the epitope being affected. Thus a new definition for steric clashes between residues was defined. Two residues were defined to have a steric clash if any of the following conditions were met: 1) Two heavy (i.e., non-hydrogen) atoms were closer than 1.3 Å, 2) More than one pair of heavy atoms were closer than 1.8 Å, or 3) One pair of heavy atoms was closer than 1.8 Å, with at least one of the two atoms being a backbone atom. Instead of calculating vdW energy between atoms like conventional approaches, AUBIE used these relaxed clash constraints under the assumption that the flexibility of the proteins would compensate for the minor steric clashes introduced during antigen placement. These clash definitions were selected because they consistently allowed CHARMM36 energy minimization to rectify the clashing structures, which was used as a proxy for the proteins' flexibility. Designs that had steric clashes between antigen and framework residues were deleted. Designs that had steric clashes between the antigen and binding loops were passed through a clash resolving mutation step.

Mutation of clashing residues

Once a design was identified, a novel feature-based approach was used to mutate binding loop residues for one of two purposes: 1) resolving clashing variable side chains and 2) improving binding features. The mutation decisions were based on features obtained after analyzing the database of antibody-antigen complexes. The affinity maturation mutations were conservative in nature and made changes with high confidence of creating either nonpolar or polar interactions since the approach avoids any energy minimizations.

The following steps were taken for the binding loop residues that had steric clashes with the antigen:

1. Identify an alternate rotamer of the native amino acid that does not have steric clash with any surrounding antigen or scaffold residue.
2. If step 1 fails, search for an ASP rotamer that forms a salt bridge with an antigen residue. Salt bridges are prioritized because they are the strongest polar interactions and are less orientation dependent than H-bonds. Each antigen ARG/LYS is allowed to form a salt bridge with only one mutated paratope residue to avoid the formation of closely placed negatively charged residues.
3. If step 2 fails, search for a GLU rotamer that forms a salt bridge with an antigen residue. ASP is prioritized over GLU for salt bridge formation since it is a smaller side chain and thus its stabilization involves a lower entropic cost. Moreover, ASP salt bridges are more frequent than GLU salt bridges in known antibody-antigen complexes [1]. AUBIE does not consider introducing ARG/LYS mutations for salt bridge formation due to their high side chain flexibilities.
4. If step 3 fails, rotamers that can introduce new H-bonds are searched for. If a polar rotamer does not have steric clashes and forms the required type and number of H-bonds with the

antigen, it is selected. Polar rotamers are searched in the sequence: SER, THR, ASN, ASP and HIS. The required minimum number of H-bonds is one for SER and THR and two for ASP, ASN and HIS. It was observed in the antibody-antigen database that ASN rotamers frequently make hydrogen bonds with backbone atoms. Hence ASN rotamers making single H-bonds with antigen backbone atoms were also selected. GLN and GLU rotamers were not analyzed due to their relatively high entropy side chains.

5. If step 4 fails, rotamers that introduce nonpolar contacts are searched for. The nonpolar amino acids analyzed in this step are ALA, VAL, LEU, ILE and PHE. A side chain atom is defined to be making a nonpolar contact if it is either a C or S atom and is less than 4.5 Å away from an antigen C or S atom. MET is not considered due to the relatively high entropy of its side chain. TRP and TYR are also not considered because they need to satisfy both polar and nonpolar requirements when buried in an interface. The rotamer make the most nonpolar contacts is selected.
6. If step 5 fails, the pose is rejected because it has an irreconcilable steric clash.

Only one mutation per binding loop was allowed to lower the odds of the mutations impacting the loop structures. Following the mutations, the clash-free mutated designs were output in PDB format. In the current version of AUBIE, no further pose refinement or ranking is performed and rotamer repacking/energy minimization and pose ranking strategy is left to the user.

Solvent exposed epitope atoms

In the previous AUBIE approach, all the epitope residue atoms were searched through during solution identification steps. This led to buried atoms being part of interactions that were always destined to have steric clashes. In order to avoid wasting computational load on such

solutions, only epitope atoms with non-zero SASA were allowed to be a part of any AUBIE solution.

Solution similarity

In the previous AUBIE approach, the maximum total CDR loop similarity allowed between two unique designs was 60%. This limit was considered to be relatively relaxed since the chances of generating antibody designs with three same CDR loops is increased when the number of interactions searched for is reduced to four or lower. The maximum limit was reduced to 40% to increase solution diversity.

Results

The modified AUBIE approach was tested through antibody design runs for known antibody-binding epitopes. For the database generation step, antibody structure from PDB file 1N8Z was used. It was selected since its HCDR3 loop allowed reasonable loop length to replace without mutating key hydrophobic residues important for VH-VL domain binding. Similar to the database from generated in Chapter 3, the six CDR regions were selected as binding regions with the modification that more conservative CDR3 hinge residues were selected. The selected binding loops and framework structures are shown in Figure 5.1. Maximum limits of six residues and 750 loops was set for each binding region. The generated database consisted of 212 loops for HCDR1 region and 750 loops for all the other CDRs. The generated database was used to design antibodies for known epitopes on 25 randomly selected antigens from the antibody-antigen database as listed in Table 5.2. Epitopes for the 25 design runs were manually selected from their native antibody-binding complexes and ranged from 10 to 15 residues as listed in Table 5.2. A minimum limit of four interactions was enforced for all design runs. Upon testing the three force fields, it was

observed that CHARMM36 was better than AMBER and Rosetta in resolving the minor steric clashes that exist in AUBIE solutions. Hence, fixed backbone vacuum energy minimization with CHARMM36 force field [2] was used to resolve steric clashes and then an all-atom energy minimization with Rosetta's REF2015 force field [3] was performed to relax the complex since it the most widely used force field in protein engineering. All structures from the antibody-antigen database were run through the same minimization protocol before using their binding metrics for comparison purposes. All binding metrics were calculated by Rosetta's InterfaceAnalyzer module [4]. All design runs were performed on single Intel Xeon Gold 6248R processors with 3.00 GHz speed.

The binding metrics of the top AUBIE designs were analyzed and compared to the binding metrics from known and wild-type (WT) antibody complexes. The runtimes for 25 design runs ranged from 6 hours, for antigen from 3P30, to 46.5 hours, for antigen from 4HEP. As mentioned in Chapter 3, AUBIE runtimes depend on the number of solvent exposed polar epitope atoms and initial orientation of binding loops towards epitope. Table 5.3 lists the number of designs and binding metrics of the WT complex and the best AUBIE design generated for each design run. Apart from the antigen from complex 3P30, AUBIE was able to generate at least 150 designs for each run. The antigen from complex 3P30 consists of two helices and hence does not have any backbone atoms to offer for H-bond formation. This led to a substantial limitation in the solution space of low entropy H-bonds being targeted and thus resulted in only 13 designs and low runtime. AUBIE was able to generate two top designs that had binding energies > 50 kcal/mol and seven other top designs that had binding energies > 40 kcal/mol. 18 out of the 25 top designs had binding energies within one standard deviation of the mean binding energies from the antibody antigen database (53.11 ± 16.00 kcal/mol). These results show that the modified version of AUBIE is

capable of generating designs with comparable performance to known antibodies but is not able to generate superior designs with binding energies > 60 kcal/mol.

Figures 5.2 and 5.3 show the pre-minimization interfaces of five top AUBIE designs with their targeted interactions. When compared with the WT complexes, the top AUBIE design had better (>3 kcal/mol difference), similar and worse binding energies in 4, 4, and 17 cases respectively. AUBIE was not able to design antibodies that could match up to the WT complexes due to multiple reasons. AUBIE targets all its interactions through the CDRs, while natural antibodies can also use framework residues to form interactions with the antigen (Figure 5.4), Moreover, natural antibodies often use long HCDR3 loops to form multiple interactions with the epitope. Meanwhile, AUBIE does not perform loop targeted design and if it were to find designs with long CDR3s making many interactions, it does so by chance. Hence WT complexes can form larger binding interfaces and better binding energies than AUBIE designed complexes.

The other binding metric analyzed was shape complementarity (Sc), which is an indicator for how well the epitope and paratope surfaces structurally fit with each other. While half of the top AUBIE designs had Sc scores below 0.6, AUBIE was able to generate other complexes with higher Sc scores for each design run. AUBIE designs can have relatively poor Sc scores when compared with Sc values from known complexes (Database average 0.68 ± 0.07) since shape fitting is not a targeted feature in the design steps.

Along with the binding metrics, the number and types of mutations being made during the clash resolution step were also analyzed. Table 5.4 lists the fraction of designs, on a per-CDR basis, that had zero mutations for each design run. The results showed that a small fraction of CDR loops required a single mutation to resolve the steric clashes. The largest fraction of loops that required single mutations, around 14%, belonged to the CDR2 loops in the antibodies designed to

bind the antigen from 2XT1. This low presence of mutations is another reason why most AUBIE designs do not have better than average binding metrics. Natural antibodies go through somatic affinity maturation to increase binding affinity [5]. Meanwhile, AUBIE designs antibodies by combining compatible loops from the PDB without any further affinity maturation. Table 5.5 lists the initial and final amino acid distributions in residues that were mutated in all the design runs. The amino acids that were most frequently mutated include large inflexible chain amino acids such as TYR, TRP, HIS and PHE. Mutations were mainly being made to LEU to fill up cavities left by the native large side chains and to other small amino acids like ASP, ASN, ALA and SER.

Next, the current and old Chapter 3 AUBIE designs were compared in terms of conformational stability before binding. RL-SASA values for the top designs for the 25 cases were computed and averages and standard deviations were plotted as shown in Figure 5.5. The average AUBIE designed paratope had lower RL-SASAs than the ones from the old AUBIE designs but were still larger than RL-SASAs from known antibody paratopes. The large standard deviation of RL-SASAs from new AUBIE designs indicated that AUBIE was now able to generate top designs that had comparable paratope rigidity to known antibodies. The mutations to LEU were a likely contributor to increasing the RL-SASAs of the new AUBIE designs.

Unlike other antibody design approaches, the current version of AUBIE does not perform any computational affinity maturation. The improvements in binding energies upon affinity maturation of AUBIE designs were explored. RAbD's sequence design feature [6] was used to modify the CDR sequence of 13 top designs from Table 5.3 that had large binding energy differences with their WT complexes. The inner_kt, outer_kt, seq_design_profile_samples, nstruct and outer_cycle_rounds parameters were set as 2.0, 2.0, 50, 300 and 50 respectively. The results of the affinity maturation are listed in Table 5.6. For nine designs, the affinity maturations

improved the binding energies and for five of these designs, the affinity matured designs had better binding energies than the WT complexes. For four designs, RAbD's random search approach was unable to find any beneficial mutations. RAbD made a large number of mutations in all improved designs. The native and improved interfaces with the mutated side chains of 3P9W-antigen binding AUBIE design is shown in Figure 5.6. Several mutations were made in the stem regions of CDRs which were not involved in antigen binding. Despite the large number of mutations, several AUBIE-designed loops were unchanged as shown in the figure.

AUBIE was compared with OptMAVEn 2.0 [7], another *de novo* antibody design approach for target epitopes. Both the approaches were used to design antibodies binding to the Herceptin epitope on HER2. The OptMAVEn 2.0 designs run were done by Dr. Ratul Chowdhury, the program's lead developer. Designs generated by both methods were run through the same minimization protocol. The top three binding energies from the AUBIE and OptMAVEn 2.0 simulations were -40.84, -40.66, -40.43 kcal/mol and -43.45, -37.42, -33.82 kcal/mol. Even without affinity maturation, AUBIE was able to generate designs with comparable binding energies to the ones made by OptMAVEn 2.0.

MD simulations were used to further test the quality of the top AUBIE designed binding interfaces. Implicit solvent NpT (isobaric and isothermal) MD simulations with 5 ns long equilibration steps [8] and 50 ns long production steps [7]using NAMD 2.14 were performed for AUBIE designs with the best and worst binding energies: -52.07 kcal/mol from 3BDY-ab and -31.70 kcal/mol from 4LMQ-ab. The configuration files were set up using QwikMD and default parameters were used. Both the complexes remained bound at the end of the MD simulations. Figures 5.7 compare the initial and final frames after aligning the antibody structures of both the MD simulations. While the AUBIE design for 4LMQ antigen remained bound, the MD simulation

substantially rearranged the binding interface to form new strong beneficial interactions that were not designed for. Thus it was concluded that the complex remained bound mainly by chance. Similar to 4LMQ MD simulation, the MD simulations for 3BDY-AUBIE design also identified new strong beneficial interactions that helped stabilized the complex but few interactions from the starting complex persisted throughout the MD simulation. These interactions included salt bridge between 56D and 35K and hydrophobic interactions between 38L and 68M. Moreover, 107N and 76Q were in close proximity even though the initial H-bond between them was not present in the final frame. These interactions are shown in Figure 5.8.

Discussion

AUBIE is a first-of-its-kind computational method that can design antibodies for any specific epitope. The workflow has been divided into two broad steps: database generation and protein design. The diversity of naturally produced antibody structures and sequences is limited by the genetic code. By combining loops available in the PDB, AUBIE explores a design space that the human immune system cannot search for and hence generates never-seen-before designs. AUBIE uses this innovative *de novo* approach to rapidly generate high affinity designs without any energy functions or optimization formulations as seen in other software. Current software often use such energy functions to predict structure changes due to residue mutations [6], [7], [9]. Rather than determining backbone structures using such energy functions, AUBIE makes use of the sequence and structure information already available in the PDB in the database generation step. Antibodies are known to bind via few strong "hotspot" interactions. Our results from Chapter 2 supported this conclusion. It was hypothesized that one could still generate high affinity solutions by engineering known geometric requirements of strong interactions and avoiding obvious

energetically expensive steric and charge clashes. Aided with the knowledge from Chapter 3, these strong interactions were defined to be low entropy H-bonds. This hypothesis was supported by the results from this Chapter. Without using any affinity maturation, AUBIE was able to design antibodies that had comparable binding energies to known antibody-antigen binding complexes. The close proximity of the two binding proteins allowed for the formation of extra H-bonds and hydrophobic interactions that lead to the acceptable binding energies in the top designs. Furthermore, the modifications made in this Chapter led to the desired rigidification of AUBIE paratopes. The top AUBIE designs had poor Sc scores when compared with the antibody-antigen database. While the design approach has negative design elements such as rejecting solutions with steric and charge clashes, it lacks positive design steps to encourage the selection of well-docked solutions. This shortcoming requires the user to manually compare individual Sc values while deciding which designs to experimentally test.

When compared with the current state-of-the-art software in computational *de novo* antibody design, AUBIE offers several advantages. An advantage AUBIE possess over other software is its lack of iterative energy calculations in the design workflow, which could potentially lead to considerable reduction in runtimes once further updates are made. One of these updates includes moving the AUBIE code from Cython to C++. Furthermore, both OptMAVEN 2.0 and RAbD have been developed to design only antibody structures while AUBIE is capable of generating structures of protein binders, including antibodies, which have modular parts such as CDRs. An example of such a binding scaffold is 10^{th} type human fibronectin domain. AbDesign can be used to design alternative binding scaffolds but requires a database of protein structures similar to the scaffold. Such databases are not readily available for several scaffolds. AUBIE requires no such database and uses only the single input scaffold structure provided. Though I

acknowledge that AUBIE and Baker et al.'s approach share a similar step of modelling target high affinity residue positions, there are differences in how this step is implemented. While Baker et al.'s work identifies ideal paratope residue positions using RosettaDock [10], AUBIE identifies ideal epitope atom positions using simple distance and angle information from literature for various interaction types. The subsequent steps in both the approaches are considerably different. Our primary goal with regards to AUBIE was to develop an approach that designed binding antibodies rapidly. There is still significant improvement needed even though AUBIE is now able to generate designs with acceptable metrics. I believe that AUBIE's sequential search approach and simple distance matching algorithm can eventually lead to faster design generation after further modifications.

Table 5.1 Antibody and antigen amino acid types allowed to form H-bonds in AUBIE solutions.

Antibody	Antigen
Backbone	Backbone
Backbone	All side chains
All side chains except R, K and Q	Backbone
S, T, Y	D
D, E	R, K

Table 5.2 Epitope residue numbers in the 25 AUBIE test cases and HER-2 binding design run.

Ag PDB ID	Epitope residue numbers
3L5Y	2, 6, 7, 9, 10, 13, 74, 75, 76, 78, 79, 81, 82, 83, 84
5EPM	9, 11, 12, 23, 24, 25, 26, 27, 30, 32
4LMQ	9, 10, 11, 12, 13, 14, 15, 17, 35, 36, 37, 38, 39, 40
4LQF	19, 20, 22, 23, 57, 58, 59, 64, 66, 67, 68, 69, 70
3X3F	6, 16, 17, 18, 19, 29, 33, 35, 36, 37, 38, 39, 40, 41, 42
4N9G	7, 10, 11, 12, 13, 14, 16, 17, 19, 20, 23
3P9W	28, 29, 30, 31, 62, 63, 64, 66, 69, 71, 80, 81, 82, 83, 84
1OB1	8, 9, 10, 11, 14, 23, 28, 36, 39, 40
4UU9	14, 16, 17, 18, 19, 25, 30, 33, 41, 43, 44, 45, 46, 47, 48
2YBR	7, 10, 13, 15, 16, 17, 18, 22, 24, 42, 43, 44, 45, 64
4HEP	3, 4, 5, 20, 21, 24, 27, 28, 33, 34, 41, 42, 43
5DFV	17, 20, 23, 24, 25, 32, 48, 49, 51, 52, 56, 59, 60
2XT1	2, 3, 4, 5, 6, 7, 26, 30, 31, 32, 33, 43, 44, 47, 48
4AL8	13, 14, 16, 17, 18, 19, 20, 24, 26, 66, 67, 68, 69, 71
2H9G	5, 6, 16, 17, 18, 19, 30, 33, 36, 38, 39, 42
5IKC	10, 11, 12, 13, 14, 44, 45, 46, 48, 80, 82, 83, 84, 85, 86
2BDN	27, 28, 29, 30, 31, 34, 35, 36, 37, 38, 52, 53, 58, 62, 65
4DGI	15, 16, 17, 18, 19, 20, 21, 22, 26, 79, 83, 87
4P2C	14, 16, 17, 18, 20, 27, 29, 31, 32, 33, 53, 59, 60, 61, 63
1JRH	37, 39, 40, 41, 42, 43, 44, 45, 46, 74, 88
3UZQ	8, 9, 10, 11, 12, 13, 14, 29, 64, 90, 91, 92
3NH7	10, 31, 32, 39, 44, 45, 46, 48, 53, 58, 59, 60, 61, 62, 64
3BDY	35, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80
3KR3	2, 6, 7, 8, 11, 14, 24, 31, 33, 36, 40, 41, 43, 45, 55
3P30	55, 56, 59, 60, 62, 63, 64, 66, 67, 69, 70, 71, 73, 74, 77
HER2 (1N8Z)	532, 558, 560, 561, 569, 570, 573, 593, 594, 597, 598, 602

Table 5.3 Total number of designs, binding energies and Sc scores of top AUBIE designs and WT complexes for the 25 antigens.

Ag PDB ID	Solutions	Top AUBIE BE (kcal/mol)	Sc of Top design	WT BE (kcal/mol)	WT Sc	WT BE - AUBIE BE (kcal/mol)
4LQF	210	-33.78	0.64	-26.59	0.63	7.20
3BDY	1016	-52.07	0.67	-45.67	0.68	6.40
4DGI	512	-32.74	0.57	-28.55	0.73	4.18
4P2C	361	-39.42	0.64	-36.26	0.62	3.16
3P30	13	-35.31	0.47	-32.73	0.65	2.58
4AL8	364	-43.32	0.59	-40.83	0.67	2.49
2BDN	365	-40.56	0.62	-41.00	0.66	-0.45
1OB1	900	-39.14	0.55	-41.98	0.69	-2.84
4N9G	564	-33.47	0.53	-36.77	0.67	-3.30
3X3F	544	-50.85	0.66	-56.51	0.76	-5.66
2H9G	637	-49.75	0.69	-56.08	0.72	-6.32
4HEP	950	-40.38	0.51	-49.12	0.75	-8.74
5EPM	619	-33.00	0.66	-41.86	0.58	-8.86
3L5Y	356	-45.76	0.48	-54.74	0.74	-8.99
5IKC	542	-38.10	0.61	-50.06	0.67	-11.96
4LMQ	953	-31.70	0.77	-45.85	0.72	-14.15
5DFV	418	-39.28	0.55	-55.47	0.63	-16.20
3P9W	582	-39.43	0.59	-59.98	0.67	-20.55
2YBR	204	-39.73	0.61	-64.50	0.70	-24.77
2XT1	154	-41.18	0.56	-66.21	0.75	-25.03
3UZQ	216	-36.87	0.41	-62.79	0.67	-25.93
1JRH	494	-42.27	0.68	-71.58	0.81	-29.31
3KR3	469	-45.61	0.63	-76.40	0.79	-30.79
3NH7	167	-38.63	0.54	-72.02	0.65	-33.38
4UU9	200	-38.58	0.64	-74.67	0.73	-36.09

Table 5.4 Percentage fraction of designed loops that were not mutated

Ag PDB ID	HCDR1	HCDR2	HCDR3	LCDR1	LCDR2	LCDR3
1JRH	95.93	92.68	94.11	90.85	94.51	97.97
1OB1	98.11	94.44	97.22	96.44	97.00	99.78
2BDN	96.97	92.01	93.39	94.21	95.04	98.90
2H9G	97.78	94.78	94.62	94.78	95.89	99.84
2XT1	94.81	86.36	94.81	93.51	96.10	98.70
2YBR	96.08	93.63	97.06	91.18	97.06	98.53
3BDY	98.13	96.85	96.36	94.59	96.65	98.43
3KR3	97.44	93.39	96.16	94.88	95.52	99.36
3L5Y	97.46	96.90	96.90	96.62	96.06	99.44
3NH7	100.00	97.01	97.60	97.60	98.20	98.80
3P30	92.31	92.31	100.00	92.31	100.00	92.31
3P9W	96.56	92.08	93.46	93.29	90.71	99.14
3UZQ	99.07	93.98	94.44	88.89	94.44	99.54
3X3F	97.42	95.57	93.36	93.36	92.25	99.63
4AL8	98.35	97.53	94.23	94.78	92.58	99.45
4DGI	97.26	97.26	94.32	92.56	93.54	100.00
4HEP	98.73	95.46	95.68	97.47	96.10	99.47
4LMQ	95.26	94.84	95.26	93.89	94.42	98.21
4LQF	97.62	94.29	87.62	89.05	94.29	98.10
4N9G	98.05	97.87	95.56	93.78	98.05	98.93
4P2C	96.94	92.76	96.10	91.92	94.99	98.61
4UU9	96.97	94.95	95.45	96.97	96.97	98.99
5DFV	96.86	92.75	94.93	93.72	92.03	98.55
5EPM	96.92	97.89	94.98	96.11	97.57	99.84
5IKC	97.23	95.19	94.27	93.16	92.79	99.45

Table 5.5 Percentage frequencies of different native and final amino acids in mutated residues.

Amino acid	Initial percentage	Final percentage
LEU	1.18	45.33
ASP	1.95	13.84
ASN	3.09	11.09
ALA	0.00	9.58
SER	0.50	9.21
GLU	0.67	5.38
ILE	1.68	3.43
THR	6.18	1.28
VAL	2.22	0.84
HIS	9.85	0.03
ARG	0.97	0.00
CYS	0.07	0.00
GLN	0.60	0.00
LYS	0.87	0.00
MET	0.10	0.00
PHE	7.69	0.00
PRO	0.00	0.00
TRP	11.79	0.00
TYR	50.57	0.00

Table 5.6 Binding energies of top mutated and native AUBIE designs and WT complexes. The differences between these energies and the number of mutations in the best mutated design are also listed.

Ag PDB ID	Native BE (kcal/mol)	WT BE (kcal/mol)	New BE (kcal/mol)	New BE - WT BE (kcal/mol)	New BE - Native BE (kcal/mol)	Mutations
3P9W	-39.43	-59.98	-82.85	-22.88	-43.43	21.00
4LMQ	-31.70	-45.85	-60.09	-14.24	-28.39	29.00
5DFV	-39.28	-55.47	-61.97	-6.50	-22.69	24.00
3L5Y	-45.76	-54.74	-67.77	-13.03	-22.02	18.00
5EPM	-33.00	-41.86	-50.69	-8.83	-17.68	22.00
3NH7	-38.63	-72.02	-53.84	18.18	-15.21	28.00
1JRH	-42.27	-71.58	-53.96	17.62	-11.69	20.00
4UU9	-38.58	-74.67	-46.18	28.50	-7.59	18.00
5IKC	-38.10	-50.06	-41.40	8.66	-3.30	15.00
3UZQ	-36.87	-62.79	-36.87	0	0	0
2YBR	-39.73	-64.50	-39.73	0	0	0
2XT1	-41.18	-66.21	-41.18	0	0	0
3KR3	-45.61	-76.40	-45.61	0	0	0

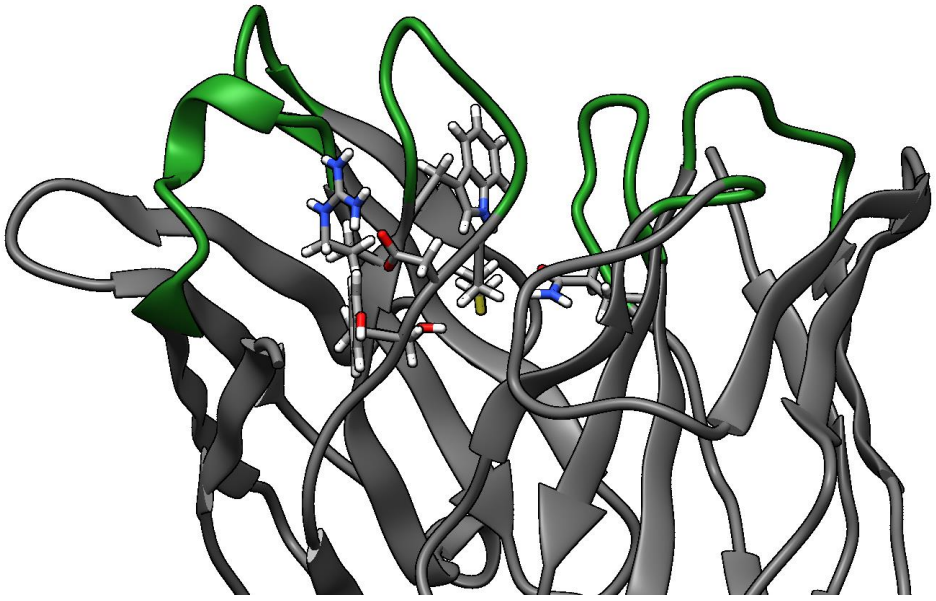

Figure 5.1 Binding regions (green) and framework (gray) structure used for database generation step. Also shown are the CDR3 loop side chain atoms that would have been classified as binding regions if IMGT definitions of CDR3 loops were used. These side chain atoms were classified to be significant for antibody stability and hence were not mutated. Structures obtained from PDB 1N8Z.

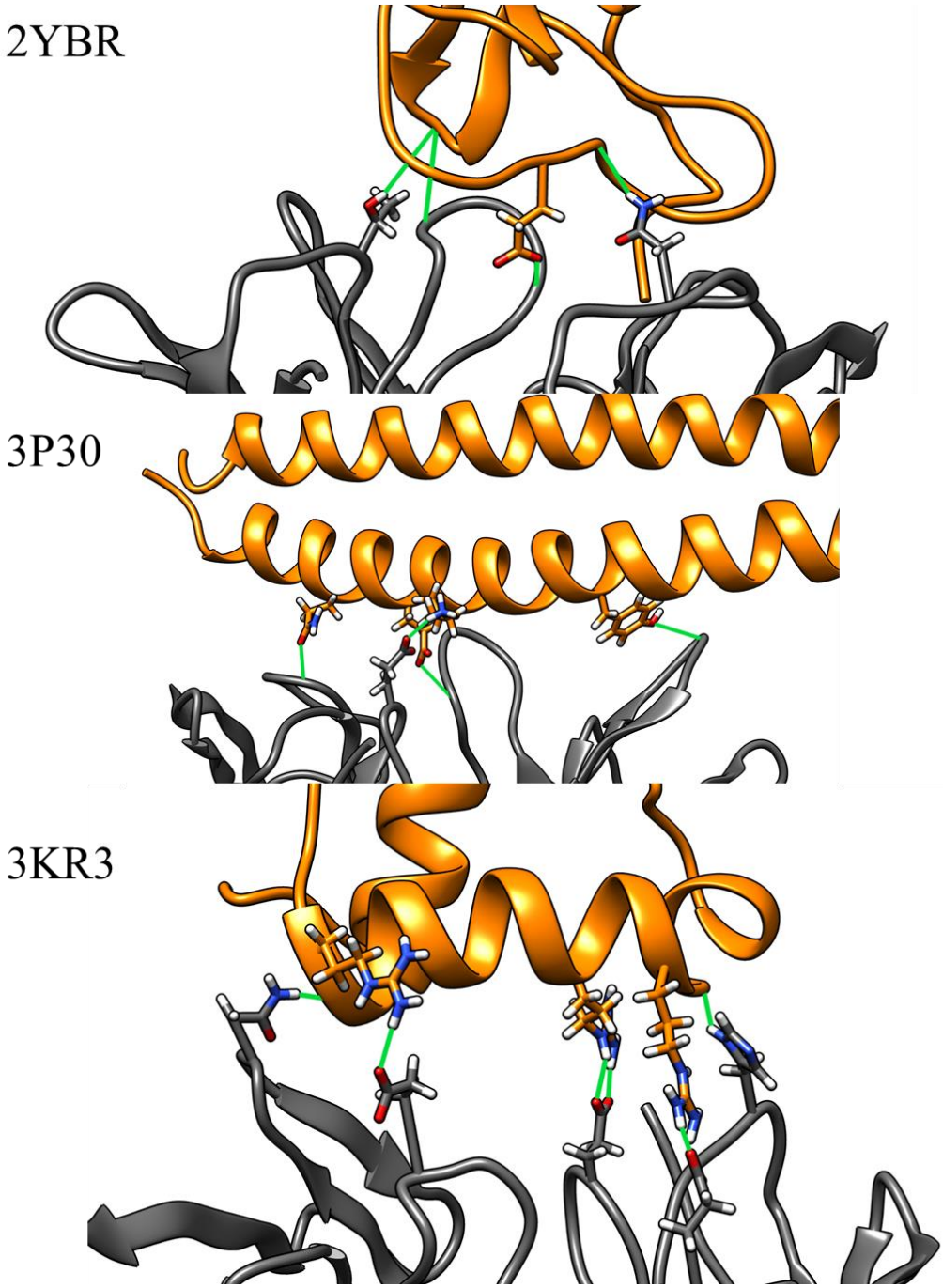

Figure 5.2 Designed low entropy H-bonds in top AUBIE designs for antigens from PDB files 2YBR, 3P30 and 3KR3 before energy minimizations. Antibodies and antigens shown in gray and orange respectively. H-bonds shown in green.

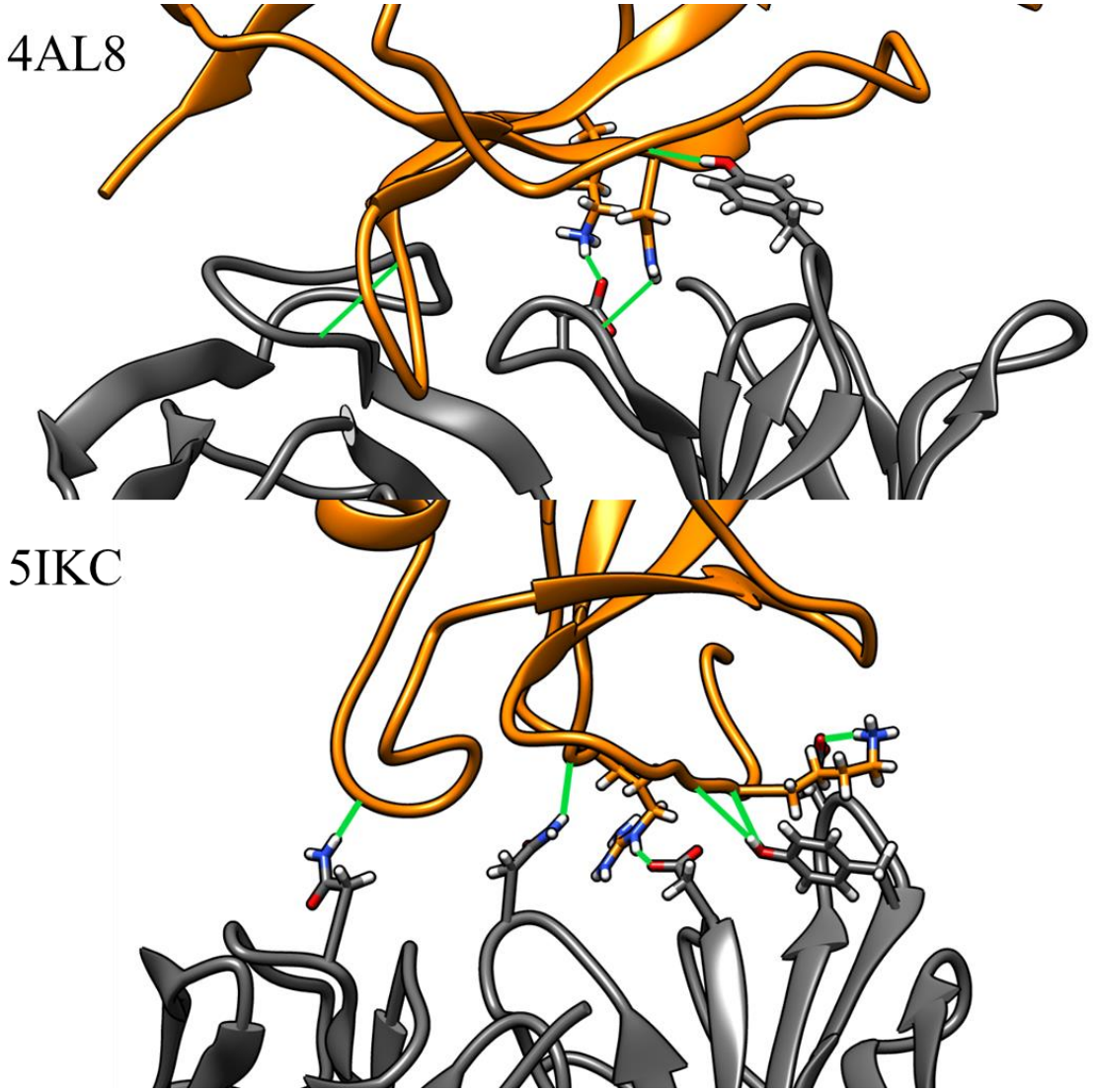

Figure 5.3 Designed low entropy H-bonds in top AUBIE designs for antigens from PDB files 2YBR, 3P30 and 3KR3 before energy minimizations. Antibodies and antigens shown in gray and orange respectively. H-bonds shown in green.

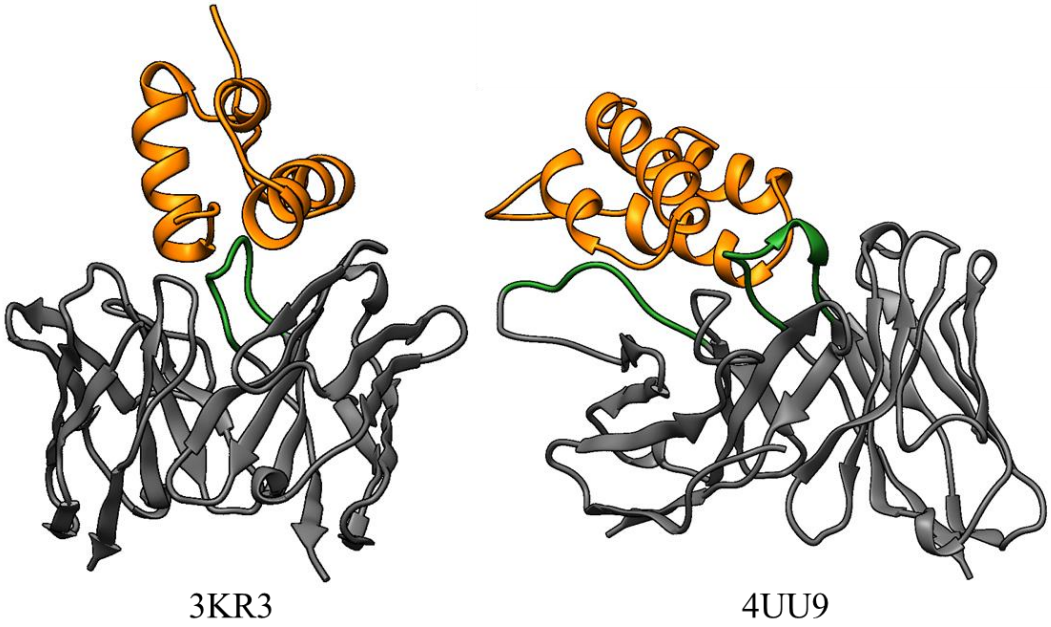

Figure 5.4 Antibody (gray) and antigen (orange) complexes from PDB files 3KR3 and 4UU9. Long CDR3 loops and interacting framework regions shown in green.

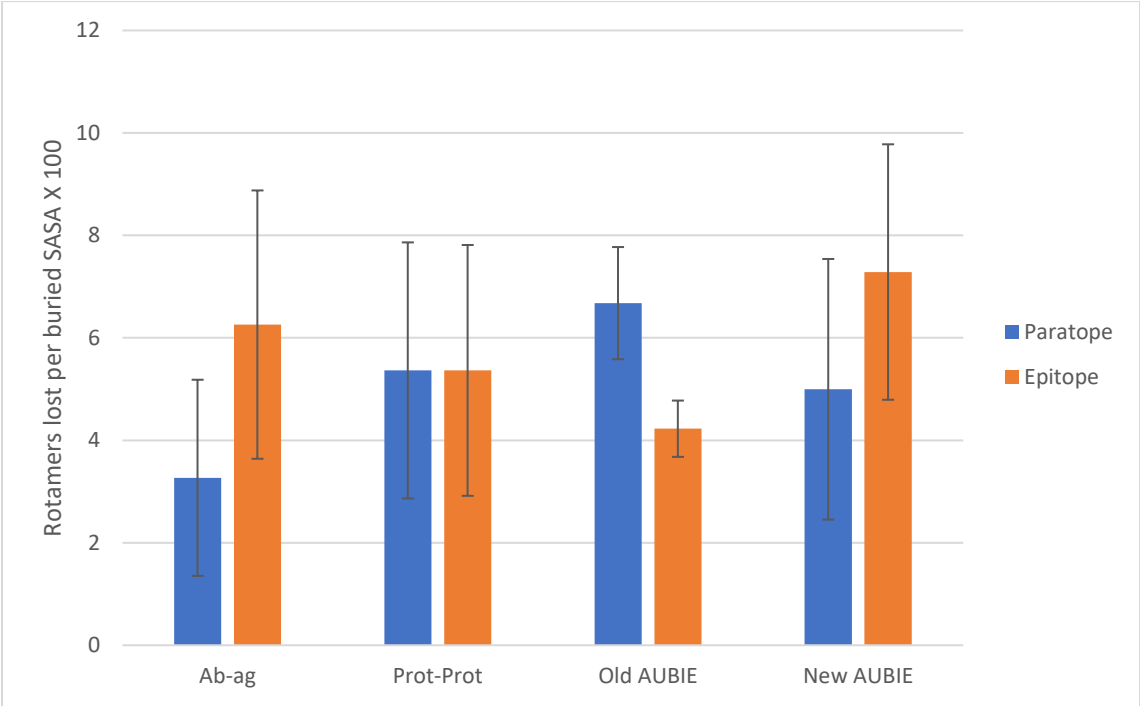

Figure 5.5 Averages and standard deviations of RL-SASAs of epitopes and paratopes of antibody-antigen, protein-protein databases, top designs from old and new AUBIE approach.

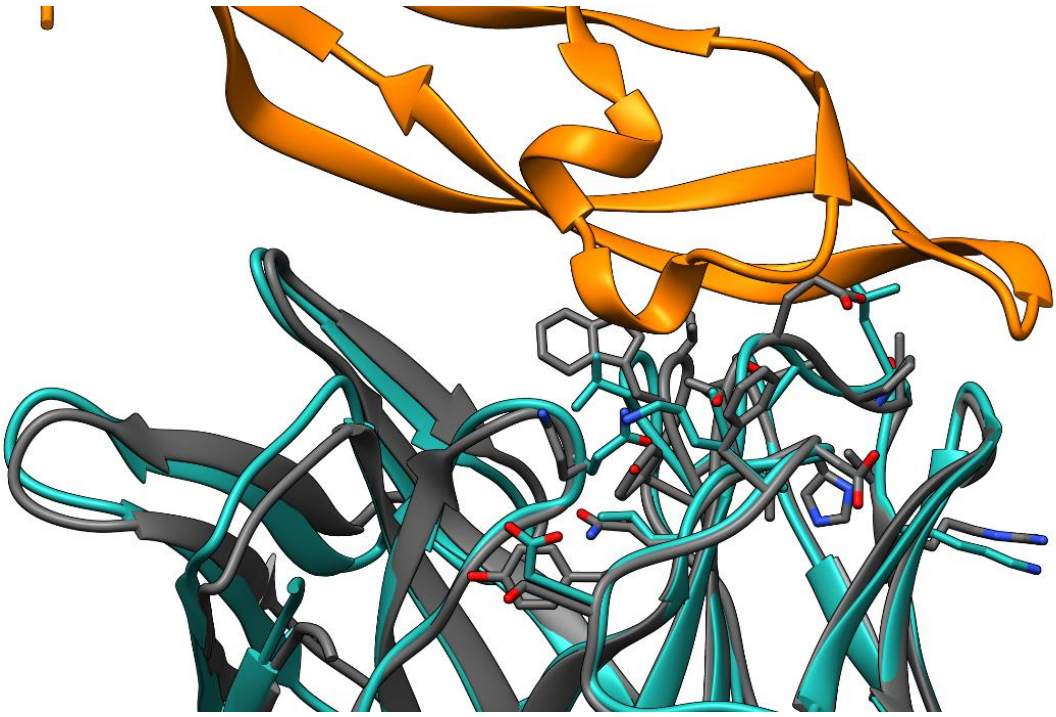

Figure 5.6 Native and mutated side chains from AUBIE design binding antigen from PDB 3P9W.

Native, mutated antibodies and antigens shown in gray, sea green and orange respectively.

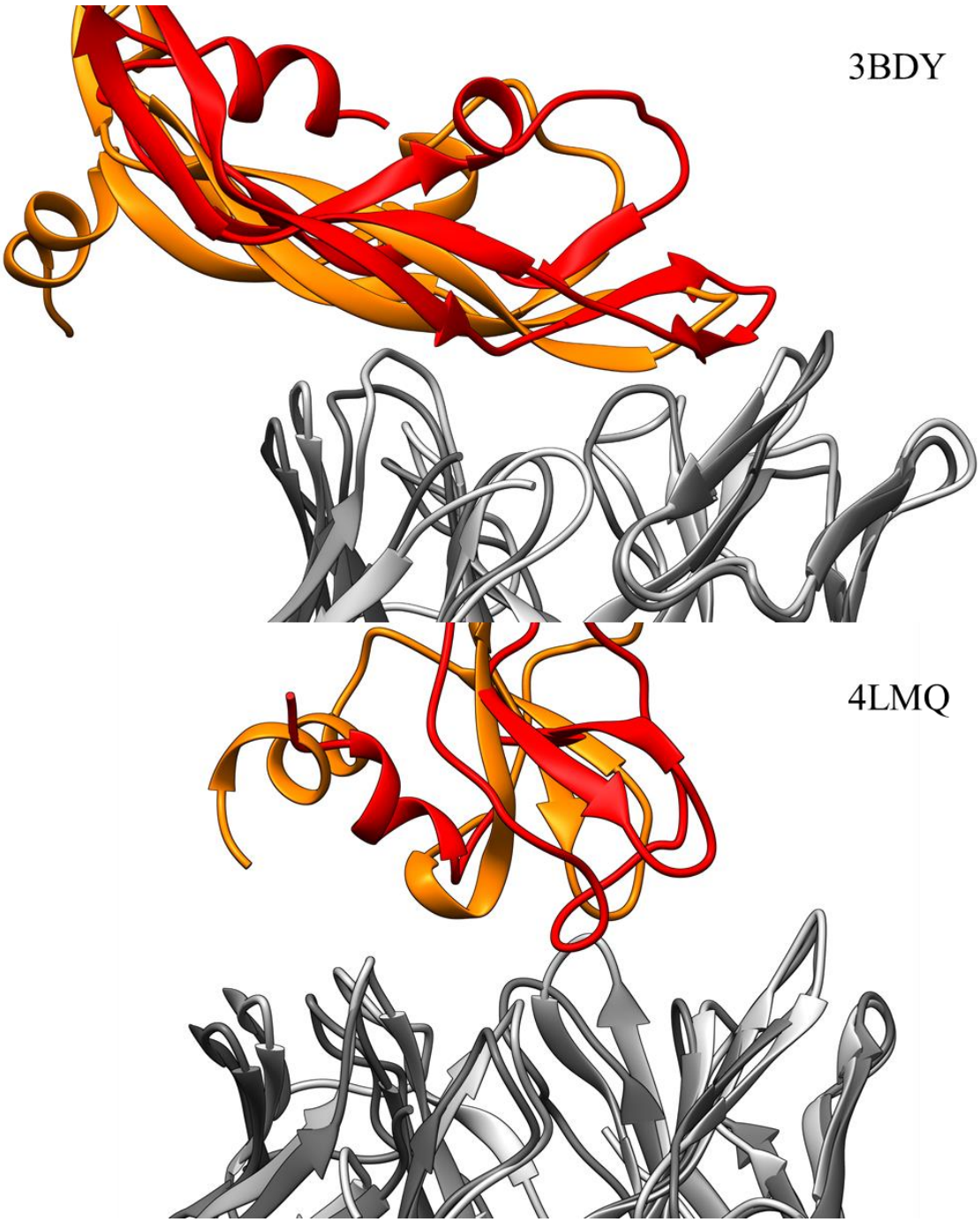

Figure 5.7 Initial and final antibody-antigen complexes from MD simulations of top AUBIE designs binding to antigens from PDB 3BDY and 4LMQ. Starting antibodies and antigens shown in dark gray and orange, while final antibodies and antigens shown in light gray and red respectively.

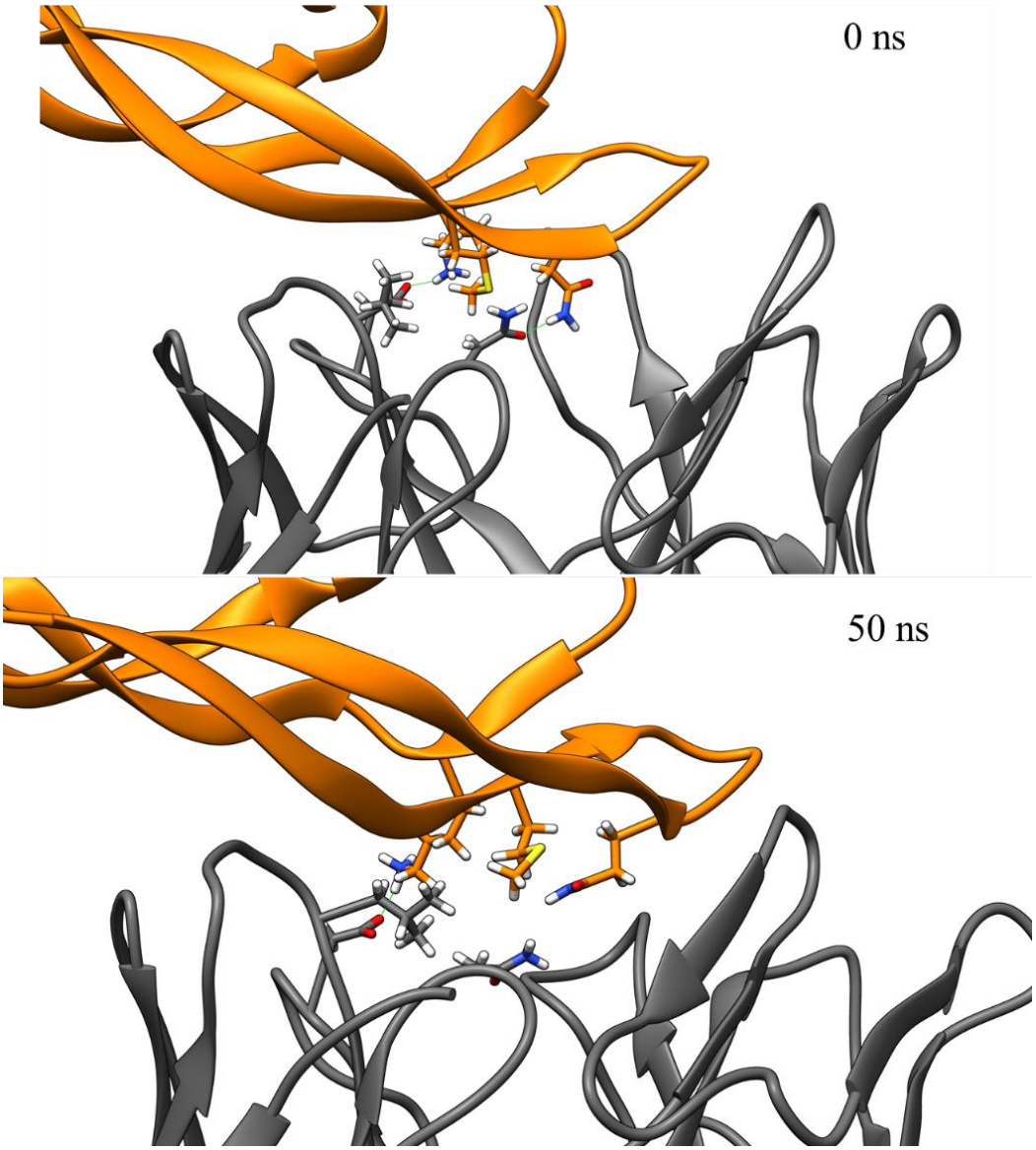

Figure 5.8 Initial and final antibody-antigen complexes from MD simulations of top AUBIE design binding to antigen from PDB 3BDY. Side chain atoms of interacting residues 56D, 38L and 107N from the antibody and 35K, 68M and 76Q from the antigen. Starting antibodies and antigens shown in dark gray and orange respectively. H-bonds shown in green.

References

[1] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” in *Computer Aided Chemical Engineering*, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds. Elsevier, 2018, pp. 2113–2118. doi: https://doi.org/10.1016/B978-0-444-64241-7.50347-5.

[2] B. R. Brooks *et al.*, “CHARMM: the biomolecular simulation program,” *J Comput Chem*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[3] R. F. Alford *et al.*, “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design,” *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, Jun. 2017, doi: 10.1021/acs.jctc.7b00125.

[4] P. B. Stranges and B. Kuhlman, “A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds,” *Protein Sci*, vol. 22, no. 1, pp. 74–82, Jan. 2013, doi: 10.1002/pro.2187.

[5] A. K. Mishra and R. A. Mariuzza, “Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing ,” *Frontiers in Immunology* , vol. 9. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2018.00117

[6] J. Adolf-Bryfogle *et al.*, “RosettaAntibodyDesign (RAbD): A general framework for computational antibody design,” *PLOS Computational Biology*, vol. 14, no. 4, pp. e1006112-, Apr. 2018, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006112

[7] R. Chowdhury, M. F. Allan, and C. D. Maranas, “OptMAVEn-2.0: De novo Design of Variable Antibody Regions against Targeted Antigen Epitopes,” *Antibodies (Basel)*, vol. 7, no. 3, p. 23, Jun. 2018, doi: 10.3390/antib7030023.

[8] G. F. Mangiatordi, D. Alberga, L. Siragusa, L. Goracci, G. Lattanzi, and O. Nicolotti, "Challenging AQP4 druggability for NMO-IgG antibody binding using molecular dynamics and molecular interaction fields," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1848, no. 7, pp. 1462–1471, 2015, doi: https://doi.org/10.1016/j.bbamem.2015.03.019.

[9] G. D. Lapidoth *et al.*, "AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences," *Proteins*, vol. 83, no. 8, pp. 1385–1406, Aug. 2015, doi: 10.1002/prot.24779.

[10] S. J. Fleishman, J. E. Corn, E. M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker, "Hotspot-centric de novo design of protein binders," *Journal of Molecular Biology*, vol. 413, no. 5, pp. 1047–1062, 2011, doi: 10.1016/j.jmb.2011.09.001.

6. Chapter 6 - MutDock

The results from Chapter 5 demonstrated that targeting for low entropy H-bonds in the AUBIE approach resulted in antibody designs that had acceptable metrics. I now turned my attention towards developing a docking approach for fixed-backbone scaffolds that built on the same design principles AUBIE was based on. In this chapter, the algorithm and initial evaluation of MutDock, a novel mutation-based docking approach, is described. Instead of translating and rotating a protein scaffold in 6D steps around the target, MutDock uses pairwise distance matching of H-bonding regions around the variable paratope and epitope to identify mutated scaffold-target poses making multiple H-bonds in a single step. Parts of this chapter have been adopted from our article titled "MutDock: A Computational Docking Approach for Fixed-Backbone Protein Scaffold Design", which is under review.

Methodology

MutDock has been developed around two major goals: 1) introducing mutations simultaneously to docking using a single geometry alignment step and 2) designing binding proteins without the use of a force field. Although they are constantly being updated, there is evidence that force fields have biases towards and against certain types of interactions and structural elements [1]–[3]. Hence, instead of using force field guided energy minimizations like conventional docking tools such as RosettaDock and HADDOCK, MutDock uses the known structural features of strong interactions such as H-bonds to guide pose identification and rotamer selection during docking. Similar to ZDOCK, where the generated poses are scored in the final step, in this manuscript force fields are used to evaluate MutDock's predicted poses but not in their identification.

The MutDock approach can be divided into two primary steps: 1) identifying docked poses with H-bonds formed by native and/or mutated residues and 2) mutating clashing side chains using a feature-based approach. This algorithm is depicted in Figure 6.1, with panels A, B, and C corresponding to step 1 and panel D corresponding to step 2.

Docking

Docking with MutDock can be further subdivided into two steps: pose identification and pose validation. The pose identification step generates antigen-scaffold poses with unique sets of H-bonds, while the pose validation step checks for steric compatibility, bond formation and solution uniqueness. Necessary information for docking includes the structures of the binding and target proteins, which residues in the binding protein are intended to interact with the target protein (i.e., the paratope), which paratope residues are mutable, and which residues in the target protein are intended to interact with the paratope (i.e., the epitope).

In the first step of pose identification, spatial coordinates that can be occupied by compatible epitope atoms for the formation of a H-bond are identified for each paratope residue. These coordinates for binding interactions, referred to as Paratope Binding Regions (PBRs), consist of one atom and three spatial positions. The atom is either a hydrogen, if the interaction is an H-bond forming from the paratope to the epitope, or an H-bond acceptor, if the H-bond is forming from the epitope to the paratope. The primary point of the PBR is the ideal position of an atom in the epitope to form an H-bond with the PBR's atom. The position of the primary point lies 1.5 Å from the PBR's atom on the vector determined by the atom and an antecedent point. The possible PBR atoms and corresponding antecedent points are listed in Table 6.1. The third spatial position of a PBR is the secondary point, 1 Å further along the vector from the primary point,

which is used for ensuring the designed H-bonds have appropriate orientations (e.g., avoiding the formation of H-bonds with acute angles).

The pose identification step of MutDock is focused on finding complexes with many H-bonds because they are prevalent in naturally occurring binding interfaces. As listed in Table 6.2, the mutable paratope residues are only allowed to change into polar residues to facilitate this search for favorable H-bonds. The rotamers used in MutDock were obtained from the Dunbrack rotamer library [4] and a maximum of ten structurally diverse rotamers for each possible mutation were used to limit final solution set diversity.

Once the PBRs are identified, the epitope residues are scanned for polar atoms capable of forming H-bonds with the scaffold in MutDock's second step of pose identification. Initially, epitope residues that have non-zero solvent accessible surface areas (SASA) are collected to avoid including buried atoms in subsequent calculations. For each solvent accessible epitope residue, Epitope Binding Atoms (EBAs) are identified. An EBA consists of two points: a primary point that is a polar hydrogen atom or an H-bond acceptor and a secondary point that is analogous to the antecedent points of PBRs.

Figure 6.2 illustrates the values that are calculated in the third step of pose identification. For each pair of PBRs, one distance and four angles are calculated. The distance is the distance between the primary points of the PBRs, while the angles are those of the quadrilateral formed by the primary and secondary points. Similar calculations are carried out for each pair of EBAs. Pairs of PBRs or EBAs that belong to the same residue are not considered in this step to encourage formation of larger binding interfaces by MutDock. A maximum limit of one positively charged residue in the paratope was enforced in step 4 to reduce the presence of highly flexible and unstable

positively charged side chains in the interface. Thus, only one ARG/LYS residue per PBR couplet is allowed.

MutDock's fourth step of pose identification searches for PBR and EBA sets that can coexist simultaneously. This begins by finding pairs of compatible PBR and EBA couplets. An initial screen eliminates from consideration incompatible interactions. A readily evident example would be H-bonds between ARG and ARG side chains, but many other potential interactions are also excluded from consideration. It is known that antibody binding interfaces are abundant with interactions made by pre-stabilized or low entropy side chains [5]. To replicate such features, H-bond type constraints are enforced to lower the chances of forming H-bonds between long chain amino acids which are unlikely to be stable. The atom and amino acid types allowed for a compatible PBR-EPA pair are listed in Table 6.2.

Interactions that are potentially compatible are then checked for geometric alignment using five constraints:

$$|D1 - D2| < d_{limit} \tag{1}$$

Where D1 and D2 are the PBR and EBA distances, respectively, as calculated in step 3 and d_{limit} is a user-defined threshold on the maximum permissible deviation in the primary point distances.

$$|\angle PBR1p - \angle EBA1p| < a_{limit} \tag{2}$$

$$|\angle PBR2p - \angle EBA2p| < a_{limit} \tag{3}$$

$$|\angle PBR1s - \angle EBA1a| < a_{limit} \tag{4}$$

$$|\angle PBR2s - \angle EBA2a| < a_{limit} \tag{5}$$

Geometric constraints 2-5 ensure the deviations in the interaction angles do not exceed a user-defined limit, a_{limit}. For this study, d_{limit} was set at 1.8 Å and a_{limit} was set at 70°. These values

were selected as the cutoffs because they permit 85% of H-bonds from the database from Chapter 2. An illustration of these constraints is shown in Figure 6.2. After compatible PBR-EBA couples are found, larger sets of PBR-EBA matches are identified by searching for groups of couples that are all mutually compatible (e.g., if AB, AC, and BC are each compatible couples, than ABC must be a compatible triple). Thus each group corresponds to a unique solution consisting of three or more predicted interface H-bonds.

In the fifth and final step of MutDock's docking protocol, each unique solution group from step 4 is positioned and analyzed for interface size, hydrogen bond geometry and steric compatibility with the scaffold and antigen. The antigen is positioned so that the RMSD between the primary atoms of the EBA and their corresponding PBR primary points is minimized. Next, the rotamers of the binding protein are changed to match those used in the PBRs, corresponding to the inclusion of any mutations identified during pose identification. These mutations are referred to as design mutations. If these newly placed rotamers have steric clashes with other PBR rotamers or native side chains, the pose is rejected. The definition of a steric clash was adopted from Chapter 3.

A coarse-grained filter is used to facilitate the rejection of low quality poses (i.e., those with small buried surface areas or major clashes between proteins). Each residue is divided into its backbone and side chain units. Each such unit is approximated as a sphere at the center of mass of the atoms with a radius equal to the distance of the farthest atom from the center of mass. Analysis of the antibody-antigen database from Chapter 2 revealed that the coarse-grained spheres should be a minimum of 3.39 Å apart and that the complexes should have a minimum of 12 spheres in contact with one another. Poses that violate either of these requirements are rejected as having irreconcilable steric clashes or too small of interface surface areas, respectively.

After the coarse-grained filter removes obviously deficient poses, an all atom pose validation is conducted. Each designed H-bond is checked, and they are accepted if the acceptor – hydrogen distance is less than 2.5 Å and the acceptor – hydrogen – donor angle is larger than 120 degrees. Poses that fail to form even one of their predicted H-bonds are rejected. Finally, the steric compatibility between the antigen and scaffold is verified. Poses that consist of steric clashes between the antigen and the binding protein's backbone and/or non-variable side chains are discarded, as are those that have clashes involving the residues forming the designed H-bonds. However, poses with clashes involving mutable paratope residues that do not form designed H-bonds are retained, as those residues can be changed in the second major step of MutDock: mutation. The mutational approach from Chapter 5 was employed here to resolve steric clashes. Following the mutations, the clash-free mutated poses are output in PDB format. In the current version of MutDock, no further pose refinement or ranking is performed and rotamer repacking/energy minimization and pose ranking strategy is left to the user.

Complex Evaluation

Force fields were used to evaluate the MutDock predictions and compare them to other docking methods, even though they were not used for energy calculations during the MutDock algorithm. All MutDock poses were relaxed through two energy minimization runs: CHARMM36 force field vacuum energy minimization [6] with fixed backbone atoms followed by 2) Rosetta force field energy minimization [7]. When docked poses are identified, they include minor steric clashes. CHARMM36 energy minimizations were able to consistently correct the clashes caused by the steric constraints used in the method, while Rosetta could only sometimes correct them. Thus, CHARMM36 was used to prepare the complexes for computational analysis while Rosetta, the most commonly used protein engineering force field, was used for comparing the quality of

poses. Complexes from the previously mentioned antibody-antigen database were run through the same energy minimization routine as the MutDock poses. The key metric analyzed in this work is binding energy, which is the difference in the Gibbs free energy of the system before and after complex formation, since the primary purpose of MutDock is to identify high affinity mutated poses. Along with computational binding energy, other binding metrics analyzed in this work such as shape complementarity and buried interface area have been calculated by the InterfaceAnalyzer application from Rosetta [8].

MutDock is compared to ZDOCK and HADDOCK. To compare the epitope specific/local docking of MutDock to that of ZDOCK, residues far from the epitope and paratope were manually selected to be blocked from being part of ZDOCK pose interfaces. The HADDOCK webserver was used for HADDOCK docking simulations and epitopes and paratopes were defined as the "active regions" on the webserver. For each run, all the 2000 ZDOCK poses and 200 HADDOCK poses from the "it1" directory were run through the energy minimization routine. All protein visualization and image generation were done via UCSF Chimera [9].

Results

MutDock's performance was tested by docking 10 randomly selected antigens from the antibody-antigen database with two scaffolds: affibody and DARPin. The affibody and DARPin structures were obtained from PDB files 3MZW and 6FPA, respectively. The identities of the 13 and 18 variable residues for affibodies and DARPins were obtained from literature [10], [11]. The paratope residues included the variable residues, as well as several surrounding residues. The epitopes from the native antibody-antigen complexes were selected as the epitopes for the docking runs. The source PDB IDs of the 10 antigen structures, residue numbers of the paratope, epitope

and variable residues are listed in Table 6.3. The docking runtimes ranged from 1 to 13 hours depending on the number of preliminary poses that needed to be filtered for various metrics. These high runtimes were expected, as the current implementation of MutDock is intended as a proof of principle method and its code has not been optimized for computational efficiency.

The predicted binding energies for the top poses of the 20 complexes are reported in Table 6.3, along with those of the native antibody-antigen complexes. Greater than 1000 poses were identified for all antigens except PDB 3P30. That antigen consists of a two-helix bundle and hence lacks solvent exposed backbone atoms for binding. Approximately 500 poses were identified for this antigen with both the affibody and DARPin scaffolds. The binding energies of top MutDock poses ranged from 35 to 51 kcal/mol. For eight of the ten antigens, the DARPin poses had stronger binding energies than the affibody poses. This is consistent with the facts that DARPins have larger paratopes than affibodies and our prior experience that computationally calculated binding energies are strongly correlated with interface size. The stronger calculated binding energies of the antibodies versus the MutDock designed proteins is also consistent with this trend.

Table 6.5 lists the percentage frequencies of poses with different numbers of mutations. It is observed that the docking approach relies heavily upon design mutations for identifying poses, as more than 90% of all poses consist of either two or three design mutations. In contrast, less than 50% of all poses had any clash mutations, with a majority of them having only one mutation. Considering that all variable side chains were allowed to clash before the mutation step, this result signifies that the rotamer repacking step (i.e., step 1 of 6 of the clash mutation calculations) was efficient at resolving most side chain clashes.

Table 6.6 lists the percentage frequency of the amino acid types in design and clash mutations. The data demonstrates expected trends, as MutDock favors the introduction of low

entropy side chains, such as SER and TYR, that can form H-bonds with backbone atoms and other low entropy side chains. In contrast, the positively charged, high entropy side chains of ARG and LYS are disfavored. The most favored clash mutation was LEU. Examples of design and clash mutations are illustrated in Figure 6.3.

The widely used docking programs ZDOCK and HADDOCK were tested on the same affibody and DARPin scaffolds against the same antigens. ZDOCK was selected for comparison with MutDock since neither method uses local pose refinement or rigid body energy minimizations [12]. ZDOCK run times lasted for 4 minutes while HADDOCK webservers took a maximum of approximately 6 hours, which also includes the time when the job was queued, for each docking run. On the other hand, MutDock runtimes ranged from 3 to 46 hours on a 3.00 GHz Intel Xeon Gold 6248R processor.

The top binding energies for each complex along with the difference with the best MutDock pose energies are listed in Table 6.7. MutDock predicted poses with binding energies at least 3 kcal / mol stronger than ZDOCK in 17 of 20 complexes and at least 10 kcal / mol stronger in 11 complexes. The only antigen ZDOCK outperformed MutDock on was 3P30, whose helical nature eliminates the possibility of the backbone H-bonds that MutDock preferentially targets.

MutDock was also compared to HADDOCK [13], [14], a docking approach that performs local rigid body energy minimizations along with further refinement using short MD simulations. The top binding energies for each complex along with the difference with the best MutDock pose energies are listed in Table 6.8. MutDock predicted better poses for nine complexes, similar quality poses (i.e., ± 3 kcal / mol) in five complexes, and worse poses in six complexes. Prior experience had demonstrated that multiple Rosetta energy minimizations with the same complex resulted in binding energies within 2 kcal/mol of each other. Thus, 3 kcal/mol was defined as the threshold

for significant binding energy difference. In the nine complexes where MutDock outperformed HADDOCK, the improvement in binding energies were lower than those obtained from the ZDOCK comparisons. Thus, HADDOCK predicted complexes with stronger binding energies than ZDOCK.

A likely cause of HADDOCK's performance being evaluated well by energy calculations is its use of such calculations to refine initial poses. To investigate whether the positional refinements HADDOCK utilizes could further improve the MutDock poses, the top MutDock designs were docked using HADDOCK. After HADDOCK docking, poses were defined as near native if they had an interface Cα RMSD less than 4 Å when compared with the respective top MutDock pose.

Table 6.8 lists the top binding energies of near-native poses, their RMSDs and the respective binding energy differences. For nine complexes, HADDOCK identified novel poses with better binding energies than the native MutDock poses. Of those nine complexes, HADDOCK identified: no near native pose for two complexes, better (i.e., by at least 3 kcal / mol) poses for three complexes, comparable poses for two complexes and worse than native poses for two complexes. For the remaining eleven complexes, HADDOCK identified: better poses for two complexes, comparable poses for six complexes, and worse poses for three complexes. Finally, the top HADDOCK poses using the MutDock designed binding proteins had stronger calculated binding energies for 16 / 20 complexes than the top HADDOCK poses with the original scaffolds.

Discussion

It is particularly notable that more than 90% of the poses MutDock identified required at least two mutations. This shows that MutDock has the ability to generate large number of poses

with multiple beneficial mutations per pose. Conventional design approaches search for beneficial mutations through iterative cycles of random point mutations [15], [16]. On the other hand, MutDock is able to simultaneously identify multiple beneficial mutations per pose in a single search step with further beneficial mutations added in the clash-removal step. Such an approach allows MutDock to search a larger solution sequence space and hence identify poses that would not be identified by fixed-sequence docking methods.

Using Rosetta-calculated binding energies as a benchmark, MutDock performed significantly better than ZDOCK for most of the complexes. Each of these methods relies on geometric criteria for identifying binding poses: H-bond formation for MutDock and shape complementarity along with molecular mechanics for ZDOCK. This is in line with expectations, as the mutations introduced by MutDock should result in improved binding energies relative to those attainable by the original scaffold. Nonetheless, this demonstrates that strictly using geometric criteria MutDock is able to identify favorable and promising binding conformations.

The comparison of the performances of MutDock and HADDOCK is more nuanced. Each did best on approximately half of the complexes in a head to head comparison. This is due in part to the fact that each has an advantage over the other: MutDock allows for mutation of the scaffold, while HADDOCK uses energy minimizations and positional refinement to maximize predicted binding energies. However, 16 / 20 complexes were improved when HADDOCK used the MutDock designed scaffolds compared to when it used the original ones, albeit not always in conformations similar to MutDock's predicted poses. This indicates that the MutDock's predicted mutations, which are unguided by energy calculations, create the potential to improve binding energies. It is notable that in a number of complexes, HADDOCK was unable to identify poses with binding energies as strong as MutDock's predictions. Given that it is demonstrable that those

poses exist, this indicates that HADDOCK's energy-based pose identification algorithm still has potential room for improvement.

Force fields are the best available tool for computationally assessing designed proteins, but they have limitations and biases. One of the primary motivations of MutDock was to explore the design of binding proteins without the use of a force field in the design decisions. Despite only using force fields to evaluate the designs and not in their generation, MutDock was able to generate poses with computational metrics comparable to known binding complexes and poses made with force field -dependent docking tools. Thus, MutDock serves as an example of a viable docking-design approach that attempts to replicate known beneficial features of binding interfaces, such as hotspot interactions in hotspot-centric design [17]. The development of such methods has been made possible by the availability of large datasets of known complex structures that be analyzed for common key structural features which can later be targeted.

The only other interaction-based docking approach I have seen in the literature is RIFdock [18], [19]. Compared to MutDock, RIFdock uses a larger library of rotamer poses and includes hydrophobic interactions as target interactions too. A major difference between the two approaches is the search strategy. RIFdock moves the receptor protein in 6D steps, with increasing resolution, around the target protein to find scaffolds poses that host multiple rotamers which make strong interactions with the epitope. On the other hand, MutDock uses pairwise distance alignment to identify groups of compatible interactions in one step. However, a more thorough comparison of the approaches was not possible since the detailed description of the RIFdock methodology is not available in a peer-reviewed article.

The recent breakthroughs of AlphaFold and RosettaFold at predicting protein structures without heavy reliance on physics-based force fields herald a change in computational protein

engineering. I believe that one of the frontiers of bioengineering will be the growth of computational protein design methods that use machine learning and engineering principles instead of force fields. MutDock demonstrates the potential of such approaches. Through spatial positioning and mutation steps, MutDock is able to identify poses that have many low entropy, favorable interactions. The results, especially those from re-docking the scaffolds with HADDOCK, indicate that the binding energies can be improved without relying on force field calculations.

Table 6.1 Polar atom and their antecedent atom names considered for PBR and EBA identification. Atom names follow the CHARMM PDB atom naming.

Amino Acid	PBR Atom	Antecedent Point
Backbone	O	C
Backbone	OT1	C
Backbone	OT2	C
Backbone	HN	N
Backbone	HN1	N
Backbone	HN2	N
Backbone	HN3	N
ARG	HE	NE
ARG	HH11	NH1
ARG	HH12	NH1
ARG	HH21	NH2
ARG	HH22	NH2
LYS	HZ1	NZ
LYS	HZ2	NZ
LYS	HZ3	NZ
ASP	OD1	CG
ASP	OD2	CG
GLU	OE1	CD
GLU	OE2	CD
SER	HG1	OG
SER	OG	Midpoint of HG1 and CB
THR	HG1	OG1
THR	OG1	Midpoint of HG1 and CB
TYR	HH	OH
TYR	OH	Midpoint of HH and CZ
ASN	OD1	CG
ASN	HD21	ND2
ASN	HD22	ND2
GLN	OE1	CD
GLN	HE21	NE2
GLN	HE22	NE2
HIS	HD1	ND1
HIS	NE2	Midpoint of CD2 and CE1
TRP	HE1	NE1

Table 6.2 Amino acid types allowed to form H-bonds in the MutDock approach.

Scaffold	Antigen
Backbone	Backbone
Backbone	All
All except ARG, LYS, GLN	Backbone
ARG, LYS *	ASP, GLU
ASP, GLU	ARG, LYS
HIS, ASN, ASP, TYR, SER, THR, TRP **	HIS, ASN, ASP, TYR, SER, THR, TRP

* Only one per solution

** No H-bonds allowed between ASN and ASP since they contain multiple polar groups

Table 6.3 Residue numbers of “active binding regions” in HADDOCK simulations. Residue numbers of epitope, paratope and variable paratope residue numbers for MutDock runs are also provided. For 4AL8-dp and 1OB1-dp MutDock simulations, the number of epitope residues were reduced to increase runtimes.

Complex	Residues
	HADDOCK
1JRH	37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 66, 67, 68, 69, 70, 72, 74, 88, 89, 90, 91, 93
1OB1	1, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 23, 24, 25, 26, 28, 31, 36, 38, 39, 40
2XT1	1, 2, 3, 4, 5, 6, 7, 9, 26, 30, 31, 32, 33, 35, 36, 39, 40, 42, 43, 44, 47, 49
3BDY	31, 32, 35, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 81
3L5Y	1, 2, 3, 6, 7, 9, 10, 13, 74, 75, 76, 78, 79, 81, 82, 83, 85
3P30	52, 55, 56, 59, 60, 62, 63, 64, 66, 67, 69, 70, 71, 73, 74, 77
3X3F	14, 15, 16, 17, 18, 19, 28, 29, 33, 35, 36, 37, 38, 39, 40, 41, 42
4AL8	13, 14, 16, 17, 18, 19, 20, 21, 22, 24, 26, 52, 54, 55, 56, 57, 58, 66, 67, 68, 69, 71
5DFV	13, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 69
5IKC	9, 10, 11, 12, 13, 14, 15, 16, 43, 44, 45, 46, 47, 48, 80, 82, 83, 84, 85, 87
DARPin	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148
Affibody	6, 7, 9, 10, 11, 13, 14, 15, 16, 17, 18, 23, 24, 25, 27, 28, 29, 31, 32, 34, 35, 36, 37
	MutDock
1JRH_aff	37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 66, 67, 68, 69, 70, 72, 74, 88, 89, 90, 91, 93
1JRH_dp	37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 66, 67, 68, 69, 70, 72, 74, 88, 89, 90, 91, 93
1OB1_aff	1, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 23, 24, 25, 26, 28, 31, 36, 38, 39, 40
1OB1_dp	8, 9, 10, 11, 13, 14, 15, 16, 23, 24, 25, 26, 28, 36, 38, 39, 40
2XT1_aff	1, 2, 3, 4, 5, 6, 7, 9, 26, 30, 31, 32, 33, 35, 36, 39, 40, 42, 43, 44, 47, 48
2XT1_dp	1, 2, 3, 4, 5, 6, 7, 9, 26, 30, 31, 32, 33, 35, 36, 39, 40, 42, 43, 44, 47, 48
3BDY_aff	31, 32, 35, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80
3BDY_dp	31, 32, 35, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80
3L5Y_aff	1, 2, 3, 6, 7, 9, 10, 13, 74, 75, 76, 78, 79, 81, 82, 83, 84
3L5Y_dp	1, 2, 3, 6, 7, 9, 10, 13, 74, 75, 76, 78, 79, 81, 82, 83, 84
3P30_aff	55, 56, 59, 60, 62, 63, 64, 66, 67, 69, 70, 71, 73, 74, 77

3P30_dp	55, 56, 59, 60, 62, 63, 64, 66, 67, 69, 70, 71, 73, 74, 77
3X3F_aff	6, 14, 16, 17, 18, 19, 29, 33, 35, 36, 37, 38, 39, 40, 41, 42
3X3F_dp	6, 14, 16, 17, 18, 19, 29, 33, 35, 36, 37, 38, 39, 40, 41, 42
4AL8_aff	13, 14, 16, 17, 18, 19, 20, 21, 22, 24, 26, 52, 54, 55, 56, 57, 58, 66, 67, 68, 69, 71
4AL8_dp	13, 14, 17, 18, 19, 20, 21, 22, 26, 52, 54, 56, 66, 67, 68, 69, 71
5DFV_aff	13, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 68
5DFV_dp	13, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 49, 50, 51, 52, 53, 55, 56, 57, 58, 59, 60, 68
5IKC_aff	9, 10, 11, 12, 13, 14, 15, 16, 43, 44, 45, 46, 47, 48, 80, 82, 83, 84, 85, 86
5IKC_dp	9, 10, 11, 12, 13, 14, 15, 16, 43, 44, 45, 46, 47, 48, 80, 82, 83, 84, 85, 86
DARPin	6, 7, 10, 14, 15, 16, 33, 34, 35, 36, 37, 38, 43, 46, 47, 66, 68, 69, 70, 71, 79, 80, 99, 101, 102, 103, 104, 112, 113, 132, 134, 135, 136, 137, 145, 146, 147
Affibody	6, 9, 10, 11, 13, 14, 15, 17, 18, 24, 25, 27, 28, 31, 32, 35, 36, 37, 38

	MuDock variable residues
Affibody	9, 10, 11, 13, 14, 17, 18, 24, 25, 27, 28, 32, 35
DARPin	33, 35, 36, 38, 46, 47, 66, 68, 69, 71, 79, 80, 99, 101, 102, 104, 112, 113

Table 6.4 Top binding energies (BE) of MutDock poses and their native wild type structures for the 20 antigen-scaffold complexes. The shape complementarity values of these top MutDock poses and the native structures are also listed.

Antigen PDB	Scaffold	Top MutDock BE (kcal/mol)	Sc of top MutDock pose	WT BE (kcal/mol)	WT Sc
1JRH	Affibody	-41.31	0.64	-71.58	0.81
1JRH	DARPin	-48.67	0.61	-71.58	0.81
1OB1	Affibody	-34.95	0.53	-41.98	0.69
1OB1	DARPin	-39.54	0.59	-41.98	0.69
2XT1	Affibody	-46.83	0.74	-66.21	0.75
2XT1	DARPin	-46.38	0.66	-66.21	0.75
3BDY	Affibody	-35.07	0.60	-45.67	0.68
3BDY	DARPin	-43.44	0.56	-45.67	0.68
3L5Y	Affibody	-35.06	0.62	-54.74	0.74
3L5Y	DARPin	-42.92	0.65	-54.74	0.74
3P30	Affibody	-32.88	0.71	-32.73	0.65
3P30	DARPin	-35.11	0.50	-32.73	0.65
3X3F	Affibody	-42.95	0.64	-56.51	0.76
3X3F	DARPin	-46.56	0.59	-56.51	0.76
4AL8	Affibody	-49.89	0.61	-40.83	0.67
4AL8	DARPin	-46.66	0.70	-40.83	0.67
5DFV	Affibody	-38.38	0.56	-55.47	0.63
5DFV	DARPin	-51.02	0.56	-55.47	0.63
5IKC	Affibody	-37.87	0.62	-50.06	0.67
5IKC	DARPin	-42.75	0.61	-50.06	0.67

Table 6.5 Total number of poses generated and percentage frequencies of different number of design and clash mutations for each of the 20 MutDock simulations. Per., Freq., Des., and Mut. refer to percentage, frequency, design and mutations respectively. Aff and DPn refer to Affibody and DARPin respectively.

Antigen PDB	Scaf-fold	Total poses	Per. Freq. of 3 Des. Mut.	Per. Freq. of 2 Des. Mut.	Per. Freq. of 1 Des. Mut.	Per. Freq. of 0 Des. Mut.	Per. Freq. of any clash Mut.
1JRH	Aff	2000	41.05	54.55	4.40	0.00	47.00
1JRH	DPn	2000	19.40	72.40	8.10	0.10	22.85
1OB1	Aff	2000	52.85	42.90	4.25	0.00	35.10
1OB1	DPn	2000	57.70	36.80	5.20	0.30	17.60
2XT1	Aff	1937	71.55	25.97	2.48	0.00	39.29
2XT1	DPn	1654	71.28	26.36	2.36	0.00	18.38
3BDY	Aff	1607	70.82	26.45	2.61	0.12	43.25
3BDY	DPn	1806	62.57	32.67	4.71	0.06	22.54
3L5Y	Aff	2000	73.75	25.00	1.25	0.00	39.05
3L5Y	DPn	2000	66.15	30.30	3.45	0.10	20.25
3P30	Aff	503	70.38	26.44	2.98	0.20	34.19
3P30	DPn	564	43.44	45.57	10.11	0.89	18.26
3X3F	Aff	1575	69.52	28.38	2.10	0.00	38.67
3X3F	DPn	1540	64.94	31.62	3.38	0.06	20.71
4AL8	Aff	2000	64.45	31.90	3.55	0.10	39.35
4AL8	DPn	1120	61.61	33.04	5.27	0.09	17.95
5DFV	Aff	2000	66.65	30.25	2.95	0.15	39.85
5DFV	DPn	2000	67.85	29.70	2.35	0.10	16.80
5IKC	Aff	2000	63.90	34.00	2.05	0.05	40.25
5IKC	DPn	2000	62.55	33.80	3.50	0.15	21.65

Table 6.6 Percentage frequencies of different amino acids in design mutations, native and final clash mutations in all 20 MutDock simulations.

Amino acid	Percentage frequency in design mutations	Percentage frequency in native clash mutation residues	Percentage frequency in clash mutation
ARG	9.86	0.70	0.00
ALA	0.00	0.00	0.81
ASN	13.48	0.29	13.41
ASP	11.51	0.00	14.74
GLN	0.00	0.05	0.00
GLU	13.34	0.06	5.78
GLY	0.00	0.00	0.00
HIS	0.00	2.49	0.03
ILE	0.00	0.00	1.46
LEU	0.00	0.07	54.89
LYS	7.48	0.02	0.00
MET	0.00	0.20	0.00
PHE	0.00	20.33	0.00
PRO	0.00	0.00	0.00
SER	14.82	0.00	7.58
THR	3.28	0.03	1.08
TRP	7.55	25.14	0.00
TYR	18.69	50.37	0.00
VAL	0.00	0.25	0.23

Table 6.7 Top binding energies (BE) from the MutDock and Zdock docking simulations along with their energy differences for the 20 antigen-scaffold complexes.

Antigen PDB	Scaffold	Top MutDock BE (kcal/mol)	Top Zdock BE (kcal/mol)	Top Zdock BE - Top MutDock BE (kcal/mol)
1JRH	Affibody	-41.3	-28.9	12.4
1JRH	DARPin	-48.7	-28.8	19.8
1OB1	Affibody	-35.0	-30.7	4.2
1OB1	DARPin	-39.5	-33.2	6.4
2XT1	Affibody	-46.8	-31.1	15.8
2XT1	DARPin	-46.4	-31.3	15.0
3BDY	Affibody	-35.1	-30.8	4.3
3BDY	DARPin	-43.4	-36.5	7.0
3L5Y	Affibody	-35.1	-26.5	8.6
3L5Y	DARPin	-42.9	-31.3	11.6
3P30	Affibody	-32.9	-38.6	-5.7
3P30	DARPin	-35.1	-39.3	-4.2
3X3F	Affibody	-43.0	-30.7	12.3
3X3F	DARPin	-46.6	-31.4	15.1
4AL8	Affibody	-49.9	-26.2	23.7
4AL8	DARPin	-46.7	-26.1	20.5
5DFV	Affibody	-38.4	-38.2	0.2
5DFV	DARPin	-51.0	-43.5	7.5
5IKC	Affibody	-37.9	-23.6	14.3
5IKC	DARPin	-42.8	-28.0	14.8

Table 6.8 Top binding energies (BE) from the MutDock (Mut) and HADDOCK (HAD) with native and MutDock scaffold docking simulations along with their energy differences for the 20 antigen-scaffold complexes. Also listed are the best binding energies of near native poses from the HADDOCK docking with MutDock scaffolds and the RMSD of these poses. All binding energies are reported in kcal/mol.

Antigen PDB	Scaf-fold	Top HAD BE	Top HAD BE - Top Mut BE	Top HAD w/ Mut scaffold (HAD-Mut) BE	Top HAD BE - Top HAD-Mut BE	Top near native BE	RMSD (Å)	Top Mut BE - Top HAD-Mut BE
1JRH	Aff	-36.7	4.6	-48.5	11.8	-48.5	0.7	7.2
1JRH	DP	-42.7	6.0	-48.9	6.2	-41.5	1.8	0.2
1OB1	Aff	-44.7	-9.7	-46.1	1.5	None	None	11.2
1OB1	DPn	-42.6	-3.1	-54.9	12.3	-45.2	0.6	15.3
2XT1	Aff	-43.0	3.9	-42.9	-0.1	-42.9	0.7	-3.9
2XT1	DPn	-46.3	0.1	-55.9	9.6	-41.7	3.0	9.5
3BDY	Aff	-27.4	7.7	-39.2	11.8	-35.0	0.6	4.2
3BDY	DPn	-45.9	-2.4	-40.4	-5.5	None	None	-3.0
3L5Y	Aff	-32.8	2.2	-37.4	4.6	-37.4	1.5	2.4
3L5Y	DPn	-47.8	-4.9	-54.2	6.3	-41.9	2.4	11.3
3P30	Aff	-39.1	-6.3	-48.4	9.2	-27.5	3.7	15.5
3P30	DPn	-52.1	-17.0	-53.6	1.5	None	None	18.5
3X3F	Aff	-35.3	7.7	-45.3	10.0	None	None	2.3
3X3F	DPn	-49.9	-3.4	-56.5	6.6	-50.7	1.0	9.9
4AL8	Aff	-34.3	15.6	-39.7	5.4	-39.7	0.7	-10.2
4AL8	DPn	-40.0	6.6	-46.9	6.9	-46.9	0.6	0.3
5DFV	Aff	-39.0	-0.7	-44.3	5.3	-41.9	3.1	6.0
5DFV	DPn	-47.7	3.3	-53.4	5.7	-53.4	1.1	2.4
5IKC	Aff	-34.6	3.3	-40.7	6.2	-34.0	2.4	2.9
5IKC	DPn	-40.0	2.8	-51.7	11.8	-49.6	3.4	9.0

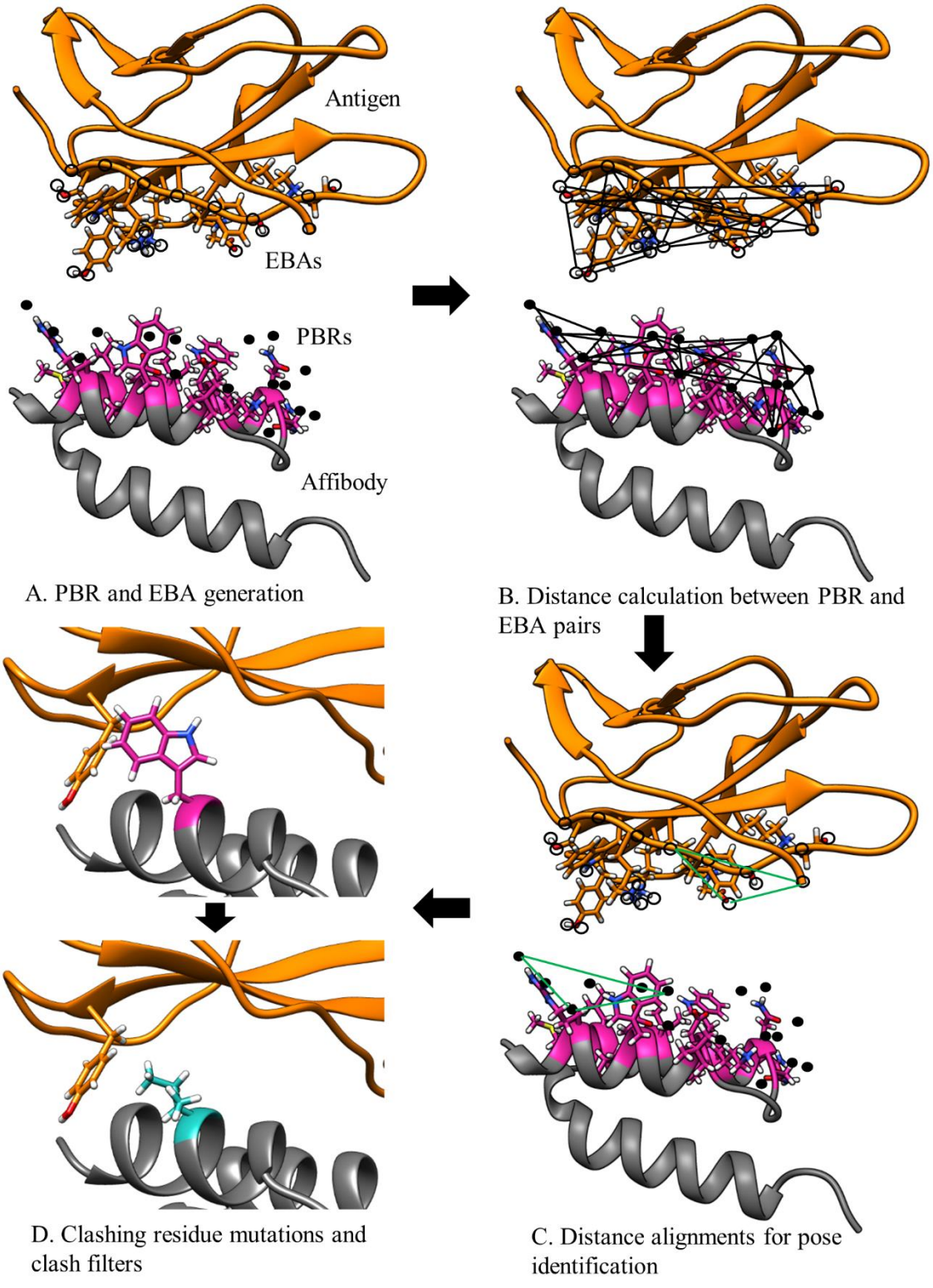
Antigen
EBAs
PBRs
Affibody
A. PBR and EBA generation
B. Distance calculation between PBR and EBA pairs
C. Distance alignments for pose identification
D. Clashing residue mutations and clash filters

Figure 6.1 The MutDock workflow. MutDock can be divided into two main steps: pose identification (panels A-C) and pose validation (panel D). Step A: PBRs are identified for all paratope residues and all other rotamers of variable residues (shown in pink). Similarly, EBAs are identified for all epitope residues. Step B: Pairwise distance calculations within the sets of PBRs and EBAs. Step C: Pairwise distance matching between PBR pairs and EBA pairs to identify groups of compatible low entropy interactions. Step D: Each pose from Step C is passed through steric clash filters and clashing variable side chains are mutated.

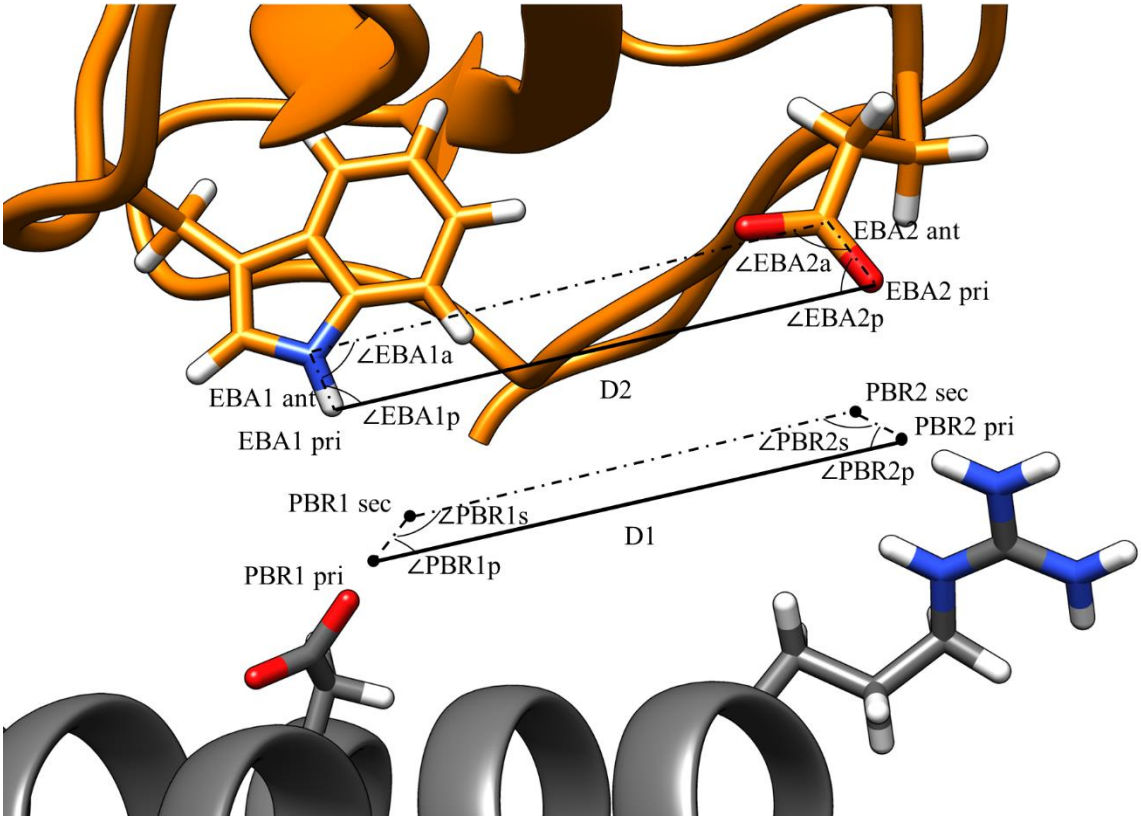

Figure 6.2 PBR and EBA pairwise distance and angle calculations. PBRs generated for paratope ASP and ARG. EBA identified for epitope TRP and GLU. The interactions being considered here are H-bonds between 1) ASP and TRP and 2) ARG and GLU. For the two interactions to be

compatible, |D1 – D2| < 1.8 Å, |∠PBR1p - ∠EBA1p| < 70°, |∠PBR2p - ∠EBA2p| < 70°, |∠PBR1s - ∠EBA1a| < 70° and |∠PBR2s - ∠EBA2a| < 70°.

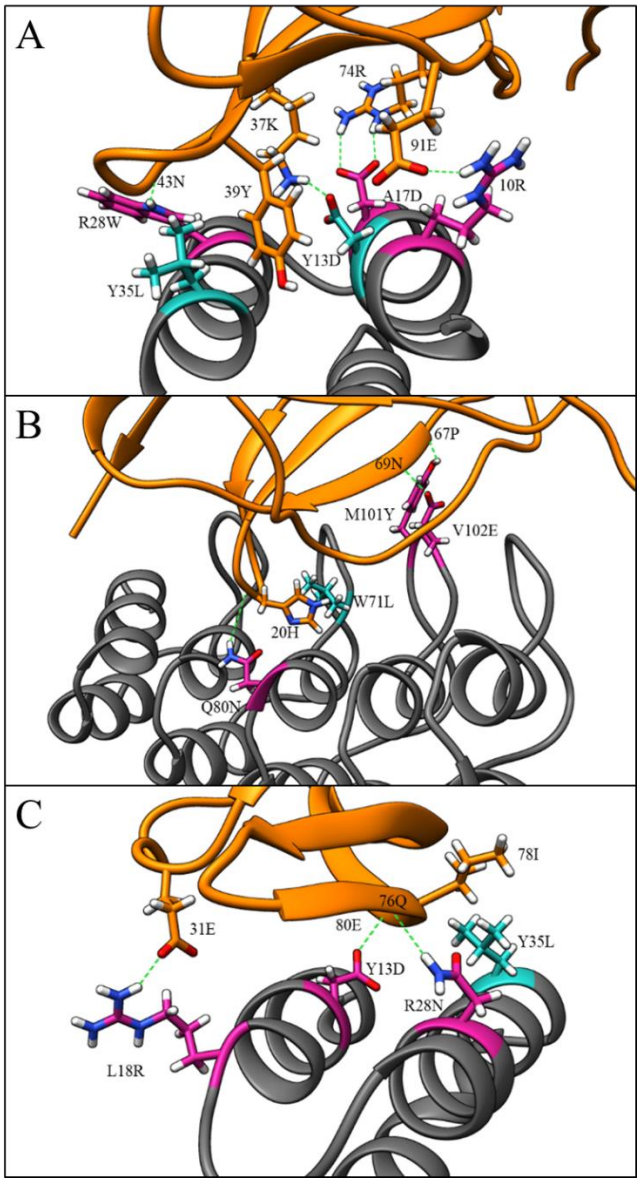

Figure 6.3 Example design and clash mutations in three MutDock designs. Design mutations are shown in pink, clash mutations are shown in dark cyan, and H-bonds are shown in broken green lines. (A). 1JRH-affibody. Native residue 10ARG and design mutations ARG28TRP andALA17ASP making H-bonds with 91GLU, 43ASN and 46TRP respectively. Clash mutation

TYR13ASP makes H-bond with 37LYS. Clash mutation TYR35LEU makes hydrophobic interaction with 39TYR. (B). 4AL8-DARPin. Design mutations VAL102GLU, GLN80ASN and MET101TYR making H-bonds with 69ASN, 20HIS and 67PRO respectively. Clash mutation TRP71LEU makes hydrophobic interaction with 20HIS. (C). 3BDY-affibody. Design mutations ARG28ASN, TYR13ASP and LEU18ARG making H-bonds with 76GLN, 80GLU and 31GLU respectively. Clash mutation TYR35LEU makes hydrophobic interaction with 78ILE.

References

[1] M. D. Smith, J. S. Rao, E. Segelken, and L. Cruz, “Force-Field Induced Bias in the Structure of Aβ21–30: A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields,” *Journal of Chemical Information and Modeling*, vol. 55, no. 12, pp. 2587–2595, Dec. 2015, doi: 10.1021/acs.jcim.5b00308.

[2] A. B. Rubenstein, K. Blacklock, H. Nguyen, D. A. Case, and S. D. Khare, “Systematic Comparison of Amber and Rosetta Energy Functions for Protein Structure Evaluation,” *Journal of Chemical Theory and Computation*, vol. 14, no. 11, pp. 6015–6025, Nov. 2018, doi: 10.1021/acs.jctc.8b00303.

[3] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” in *Computer Aided Chemical Engineering*, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds. Elsevier, 2018, pp. 2113–2118. doi: https://doi.org/10.1016/B978-0-444-64241-7.50347-5.

[4] R. L. Dunbrack Jr and F. E. Cohen, “Bayesian statistical analysis of protein side-chain rotamer preferences,” *Protein Sci*, vol. 6, no. 8, pp. 1661–1681, Aug. 1997, doi: 10.1002/pro.5560060807.

[5] S. J. Fleishman, S. D. Khare, N. Koga, and D. Baker, “Restricted side chain plasticity in the structures of native proteins and complexes,” *Protein Science*, vol. 20, no. 4, pp. 753–757, Apr. 2011, doi: https://doi.org/10.1002/pro.604.

[6] K. Vanommeslaeghe *et al.*, “CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields,” *Journal of*

Computational Chemistry, vol. 31, no. 4, pp. 671–690, Mar. 2010, doi: https://doi.org/10.1002/jcc.21367.

[7] R. F. Alford *et al.*, "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design," *Journal of Chemical Theory and Computation*, vol. 13, no. 6, pp. 3031–3048, Jun. 2017, doi: 10.1021/acs.jctc.7b00125.

[8] P. B. Stranges and B. Kuhlman, "A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds," *Protein Sci*, vol. 22, no. 1, pp. 74–82, Jan. 2013, doi: 10.1002/pro.2187.

[9] E. F. Pettersen *et al.*, "UCSF Chimera—A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: https://doi.org/10.1002/jcc.20084.

[10] S. Ståhl, T. Gräslund, A. Eriksson Karlström, F. Y. Frejd, P.-Å. Nygren, and J. Löfblom, "Affibody Molecules in Biotechnological and Medical Applications," *Trends in Biotechnology*, vol. 35, no. 8, pp. 691–712, 2017, doi: https://doi.org/10.1016/j.tibtech.2017.04.007.

[11] A. Plückthun, "Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy," *Annual Review of Pharmacology and Toxicology*, vol. 55, no. 1, pp. 489–511, Jan. 2015, doi: 10.1146/annurev-pharmtox-010611-134654.

[12] B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng, "ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers," *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, Jun. 2014, doi: 10.1093/bioinformatics/btu097.

[13] G. C. P. van Zundert *et al.*, "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes," *Journal of Molecular Biology*, vol. 428, no. 4, pp. 720–725, 2016, doi: https://doi.org/10.1016/j.jmb.2015.09.014.

[14] R. v Honorato *et al.*, “Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem,” *Frontiers in Molecular Biosciences*, vol. 8, 2021, [Online]. Available: https://www.frontiersin.org/article/10.3389/fmolb.2021.729513

[15] R. J. Pantazes, M. J. Grisewood, T. Li, N. P. Gifford, and C. D. Maranas, “The Iterative Protein Redesign and Optimization (IPRO) suite of programs,” *Journal of Computational Chemistry*, vol. 36, no. 4, pp. 251–263, Feb. 2015, doi: https://doi.org/10.1002/jcc.23796.

[16] J. Adolf-Bryfogle *et al.*, “RosettaAntibodyDesign (RAbD): A general framework for computational antibody design,” *PLOS Computational Biology*, vol. 14, no. 4, pp. e1006112-, Apr. 2018, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006112

[17] S. J. Fleishman, J. E. Corn, E.-M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker, “Hotspot-centric de novo design of protein binders,” *J Mol Biol*, vol. 413, no. 5, pp. 1047–1062, Nov. 2011, doi: 10.1016/j.jmb.2011.09.001.

[18] J. Dou *et al.*, “De novo design of a fluorescence-activating β-barrel,” *Nature*, vol. 561, no. 7724, pp. 485–491, 2018, doi: 10.1038/s41586-018-0509-0.

[19] C. Longxing *et al.*, “De novo design of picomolar SARS-CoV-2 miniprotein inhibitors,” *Science (1979)*, vol. 370, no. 6515, pp. 426–431, Oct. 2020, doi: 10.1126/science.abd9909.

7. Chapter 7 - Summary and Future work

Over the past two decades, significant work has been done towards developing computational tools for the design and docking of proteins. These design tools depend on computationally expensive iterative cycles of force field dependent energy calculations and optimizations. The overall purpose of this dissertation was two-fold: 1) to gain a better understanding of protein-protein residue interactions and 2) to implement a new design paradigm that targets the formation of groups of strong interactions via simple geometry alignments. This paradigm was implemented for both antibodies and other fixed-backbone binding proteins.

Chapters 2 and 3: Database development and AUBIE

In Chapter 2, I developed a non-redundant database of antibody-antigen complexes from the IMGT database and analyzed pairwise residue energies between epitopes and paratopes in three force fields. This database served as a reference set for all comparison studies done in this dissertation. According to all the force fields, five residues contributed, on average, more than half of the total binding energy. It was hypothesized that a quick and general computational design approach for binding proteins could be developed that targeted these few "hotspot" interactions in a sequential step method. Unlike conventional design tools, this method would be independent of any computationally expensive, force field dependent iterative design cycles. In Chapter 3, I developed AUBIE, a first-of-its-kind computational method that can rapidly design antibodies and other binding proteins for any specific epitope. The workflow has been divided into two broad steps: database generation and protein design. To generate the database, non-redundant loops that match the geometric criteria of the binding loop attachment points are identified from the PDB database. Optimal binding positions for various interactions types are then identified for all the

loops in the database. The initial steps of the design section include positioning the antigen to face the epitope towards the framework and rotating it around the z axis in one degree steps. For each rotation, distances between the antigen residues and SBRs are calculated to identify equivalent movements which correspond to a simultaneous fit for that combination of epitope residues. Final solution structures are identified after checking for steric and electrostatic compatibility between the antigen, loops and framework.

AUBIE was used to design HER2 binding antibodies in a collaborative project with Dr. Spangler at Johns Hopkins University. Although the top AUBIE designs were computationally predicted to be statistically significantly better than naturally occurring antibodies, they did not demonstrate binding when tested experimentally. Comparing AUBIE-designed and known antibody paratopes revealed that AUBIE designed paratopes had a larger number of ARG, LYS and GLN residues than the average antibody paratope. The common property shared by these amino acids is their large conformational variability due to their highly flexible side chains. From a thermodynamics point of view, the change in Gibbs free energy is the sum of the enthalpic loss in energy, due to bond formation and solvation, and the entropic loss in energy, due to the loss of motion of binding atoms. The top AUBIE designs consisted of large beneficial enthalpic energy contributions due to several H-bond and salt bridge formations. Their failure to bind to HER2 demonstrated that the enthalpic loss in energies were not enough to compensate for the larger gains in entropic energy of the designed paratopes.

Chapter 4: Pre binding conformational stability

Upon learning that the AUBIE designs contain a large number of highly flexible side chains that required very strong interactions to compensate for their large entropy loss upon binding, I set

out to get a better understanding of conformational stability in binding interfaces. In Chapter 4, databases of antibody-antigen and non-antibody protein-protein complexes were analyzed for various prestabilization related features. Residue side chains were defined as either stable or unstable before binding and quantified the different interactions they made. The results demonstrated that most interactions in the binding interfaces consist of either one or two stable interacting residue parts. Moreover, this preference for stability is selected and not an expected result of the stabilities of paratopes and epitopes before binding. This binding surface conformational stability was defined as rotamers lost per buried SASA and was substantially larger for paratopes in AUBIE solutions than in known antibodies. Lastly, key antibody residue positions were identified that are more likely to host certain stable amino acid side chain interactions than other positions.

The computational analysis of the conformational stability of residues by themselves and in groups can be explored in several directions and depths in future work. In this work, very simple definitions of stability based on the fraction of clashing rotamers were used. Side chain stability is a dynamic property. This property was estimated from the minimized complex structure since it was not possible to perform MD simulations for each database complex. B-factor values are often used to predict protein atom stabilities [1]. B-factor prediction tools are being developed [2] and can be incorporated into binding metrics to judge designed interfaces by comparing them to values from known antibody-antigen complexes. This work focused mainly on interactions between protein atoms. Water molecules are known to form H-bonds within binding interfaces [3]. The features that were analyzed can be extended to complexes with interface waters and the stability of side chains forming H-bond networks via water molecules can be studied. The interaction patterns determined from this work can be used to guide binding protein design. As mentioned

earlier, a common strategy for binding protein design is to dock proteins or single loops and to mutate. Liu et al. designed Keap1 binding proteins via this approach [4]. The frequency of CDR residues to host stable side chains or the interaction types they make can be used to guide this docking step. For example, to dock a CDR or antibody towards a hotspot unstable ARG or LYS in the center of epitope, aromatic side chains from positions 37 and 38 of CDR1 loops can be targeted to make hydrophobic interactions first rather than searching through all the CDR positions. Similarly, knowledge from Chapter 4 can also be used to dock CDRs in such a way that they use two CDR residues to make interactions with the ARG or LYS side chain. Another approach that can be used for binding protein design is to use sequence homology between the target epitope and known antibody binding epitopes. Akbar et al. analyzed frequent sequence motifs between interacting epitope and paratope regions in antibody-antigen complexes [5]. Stability related information from this work can incorporated into such models to increase chances of designing stable interfaces. The work presented in this chapter aided us in understanding the key role played by the entropic component of the change in Gibbs free energy during antibody binding, which was not adequately captured by the forcefields used in Chapter 3.

Chapter 5: AUBIE modifications

In Chapter 5, the AUBIE approach was modified to only target low entropy H-bonds and salt bridges made by paratope ASP and GLU. Along with other changes, the improved AUBIE approach was tested by using it to design antibodies for 25 randomly selected antigens. 18 / 25 top AUBIE designs that binding energies that were within one standard deviation of the mean of known antibody-antigen binding energies. For eight designs, the top AUBIE complex has better or similar binding energies than the WT complex. When compared with OptMAVEn 2.0, AUBIE

was able to generate designs with similar binding energies for the same epitope. AUBIE-designed interactions in a top design persisted during a 50 ns long MD simulation.

One way of doing force field - independent design is to replicate known structural and sequence features. In AUBIE, that was done by replicating low entropy H-bonds, interactions often seen in antibody-antigen complexes. However, doing affinity maturation without force fields is challenging since there is a need to minimize total energy to predict long and short range effects of mutations. AUBIE is able to generate designs with acceptable metrics with no affinity maturation and few mutations made only due to clashing residues. This approach now accounted for both the enthalpic and entropic elements of the change in Gibbs free energy by targeting only low entropy H-bonds. As shown in the results, AUBIE designs can be improved with further, force field guided, computational affinity maturation.

The current version of AUBIE can be improved in various ways. The database generation step in AUBIE defines loop compatibility based on the lack of steric clashes, which can be improved by including inter-loop interactions. Doing so would encourage the identification of groups of loops that have interactions with each other and increase the overall stability of the paratope. The current version of AUBIE rotates the antigen around the z axis in one degree steps. This search method needs to be made independent of stepwise movements or rotations and transformed into a MutDock-like search approach. This way, a larger solution space can be searched and better designs can be identified. Furthermore, AUBIE searches for H-bonds in all CDR residues positions. This search can be made more efficient by first searching through high probability positions identified in Chapter 4. To reduce runtimes and computational load, the AUBIE code will be transferred from Cython to C++. Hydrophobic interactions are as crucial as H-bonds or salt bridges in protein binding and AUBIE needs to be improved to target these

interactions too. One way to overcome the challenge of their lack of specific geometries is to replicate orientations of nonpolar contacts from known binding complexes. The modification bought about in this chapter rigidified the designed paratopes and brought them one step closer to known antibody paratope stabilities. One possible route to replicate known paratope stability is to obtain binding loops only from antibodies and/or to enforce constraints on aromatic or small polar amino acid frequencies in CDRs. Recently improved protein sequence-to-structure prediction tools like AlphaFold2 [6] can be used to validate the antibody structure predictions by *de novo* design tools like AUBIE. As a final verification of AUBIE's efficacy, the designed antibodies need to be tested experimentally for different antigens.

Chapter 6: MutDock

In Chapter 6, I developed MutDock, a novel computational approach for the generation of mutated, docked scaffolds designed to bind target epitopes. The approach identifies regions around the scaffold paratope which can host polar epitope atoms to form H-bonds. Similar to AUBIE, the MutDock approach attempts to balance entropic and enthalpic elements of the change in Gibbs free energy by targeting only low entropy H-bonds. Pairwise distance alignment between the epitope atoms and H-bond regions is used to obtain groups of low entropy H-bonds that can be formed simultaneously. Each group constitutes a unique pose that is passed through several compatibility filters. MutDock was benchmarked by docking ten antigens with two scaffolds. The predicted binding energies of the top MutDock poses were comparable to those of known binding complexes when accounting for the influence of interface size on calculated energies. The MutDock poses were significantly better than ZDOCK's results for 17 of the 20 predicted complexes. When compared to HADDOCK, MutDock performed better in 9 / 20 complexes,

comparably in 5 / 20, and worse in the remaining 6 / 20 complexes. However, the HADDOCK scores improved for 16 / 20 complexes when HADDOCK was used to dock the MutDock-designed scaffolds.

MutDock could also be used for antibody design. Nimrod et al. docked an antibody database to the IL-17A epitope using ZDOCK and Hex [7]. Similarly, MutDock can be used to dock an antibody to a target epitope, where CDR side chains that do not interact with other residues can be identified and labelled as variable residues. I believe that single or double mutations to non-interacting side chains in antibody CDR1 and CDR3 loops will have minimal effect on antibody backbone structure. However, unlike the numerous antibody-only protein design programs, MutDock is also able to design proteins that bind using variable surfaces, such as affibodies and DARPins.

Several improvements need to be made to the current MutDock implementation. To reduce the runtimes, the source code will be shifted to C++ and more coarse-grained steric clash filters will be used. Furthermore, an additional interface size filter based on atom-atom contacts will also be added to prevent generation of poses with low buried interface areas. Similar to RIFdock, hydrophobic interactions and a larger rotamer library will be incorporated into MutDock. Relaxed steric clash constraints between heavy atoms of interface side chains were used since the approach did not perform any energy calculations or minimizations during the design process except at the end. Such relaxed clash constraints were also observed in ZDOCK poses as several heavy atoms were closer than 2 Å in a single pose. The use of strict clash constraints through vdW force calculations can lead to the loss of good solutions even though is a safer approach. There is a need for smarter but simple steric clash constraints that can predict if energy minimizations would

resolve specific clashes. One possible strategy is to use clash constraints that are a function of the amino acid types involved.

Contributions

I believe that this dissertation has three major contributions. 1) The first contribution is the quantitative understanding of pre-binding stability of protein binding interfaces. The knowledge gained from Chapter 5 can open the doors for optimization-free feature-driven design approaches for binding proteins. The other two contributions include the application of simple pairwise distance alignments for targeted interaction formation for the 2) *de novo* design of antibodies and 3) mutational docking of fixed-backbone protein binders. This work serves as an example that geometry-based single step calculations can be used to tackle complex combinatorial problems of binding loop and residue selection in different types of binding proteins. To the best of our knowledge, no other work has attempted to solve the epitope-specific binding protein design problem using such geometrical alignment-based techniques and with further development, these techniques can become the standard way to design binding proteins. There are various applications of these contributions beyond the direct scope of the work presented in this Chapter. The pairwise distance alignment approach for targeted design can be used for the design of other protein functionalities too such as pH depended binding and enzyme active sites. pH-dependent binding interfaces are rich in HIS residues, since its side chain is the only amino acid with multiple protonation sites [8]. MutDock and AUBIE design techniques can be modified to target the formation of H-bonds made by HIS side chains for pH dependent binding interface design. Enzyme active sites have stable precise geometries before substrate attachment [9]. Side chain orientations

of active side triads can be replicated from an enzyme to another scaffold using distance matching techniques used in this work.

ML based tools are being developed at rapid pace for different antibody and protein prediction purposes [10], [11]. There are several reasons why the MutDock and AUBIE methods differ from any ML-based design approach. A ML-based tool will rely only on the frequency of structural elements to base its design decisions on. On the other hand, the approaches developed in this work are based on thermodynamics and ease of computational design. For example, a ML based antibody design tool may target long HCDR3 loops since they are prevalent in known antibodies, but AUBIE does not target such long loops due to the difficulty in predicting their pre-binding conformations. Furthermore, using ML tools like neural networks often result in black box-like models which are difficult to interpret [12]. On the other hand, every step of the approach that we have developed is based on either thermodynamics or computational load related reasons that are easy to understand. A major issue with using ML methods for binding protein related prediction tools is the lack of large learning sets of binding interface structures. One way to overcome this challenge is to combine biophysics-based constraints, like the ones identified in Chapter 4, with ML-based learning abilities to develop a new class of computational tools.

References

[1] P. Radivojac *et al.*, “Protein flexibility and intrinsic disorder,” *Protein Science*, vol. 13, no. 1, pp. 71–80, Jan. 2004, doi: https://doi.org/10.1110/ps.03128904.

[2] D. Bramer and G.-W. Wei, “Blind prediction of protein B-factor and flexibility,” *The Journal of Chemical Physics*, vol. 149, no. 13, p. 134107, Oct. 2018, doi: 10.1063/1.5048469.

[3] B. C. Braden, B. A. Fields, and R. J. Poljak, “Conservation of water molecules in an antibody–antigen interaction,” *Journal of Molecular Recognition*, vol. 8, no. 5, pp. 317–325, Sep. 1995, doi: https://doi.org/10.1002/jmr.300080505.

[4] X. Liu *et al.*, “Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping,” *Scientific Reports*, vol. 7, no. 1, p. 41306, 2017, doi: 10.1038/srep41306.

[5] R. Akbar *et al.*, “A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding,” *Cell Reports*, vol. 34, no. 11, p. 108856, 2021, doi: https://doi.org/10.1016/j.celrep.2021.108856.

[6] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[7] G. Nimrod *et al.*, “Computational Design of Epitope-Specific Functional Antibodies,” *Cell Reports*, vol. 25, no. 8, pp. 2121-2131.e5, 2018, doi: https://doi.org/10.1016/j.celrep.2018.10.081.

[8] T. Igawa, F. Mimoto, and K. Hattori, “pH-dependent antigen-binding antibodies as a novel therapeutic modality,” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1844, no. 11, pp. 1943–1950, 2014, doi: https://doi.org/10.1016/j.bbapap.2014.08.003.

[9] Y. Xie *et al.*, “Enhanced Enzyme Kinetic Stability by Increasing Rigidity within the Active Site*,” *Journal of Biological Chemistry*, vol. 289, no. 11, pp. 7994–8006, 2014, doi: https://doi.org/10.1074/jbc.M113.536045.

[10] R. Akbar *et al.*, “In silico proof of principle of machine learning-based antibody design at unconstrained scale,” *MAbs*, vol. 14, no. 1, p. 2031482, Dec. 2022, doi: 10.1080/19420862.2022.2031482.

[11] G. Liu *et al.*, “Antibody complementarity determining region design using high-capacity machine learning,” *Bioinformatics*, vol. 36, no. 7, pp. 2126–2133, Apr. 2020, doi: 10.1093/bioinformatics/btz895.

[12] Y. Sheu, “Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research,” *Frontiers in Psychiatry*, vol. 11, 2020, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.551299