Application of *In-Silico* Protein Engineering and Optimization Methods to Design Target-Specific Antibody Mimetics

by

Ritankar Bhattacharya

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 3, 2024

Keywords: proteins, binders, antigens, therapeutics,
binding energy, protein engineering

Approved by

Robert J. Pantazes, Chair, Ph.D. Department of Chemical Engineering
Christopher A. Kieslich, Ph.D. Department of Chemical Engineering
Allan E. David, Ph.D. Department of Chemical Engineering
Jacek Wower, Ph.D. Department of Animal and Dairy Sciences
Mohtadin Hashemi, Ph.D. Department of Physics

Abstract

Antibodies (Abs) are proteins that bind target antigens (Ags) and interact with immune receptors. Over the years they have been developed into therapeutics with tremendous potential to treat various diseases. Their global market share is projected to cross $500 Billion by 2030. As excellent as therapeutic antibodies are when administered to patients, they exhibit limitations associated with solubility, aggregation propensity, and degradation tendency when working outside the body. Therefore, it is important to move beyond conventional Abs and develop alternative protein binders that can mimic their functions while being able to address these limitations at the same time.

Previously, several works have developed *in-silico* algorithms to design functioning proteins to bind their targets. However, they typically generate many predictions, which are practically unfeasible for experimental testing due to high cost and time constraints. This dissertation demonstrates the application of complex scientific protocols to shortlist the most promising computational designs from a large dataset of predictions most worthy of experimental testing.

Chapter Two will use a novel PETEI algorithm to design thousands of 10^{th} fibronectin type III domains to bind 3 distinct tag epitope peptides, FLAG, HA, and MYC. Since it is unfeasible to test all the designs, they will be energy minimized using CHARMM and Rosetta forcefields based on three or fewer positive CDR residues. The Rosetta Interface Analyzer will calculate and analyze their binding energy per buried surface area. Similar binding metrics will be calculated for an energy minimized, non-redundant 231 Ab-protein database and compared to the designs. The designs will be rank ordered and the top 30 will be selected for experiments. This chapter is part of a bigger project, submitted for publication as of July 2024.

Chapter Three will use another novel algorithm called MutDock to design 2145 DARPin – α-Cobratoxin complexes. They will be down-selected to the top 39 using the scientific protocols mentioned previously, and affinity matured using specific RosettaDesign protocols to improve binding. 3900 affinity matured designs will be narrowed down to 73 by applying similar protocols and visually inspected in UCSF Chimera. Short 5 ns NAMD MD simulations will be conducted to assess their structural integrity over time and conformational binding metric trajectory will be plotted against time to analyze their characteristics. Their standard deviations will be calculated and this analysis will assist in down-selecting the top 51 designs for experiments.

Chapter Four will use tools like CamSol, HoTMuSiC, SCooP, IEDB De-immunization, and RosettaDesign to fine-tune the physicochemical properties of 5 α-helical bundles to arrive at the desired protein. These top 5 were shortlisted from a list of 75 similar structures after discarding the ones with Cys residues, *de novo* designs, extra end loop, and the position of an end loop to cause steric hindrance. RosettaFold predicted 1 out of 5 re-designed sequences to fold to their native state. This binder will be the starting point for the next 2 more iterative rounds of fine-tuning to generate 16 different 'ideal' binding proteins. Therefore, observably, the level of complexity of the development and applied scientific protocols implemented in the subsequent chapters keeps increasing progressively which leads to a more sophisticated down-selection process for designs worthy of experiments.

Chapter Five will elucidate the usage of I-Tasser, Robetta, and trRosetta to model 3 anti-CTLA4 nanobodies and canine CTLA4 protein. ZDOCK was used to dock them, which resulted in 85 of the 90 predictions showing the nanobodies binding to the CTLA4 conserved epitope. This was a modeling of experimental results, and it was published in Scientific Reports. Chapter Six will explain the future directions of the current projects.

Acknowledgments

Firstly, I would like to acknowledge my advisor, Robert J. Pantazes (Ph.D.), whose continuous support, motivation and words of encouragement was crucial towards my entire time as a graduate student. He is an extremely reasonable, humble and patient human being who provided all kinds of support be it guidance in research and courses, financial support for attending conferences and workshops to help my professional growth as well as mental support during tough times especially during the COVID-19 pandemic. Next, I would like to thank my parents, without whom I would not have been able to make it to the United States to pursue my Doctoral degree. They have always supported and guided me to help bring out the very best in me. It is because of their continuous motivation, blessings, hard work and sacrifice that I have been able to come a long way. Next, I would like to acknowledge my friends for being a continuous support throughout my entire time in graduate school. I connected through video calls with those who lived far away and hanged out together on Friday nights and weekends with those who lived here in Auburn. I would also like to thank and be grateful to my committee members, colleagues for their continuous support, and words of motivation and to the Department of Chemical Engineering for providing me the opportunity to pursue my doctoral degree in one of the best institutions in the country. Finally, I would like to thank God. I have prayed more times to my deities during my time as a doctoral student than I have in my entire life. I have always had faith; however, I found hope in my religion during this time.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CPD	Computational Protein Design
CPE	Computational Protein Engineering
PDB	Protein Data Bank
vdW	van der Waals (interaction)
Ab	Antibody
Ag	Antigen
BE	Binding Energy
SASA	Solvent Accessible Surface Area
BSA	Buried Surface Area
Sc	Shape complementarity
MDs	Molecular Dynamics simulation
AffMat	Affinity Maturation
DARPin	Designed Ankyrin Repeat Protein
α-Cbtx	α-Cobratoxin
IEDB	Immune Epitope Database
VMD	Visual Molecular Dynamics
NAMD	Nanoscale Molecular Dynamics
CHARMM	Chemistry at HARvard Molecular Mechanics
PETEI	Protein Designed with Targeted Exceptional Interaction
Nb(s)	Nanobody(ies)
sdAb	Single Domain Antibody

1. Introduction

1.1. Research focus of this dissertation

Proteins are large biomolecules that mediate some of the fundamental life processes, and how they do so have been an area of intense research for scientists over the last few decades [1]. In particular, therapeutic proteins like antibodies (Abs) have revolutionized the field of medical science. These are defense proteins that assist the body to fight infections against disease causing pathogens. An Ab is a 'Y' shaped globular protein that is produced by the immune system, which binds an antigen (Ag) and marks it to be neutralized[2], [3], [4]. Immune proteins such as Fcγ receptors initiate the immune response through signaling pathways whereas complement receptors activate the complement pathways to defend against invading pathogens[5], [6], [7], [8], [9], [10], [11], [12], [13]. The Ab-Ag complex is then engulfed by phagocytes like macrophages. Therefore, the networking of these proteins at the cellular level is critical to keep the body healthy and fully functioning.

Therapeutic Abs have been developed to treat several diseases because of their therapeutic properties and are excellent when administered to patients. However, they have issues related to solubility, aggregation propensity, and degradation tendency when working outside the body. This dissertation chapter will acknowledge the usefulness of therapeutic Abs to treat diseases, identify their limitations, introduce alternative protein binders and *in-silico* methods as a possible solution to address the limitations, and most importantly focus on applying the appropriate complex scientific protocols necessary to finalize designs that are worthy of being experimentally tested in the laboratory. With the correct approach, a combination of *in-silico* tools to design alternative proteins can establish a faster and cost-effective technique to bind target antigens. However, these tools can make predictions based on scientific rationale and available data. The likelihood of

experimental success with fewer *in-silico* designs is extremely low. Therefore, hundreds and thousands of designs are generally predicted to increase the probability of experimental binding. However, to find the designs that bind, it is practically unfeasible to experimentally test all of them due to time constraints and monetary expenses. For this reason, it is important to apply the appropriate scientific protocols to shortlist the most promising designs from the large set of *in-silico* predictions that are most worthy of experiments.

This is the commonality in all the dissertation chapters and the level of complexity of the scientific protocols will progressively increase as they are implemented to the designed complexes in the subsequent chapters. However, before the technical contents of the individual chapters are introduced to explain the use of these protocols, the readers must have a basic understanding of this field, in particular the structure and functions of an Ab. The reason is that Abs is the motivation behind this work and unless the readers are well acquainted with their functioning, uses, production, and limitations, they will not be able to understand the need for a paradigm shift to introduce computational protein engineering to design alternative binding proteins. Therefore, a systematic approach is necessary to progress through the contents of the chapters starting with the background information.

1.2. Defense Proteins: Structure and Function

Abs or Immunoglobulins (Igs) are heavy (150 kDa) [3], heterodimeric, 'Y' shaped globular glycoproteins, produced by specialized cells of the immune system, which identify, target and attach to foreign pathogens called antigens, and mark them to be neutralized by immune cells of the body [3], [4]. The tips of the V region of the Ab are called the paratope [14] and they bind to the antigenic epitope [15], [16]. Paratopes are enriched in aromatic amino acids [14], which being bulkier have fewer degrees of freedom (DOF) compared to long-chain residues. This provides

rigidity and more stability to bind the epitopes. The paratope-epitope interaction can be thought of as a lock and key model as the two complement each other with high binding affinity and specificity, giving rise to a strong complex. Fundamentally, an Ig consists of twelve domains, which make up four polypeptide chains held together by disulfide bonds [3]. There are two heavy (H) and two light (L) polypeptide chains, which can be further divided into four variable domains, 2 V_H, and 2 V_L that bind to the antigenic determinants or epitopes, and eight constant domains (the 'tail') some of which interacts with Fc or complement receptors, thereby, signaling immune response [3], [4], [17]. Figure 1 shows a fully crystallized Ab structure with its parts labeled. Due to this extent of modularity, Abs can be cleaved using the appropriate enzymes to generate fully functional smaller Ig-like molecules. As shown in Figure 2 a protease called Papain cleaves an Ab into a fragment, antigen binding (Fab(s)) and the fragment, crystallizable (Fc), which remain attached [3], [4], [18]. Pepsin, another protease, cleaves the molecule to generate the $F(ab')_2$ fragment, consisting of both the 'V arms' joined together and a pFc' region [3], [4], [18]. Another important Ig fragment has the variable domains of the heavy and light chains linked together by a stretch of a peptide, called single chain variable, fragment, or scFv [3], [19]. As mentioned earlier, a nanobody (Nb)[20] is a sdAb, engineered from a VHH. These independent molecules have been genetically modified because of their dynamic structure, flexibility, modularity, and small size for improved therapeutic applications to treat diseases and medical anomalies.

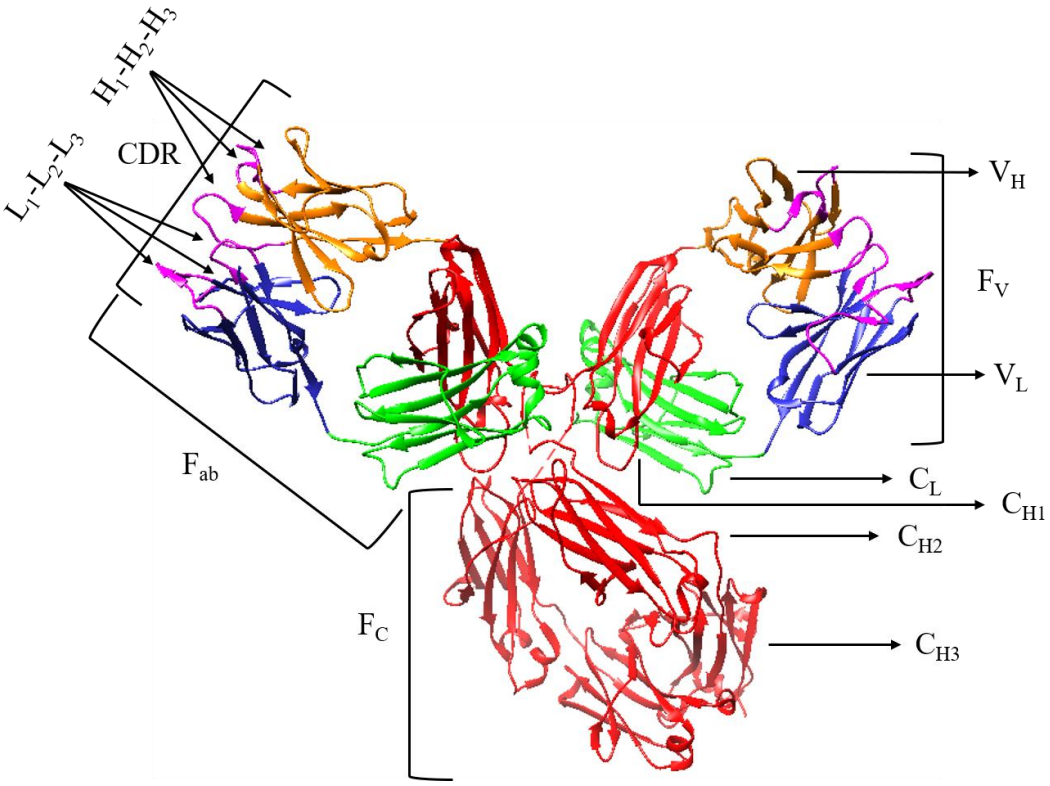

Figure 1 **A full antibody**. A fully crystallized IgG4 antibody PDB 5DK3. C_{H1-3} are highlighted in red, C_L in green, V_H and V_L are in orange and blue respectively with the CDRs marked in magenta. The L_{1-3} and H_{1-3} represents the light and heavy chain CDRs respectively. Fc and Fv represents the constant and variable domains and Fab are fragment, antigen binding.

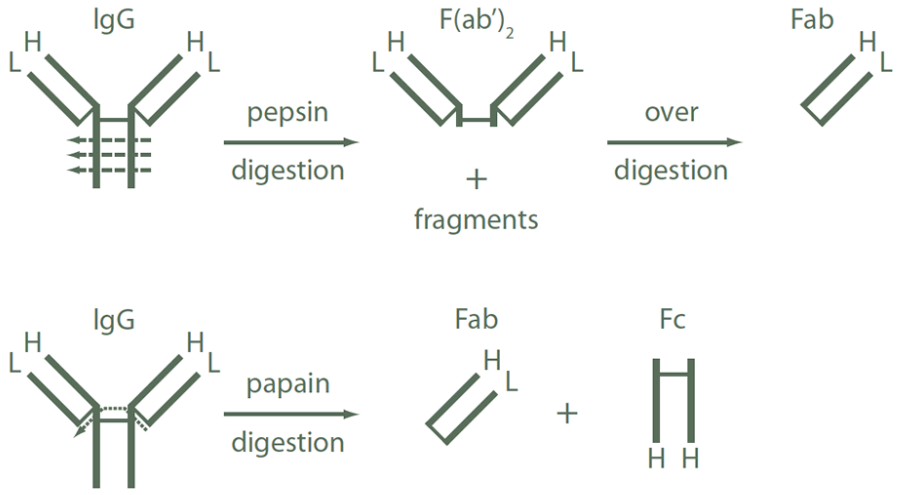

Figure 2. **Antibody fragment examples**. An IgG being cleaved by Pepsin and Papain to generate F(ab')2 and Fab segments respectively (*Source: www.stratech.co.uk/wp-content/uploads/2016/10/tech-info-fig1.gif*)

The V-shaped tips of the N-terminal variable region, V_H and V_L is a region of hypervariability known as Complementarity Determining Regions or CDR(s), which serves as the primary binding site to an epitope [4], [21]. Each 'arm' of the 'Y' structure has three CDRs loops or coils. H1-3 for the heavy chain domain, L1-3 for the light chain domain, and four stable framework regions, FR1-4. As previously described, using the lock and key mechanism, an Ab can 'tag' an Ag, to be neutralized by agents of the immune system. Apart from the CDRs, certain regions of the Ab can interact with immune receptors and signal immune response. The fragment, crystallizable, Fc region, often called the constant region of Abs near the base of the 'Y' can interact with immune receptors and assist in signaling immune response[3], [5]. This region is not antigen specific since it only interacts with immune proteins of the body and no foreign pathogen. The Fc region comprises highly conserved N-glycosylation sites where the immune receptors bind[22], [23]. The glycosylation process is essential for Ig effector functions since it plays a crucial role in the folding of such mammalian/eukaryotic glycoproteins [22], [24]. Additionally, it helps in steric protection from proteolytic degradation, and protein-protein interactions (PPI).

The information mentioned above generally sums up the basic structure and functions of an Ab. These proteins have been engineered and used in the medical industry as therapeutic Abs for decades and at a global scale have an enormous market share.

1.3. Antibodies at a global scale

The pharmaceutical industry is an enormous contributor to the global economy and it plays a significant role in the healthcare system. It is a versatile industry that is engaged in a variety of activities such as research and development, drug discovery, clinical trials, manufacturing and distribution, marketing and sales, intellectual property, quality check, and other areas. This industry has grown at a tremendous pace in the last few decades and their global sales were

reinforced further due to the pandemic. An article published in Human Antibodies (Vol 30, 2022) says that the Monoclonal Ab (mAb) revenue is expected to generate $300Bn by 2025[25]. Another article published in Antibody Therapeutics (Vol 5, 2022) has projected that the global Ab therapy market will reach $445Bn by 2028[26]. Considering these projections, their trajectory might cross more than half a trillion USD by 2030. In 2020, an article was published in JBS by Ruei-Min Lu et. al, highlighting ample market value and revenue statistics of therapeutic antibodies along with their future projections [27]. In the same year, due to the COVID-19 pandemic, the global therapeutics market observed unprecedented demands that pushed their market economy to newer heights. A report by Fortune Business Insights predicted a growth of $792.37 Bn by 2032 in the mAb global markets (Report ID: FBI102734, Source: https://www.fortunebusinessinsights.com/monoclonal-antibody-therapy-market-102734). An article was published in DARU-JPS, Springer, 2020, which further highlighted the short-term and long-term effects of COVID-19 [28]. It states that there were 113 medicines and 53 vaccines in the R&D pipelines or phase trials for patients diagnosed with the virus. Hence, the global pharmaceutical market saw soaring demands in such therapeutics in the last three years and observably therapeutic Abs has proven to be excellent binding proteins of this industry. They also have an interesting history of production and the way the Ab production technology has evolved over the years is important to understand and judge if there is any paradigm shift necessary or otherwise.

1.4. Antibody production and limitations

Originally, therapeutic Abs was developed by immunizing animals such as mice or rabbits with a target Ag. The mammal would gradually be injected with the Ag of choice in small quantities over time to help generate an amplified immune response. It has been the common

practice to develop immunity against that particular Ag. The minute quantities ensured that no harm was being caused to the animal or posing a life threat, and it was just enough to help develop Abs against the target Ag over time. The Ab was obtained from the animal by extracting blood and purifying it. Since the Abs were generated inside the animal, this process was termed an *in vivo*[29] process. As time progressed, the method of Directed Evolution[30], [31], [32] was introduced, which employed *in vitro* [33], [34], [35] experiments to generate therapeutic antibodies in the laboratory. As the name suggests, directed evolution mimics the process of natural selection to drive proteins toward the desired end goal. It is based on iterative rounds of library mutagenesis (creation of genetic variants through mutation), expressing the variants, isolating the desired ones (selection), and finally amplifying the desired ones [31], [32], [36]. The steps are repeated using the best variant from the previous round to achieve improvements and finally get directed to the desired protein. Both methods have been successful in consistently designing high affinity therapeutic Abs. Over the years, Ig-like scaffolds such as mAbs, bispecific Abs (bsAbs), Fab, Fcabs, scFvs, scFv-Fcs, etc. have been developed in the laboratory [18], [19], [37], [38]. They have demonstrated enormous potential for binding to the target epitope with high affinity and specificity[39]. mAbs have played a pivotal role as potential therapeutic agents in the treatment of cancers, autoimmune diseases, and cardiovascular diseases with high success rates [27], [39], [40]. Thus, a lot of effort has been put into developing target specific therapeutic proteins for treating autoimmune diseases and cancer models, by conducting laboratory experiments.

Although such Abs are excellent therapeutics when administered to patients, they have certain limitations when being worked with outside the body. They must be expressed in specialized eukaryotic (true nucleus) cells and require post-translational modifications (PTM) like N-glycosylation to fold and function appropriately as therapeutics [2], [23], [41]. They are relatively

large molecules (150 kDa approx.), and their size makes it difficult for them to penetrate tumors and they have slowed systemic clearance profile[2], [42]. This also causes them to have poor solubility issues, which makes it difficult to develop high concentration Ab formulations[43], [44], [45]. Due to the presence of several surface accessible drug conjugation sites, it becomes difficult to reliably control the dosage[46], [47], [48], [49]. Igs have poor thermostability, which hinders their full potential from being developed into therapeutics. Experimental approaches to design novel Abs to bind a target epitope are time and cost expensive, and such techniques rely on random luck for epitope binding specificity. A therapeutic Ab is only effective if it targets the specific epitope of the Ag, which is a matter of chance when molecular biology experiments are concerned. These factors significantly cause an increase in the production cost[2], thereby decreasing the yield. Due to such limitations, the development of conventional therapeutic Ab based drugs face hurdles for many common illnesses. While there are no current alternatives to Abs for immune-activating properties, several other protein binders and non-Ig scaffolds have been developed over the years for their binding properties.

1.5. Non-Ig binders and paradigm shift in protein design

Numerous Ig-like scaffolds such as scFv, Fabs, and Nbs have been developed over the years as potential therapeutic binders[18], [37], [38]. Being smaller compared to an Ig, they have high tumor penetrability. In addition to these engineered Ig-like binders, non-Ig scaffolds (Figure 3), also called Ab mimetics such as Affibodies, Kunitz domains, Affimers, DARPins, Fibronectins, Obodies, etc., unlike Abs, are also small (approx. $1/10^{th}$ the size of an Ab), single-domain proteins that function as therapeutic binders[35], [50], [51], [52]. They often lack disulfide bonds, do not require post-translational modifications, and can undergo direct multimerization. These recombinant proteins display additional characteristics such as high thermostability[53] and

solubility[54], ability to refold post denaturation, improved proteolytic stability, and can be administered in ways other than injections. Such properties are not observed in conventional therapeutic Abs, which makes these non-Ig excellent alternatives. However, the main challenges in developing these binders are serum half-life and tissue penetration. Understandably a relatively tiny scaffold will have a higher tissue penetrability. However, the small size leads to a low renal threshold, which results in them being largely flushed out by the kidneys after a single pass[55]. This requires a higher therapeutic uptake by the subject. To solve this issue, the scaffold can be PEGylated[35], [56] to increase the renal threshold (critical value ≈ 60 kDa), so that it can accumulate at the site of its pharmacological action. If the scaffold has a high affinity to the target, then it will distribute more effectively into a tissue, due to its small size (2-4 kDa). Thus, based on the size of conventional mAbs or Abs, smaller protein binders target tissues faster and more effectively. However, the dosage at which toxicities will occur (therapeutic index) and manufacturing cost are also important factors for developing these therapeutic binders. Multi-specificity, targeting multiple proteins or epitopes on a single protein simultaneously should be considered while designing a binder. This strategy may lead to enhanced neutralization efficacy and it remains a challenge in the field of therapeutics. Therefore, a plethora of factors need to be addressed simultaneously before approaching to design a therapeutic binder.

The point about designing such alternative binders is not that they are superior to Abs in every aspect or can revolutionize the field of biomedical engineering but that they are different. Therefore, the protocols employed for them to be developed into therapeutics are different, and each of them can be used to do different things in the field of therapeutics. Therefore, both of these scaffolds have their advantages and disadvantages, that should be thoroughly examined before developing either of them as fully functional therapeutic binders to bind target antigens.

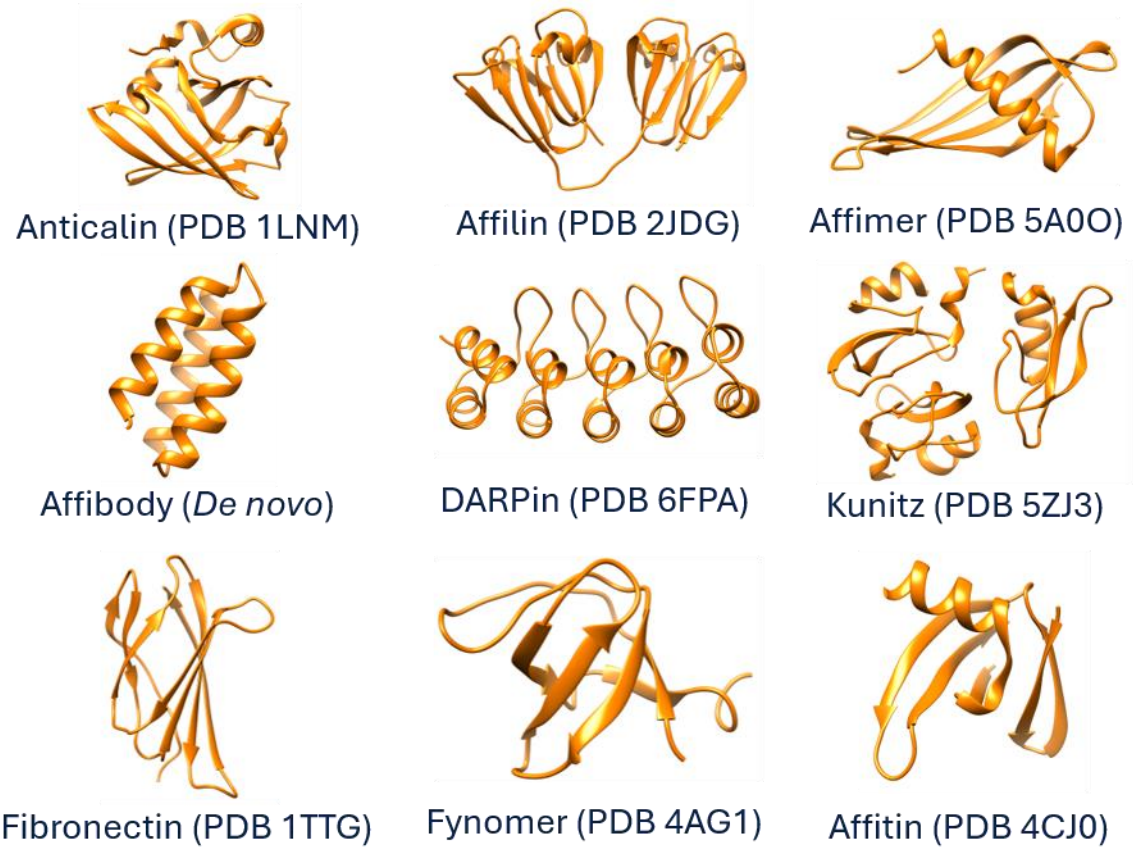

Figure 3. **Alternative scaffolds**. The figure shows a list of 9 antibody mimetics or alternative protein binders with their respective PDB IDs. All images were obtained from the RCSB PDB and viewed in UCSF Chimera. DARPin and Fibronectins are the choice of binders, which were used in the projects of this dissertation.

It is necessary to develop methodologies that will enable us to have control over the specific properties of such therapeutic binders. If the individual properties of such binders can be engineered without the need for immediate molecular biology experiments, then their full potential can be exploited to a significant extent. This gave rise to computational protein engineering and design (CPE) and (CPD), which allow users to design and engineer binding proteins by fine-tuning their properties. Several complex scientific protocols and data analysis methods are employed to shortlist the most promising designs worthy of experiments. However, these computational algorithms are based on general protein biology and chemistry. Therefore, before proceeding into CPE and CPD, one must understand the basics of protein biology.

1.6. Background of proteins

Proteins can be classified into numerous categories as shown in Figure 4, each having its functions and characteristics e.g., enzymes act as catalysts for specific chemical reactions to build (anabolism) or break (catabolism) molecules, hormones are chemical messengers, which coordinate various functions of the body e.g., thymosin secreted by the thymus gland helps develop T-cells[57], collagen strengthens bones, cartilages, ligaments, tendons, and skin [58], hemoglobin[59] acts as the oxygen carrier to the tissues and organs and membrane proteins help in the transport of ions, nutrients, fluids across the cells, and defense proteins such as antibodies. There are several types of proteins and each is responsible for its functioning. Any irregularities in their quantity inside the body can affect bodily functions to a great extent. Thus, they are very significant for the appropriate functioning of living organisms.

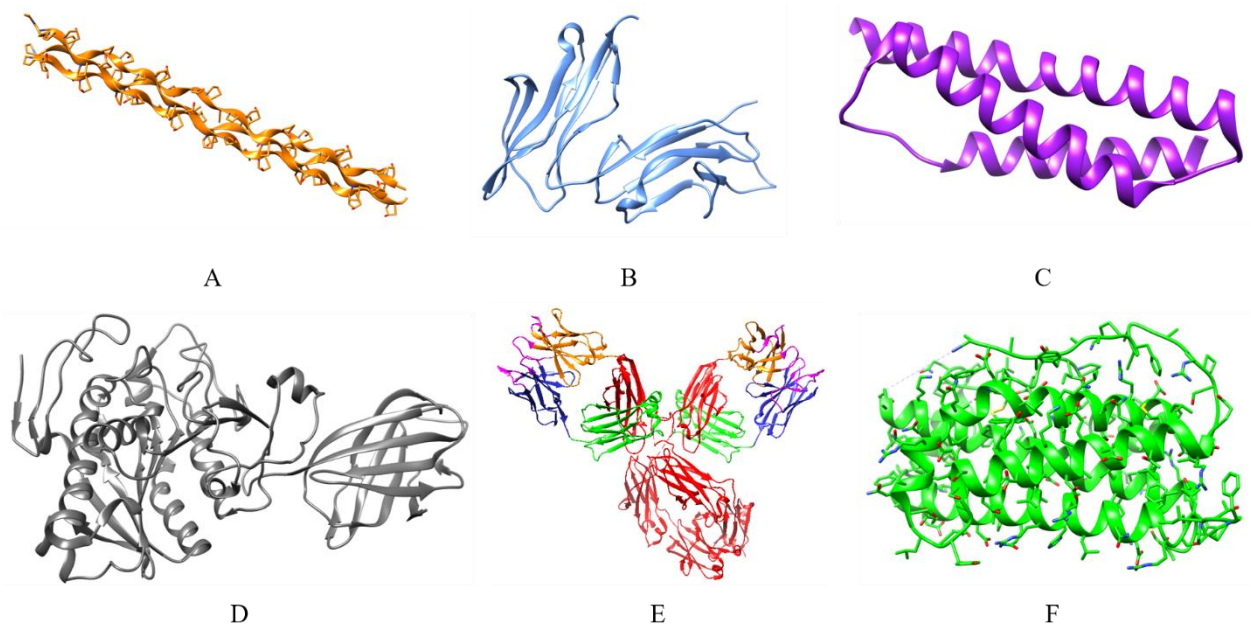

Figure 4. **Example of proteins**. Images of proteins (not drawn to scale) as described previously. A represents collagen (PDB 1CDG). B is a signaling protein or the Fc-gamma receptor III (PDB 1E4J) in this case. C is hemoglobin, which is a transport protein (3OVU). D is pancreatic lipase enzyme (PDB 1N8S). E is a defense protein or a full-length IgG4 antibody (PDB 5DK3) in this case. F is the human growth hormone (PDB 3HHR).

The fundamental unit of a protein is called an amino acid or a residue. There are twenty commonly occurring residues, which act as monomers. They connect in the form of a sequence to form a protein. A residue comprises an alpha carbon atom to a basic amino group ($-NH_2$), an acidic carboxyl group (-COOH), a hydrogen atom, and a side chain (-R for alkyl/aryl/polar/charged). This unique side chain imparts each residue its properties e.g., polarity, solubility, reactivity, hydrophobicity, melting temperature, etc. A condensation reaction occurs when the acidic group of the first amino acid reacts with the basic group of the second amino acid and results in the elimination of a water molecule, which results in the -CO-NH- linkage called the peptide bond. A series of such peptide bonds give rise to the polypeptide chain, and the bond atoms, (i.e., carbon, oxygen, and nitrogen) form the backbone of the polypeptide chain and are therefore called backbone atoms. This polypeptide chain possesses two termini i.e., the one with the amine group is called the N-terminus and the one with the carboxyl group is called the C-terminus.

The principle of protein folding, known as Anfinsen's dogma, after Nobel laureate Christian Anfinsen, is that a protein folding in a suitable environment into its native structure is governed by its amino acid sequence[60]. Thus, the folded state of a protein gives rise to a three-dimensional form and is not just a linear sequence. At the same time, it is known that all the information necessary for a protein to fold to its native state resides in its genetic sequence. This gives rise to the hierarchy of protein structures, namely, primary, secondary, tertiary, and quaternary as shown in Figure 5. The primary structure represents the specific sequence of amino acids arranged together in the polypeptide chain, the secondary structure shows the formation of alpha helices, beta sheets, and coils/loops due to the interatomic interactions within the polypeptide chain. The tertiary structure represents the overall folded structure of a protein in three-dimensional space. Several factors lead to the formation of such a structure like salt bridges, hydrogen bonds, and

disulfide bonds but the most important is the result of a hydrophobic collapse of the molecular core. Proteins also have a quaternary structure, which refers to two or more polypeptide chains in association to form a large functioning unit, so it may not apply to all protein structures. Some common examples are hemoglobin, Abs, or even the SARS-CoV-2 spike protein, a trimer (three repeated units) of a single viral crown-shaped protein. The protein folding mechanism is an interesting phenomenon as it is entirely governed by the thermodynamics of the protein and its surroundings.

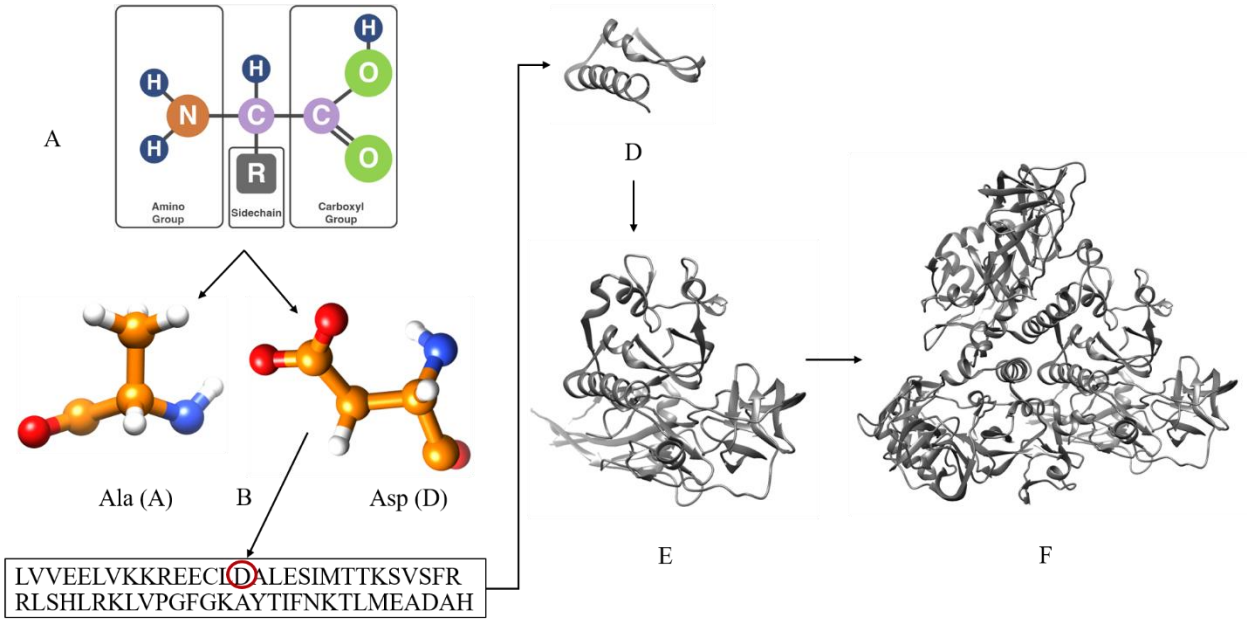

Figure 5. **Amino acids and protein hierarchy**. This figure shows the general structure of a residue followed by two examples and the protein hierarchy. A represents the general structure of an amino acid (Credit: Jazmine Aydee Torres, PROTEIN PANT(z) Group, Auburn University). B is two amino acid examples with their three letter and single letter code, namely, Alanine and Aspartic acid. The sequence of alphabets represents the primary structure. D represents the sequence folding into an alpha helix, which is a secondary structure. E is the complete globular protein or the tertiary structure. F is the trimeric form of the protein or a quaternary structure (PDB 7U9G).

1.7. Protein folding

A protein structure is not in a rigid or static state but always in a state of thermal vibrational motion due to the presence of covalent and electrostatic interactions. When interacting with another protein i.e., hormone-receptor, antibody-antigen, or enzyme-substrate, the proteins

undergo conformational changes due to the interfacial interactions to attain the lowest energy state stable complex. A protein is always surrounded by a solvent, where the protein is considered to be the thermodynamic system and the rest to be the surroundings. Assuming that the protein is completely unfolded in solution, the Surface Accessible Surface Area (SASA) is maximum i.e., maximum value of protein-solvent contact. In this situation, the solvation forces e.g., van der Waals interactions and hydrogen bonds contribute towards the entropy and enthalpy of the protein (system) and the solvent (surroundings). At the time of folding, the non-polar or hydrophobic residues will not interact with the solvent but only the polar and charged ones will do. These hydrophobic residues will begin to coil up with other hydrophobic residues, eventually collapsing into a stable hydrophobic protein core. This decreases the entropy of the protein due to core rigidity and less degrees of freedom. However, at the same time since the solvent molecules previously bound to the polar residues on the protein surface can no longer interact with them, it increases the overall entropy (system + surrounding = universe). The protein slowly stabilizes while the surrounding solvent becomes more chaotic. This results in less SASA since the protein is folding into a three-dimensional structure and no longer in maximum contact with the solvent and for this reason, the $T\Delta S$ in $\Delta G = \Delta H - T\Delta S$ (The symbols signify their usual meaning in the *Gibb's* Free energy equation) would continue to increase, and a higher magnitude of ($-T\Delta S$) would contribute more than the ΔH, which would make ΔG (of the whole system) of folding negative, showing that protein folding is mostly driven by entropy rather than enthalpy. Thus, contrary to some beliefs protein folding is not only dependent on the protein itself but also the solvent, and the entire system as a whole should be considered to prevent this thermodynamically spontaneous process from having an erroneous explanation.

Since proteins try to attain their thermodynamically lowest energy state, there are certain parts of proteins that are stabilized by themselves and are called protein domains e.g., single domain antibodies (sdAbs) or VHH domains. They can attain their lowest energy state and fold into a stabilized structure independently as compared to the entire protein. Taking advantage of this, certain Ab fragments have been genetically engineered and replaced to give rise to novel and improved chimeric proteins[61], [62]. Such proteins have enhanced binding affinity, specificity, and less immunogenicity and have proven to be extremely useful as therapeutics in the medical field. Taking advantage of this, one of these proteins have been used in the projects of this dissertation, which will be discussed later.

1.8. Computational protein engineering and design

A brief knowledge of the fundamentals of protein chemistry should be adequate in proceeding further with CPE and CPD. After considering the limitations of therapeutic Abs, a paradigm shift gave rise to computational protein design. Computational methodologies emerged as a powerful tool to design proteins, which were tested in the laboratory[63], [64], [65], [66]. It paves the way for a direct, inexpensive, and efficient method to reach the goal of a desired protein. The fundamental challenge in rational protein design was to predict the structure of a fully functional protein from the amino acid sequence up until a few years ago. The reason behind this is that proteins are of the order 10^2-10^3 amino acids and as a result, there exist innumerable possibilities for the arrangement of backbones and sidechain alignments, their angles of inclination with each other, bonded and non-bonded interactions, variety of folding and ultimately their globular structures. However, in late 2020, during the 14th Critical Assessment of Techniques for Protein Structure Prediction (CASP14) event, Google Deepmind's AlphaFold predicted the structure of a protein sequence with unprecedented accuracy (GDT of 92.4 out of 100), outperforming all other

methods[67]. This revolutionized the field of protein structure prediction, which has been a problem for decades. Ever since, it has been evolving by adding nearly 214 million protein structures to its database[68] as well as improving the AlphaFold algorithm for structure prediction.

Computational protein engineering (CPE) has evolved to a great extent in the last few decades. It can successfully design protein sequences using RFDiffusion, RosettaDesign, and RAbD[69], [70], [71], predict binding interactions between proteins by a method called protein-protein docking using ZDOCK, Haddock, RosettaDock[72], [73], [74], [75]. It can fine-tune and predict their physico-chemical properties such as solubility using CamSol[76] and thermostability using PoPMuSiC, HoTMuSiC, SCooP[77], [78], [79], [80], [81], and accomplish a plethora of other objectives. Additionally, several methods have been developed to quantify the binding interactions between proteins and other molecules and identify the potential designs to work in the laboratory experiments. These interactions are quantified based on various forms of energies as calculated by molecular mechanics forcefields such as Rosetta[82], AMBER[83], CHARMM[84], and GROMOS[85], to name a few. The fundamental law governing these interactions and therefore protein folding is based on the thermodynamic principle of free energy i.e., the amino acid sequence determines a protein's native structure, which has a minimum Gibbs free energy or in other words, it is stable[86]. The forcefield energy function calculations work based on this free energy i.e., it calculates the change in free energy between the initial and final states of the protein. The change in Gibbs free energy is given in Equation 1, where ΔH represents the enthalpic change in energy whereas $T\Delta S$ represents the change in free energy due to the entropic disorder of the system mostly due to the protein's interaction with water molecules. A detailed explanation of protein interaction with other molecules and water is given in section ***1.2 Protein Folding***.

$$\Delta G = \Delta H - T\Delta S \quad \text{Eq. 1}$$

1.9. Computational resources used in this research

The main source of PDB files was the RCSB PDB [87], home to over 221,000 experimentally determined protein structures and more than a million computationally predicted structures as of July 2024 (https://www.rcsb.org/). This served as the database from which the PDB files were obtained, studied in UCSF Chimera[88], and used as inputs in several applications. A novel tool developed in the PROTEIN PANT(z) Group is called Proteins Designed with Targeted Exceptional Interactions (PETEI) algorithm. The motivation behind developing this algorithm was that hotspot[89], [90], [91], [92] residues are responsible for protein binding interactions, where a hotspot residue is an amino acid that disproportionately contributes a large portion of the protein ligand binding energy. This algorithm designs proteins that bind using modular loops (Abs, sdAbs, scFvs) that can be grafted to the stable protein framework and it can be applied to both Ig and non-Ig protein scaffolds. The entire process can be divided into two parts: a one-time database creation step, where the algorithm searches the RCSB PDB for loops that can be used for grafting into the protein binding loops, and the protein design step where protein ligand complexes are designed based on user input epitope and paratope residues. This *in-silico* algorithm was used to design monobodies/fibronectin[93], [94], [95] to bind three different peptides.

Another novel algorithm developed in the PROTEIN PANT(z) Group is called MutDock [97]. As its name suggests, it is based of docking and mutating residues to find better binding positions. The algorithm identifies favorable binding positions of wild type residues based on the formation of hydrogen bonds since they are the strongest electrostatic interactions between molecules, and mutate residues if the interaction is unfavorable. The precedence of residue mutation is tabulated in the MutDock publication[97]. The algorithm designs the binding positions by pairwise distance

mapping calculations followed by distance alignment between the user-selected paratope and epitope residues. It was used to design DARPin – α-Cobratoxin complexes.

Both CHARMM[84], [96] and Rosetta[98] energy minimization techniques were frequently used in the projects of this dissertation, which was an important step to evaluate and shortlist the designs. There are fundamental differences between the two energy functions and for such reasons, both of them have been used when only one could have been used. Equation 2[96] represents the CHARMM energy function. It comprises the classical equations of physics and the total energy is calculated as a summation of the individual energies that account for its various parameters. The first two terms represent the energy of atoms in the quadratic form of their bond lengths and angles. The third term is a Fourier series that represents bond torsional energy or bond rotations whereas the fourth term is a quadratic form that considers improper shifts in bond angles between atoms resulting in the formation of improper dihedrals. The Urey-Bradley term is present to consider additional atomic interactions or 'stretching' that are not entirely described by dihedrals, angular shifts, etc. It provides the distances between atom pairs (1,3 in this case) with a common atom. The next two can be quite easily recognized to be the well-known terms for Coulomb's law of attraction, which describes electrostatic interactions, and the 6-12 Lennard Jones potential, which describes van der Waals (vdW) interactions. The penultimate term describes the conformational characteristics of the protein backbone for a higher degree of accuracy. In the final term, CHARMM uses the Fast Analytical Continuum Treatment of Solvation (FACTS)[99] to calculate solvation energy.

$$E_{total} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi[1 + \cos(n\varphi - \delta)] +$$

$$\sum_{\substack{improper \\ dihedrals}} K\phi(\phi - \phi_0)^2 + \sum_{Urey-Bradley} K_{UB}\left(r_{1,3} - r_{1,3;0}\right)^2 + \sum_{non-bonded} \frac{q_i q_j}{4\pi D r_{ij}} +$$

$$\varepsilon_{ij}\left[\left(\frac{R_{minij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{minij}}{r_{ij}}\right)^{6}\right] + \sum_{residues} u_{CMAP}(\Phi, \Psi) + E_{solvation} \qquad \text{Eq. 2}$$

The total energy function of the Rosetta forcefield is fairly simple and easy to comprehend. As shown in Equation 3[98], it is a linear combination of many weighted energy functions used over the summation operator.

$$\Delta E_{total} = \sum_i w_i E_i(\Theta_i, aa_i) \qquad \text{Eq. 3}$$

In Eq. 3, w_i represents the weight factor for energy term E_i, which is a function of the geometric degrees of freedom i.e., Θ_i and their respective chemical identities, aa_i. These weighted energy functions are the interactions between atom pairs e.g., the vdW interactions are measured using the 6-12 Lennard Jones potential. Several such individual energy functions are accounted for in this total energy equation, which have been described in great detail in the Rosetta manual[98]. Although these terms explain the type of energy functions utilized by these two forcefields, there is a reason for using both to minimize the designs. CHARMM[96], [100] uses the classical empirical equations of physics as components of its total energy function whereas the Rosetta forcefield incorporates both the empirical equations of physics as well as statistical components[98]. However, the ability of molecular mechanics forcefields to correctly account for entropy is very poor. Thus, none are absolute or can generate accurate values during the analysis of designed proteins. Additionally, CHARMM energy minimization consistently relaxed minor steric clashes that were found in the PETEI and MutDock designed complexes, and Rosetta energy minimization is the most commonly used method to quantify designs in this field. For this reason, both of these forcefields were used in the projects.

Several other well known *in-silico* resources were used in the projects, some of which assisted in down-selecting the most promising designs for experiments. They have been discussed in detail in their corresponding chapters e.g., Visual Molecular Dynamics (VMD)[101] and Nanoscale Molecular Dynamics (NAMD)[102] software was used in the DARPin - α-Cobratoxin project to screen the potential good designs from the bad ones. Their steps have been discussed in detail in that chapter, the results that were generated from the MD simulations, and how the results were analyzed to down-select the designs for experimental testing in the laboratory were all discussed thoroughly.

Other tools such as CamSol server's combination automated designs[103] and solubility predictor[104] were used to enhance protein solubility by incorporating stable and solubility enhancing point mutations and predict their solubility scores before and after the solubility enhancement respectively. Protein thermostability was predicted using SCooP, which is a quick and accurate method for predicting protein thermostability[79], [80], [81] at room temperature. It is based on the Gibbs-Helmholtz equation associated with the protein's folding process. Additionally, it also generates other thermodynamic properties of the folding process such as changes in enthalpy, heat capacity, melting temperature, and free energy. Its database has a collection of 14000 host organisms, which is one of the required user inputs.

Any enhancement in protein thermostability was accomplished using the Dezyme PoPMuSiC[78] and HoTMuSiC[77], [105] servers. PoPMuSiC is an easy-to-use, fast performance web server that can predict the thermodynamic stability changes in proteins due to point mutations based on the solvent accessibility of the mutated residues. HoTMuSiC is another web server similar to PoPMuSiC, and it can quickly and accurately predict the effect of point mutations on the protein melting temperature. This makes it an instrument to manually engineer the thermal

stability of proteins, which can lead to increased heat resistance. The results generated are similar to that of PoPMuSiC and the user can choose the point mutation based on the overall effect of the melting temperature of the protein. This tool can be reliably used for structures with high resolutions, low resolutions, and modeled structures as mentioned in its publication[77]. The accuracy and rapidness of such tools are critical for designing thermally stable proteins and have been described in a detailed manner in the 'ideal' binding proteins chapter.

Immunogenicity is also an important aspect of binding proteins. *De novo* designed proteins must not trigger unnecessary immune responses when injected into a tissue. The Immune Epitope Database (IEDB) Analysis Resource[106] (http://www.iedb.org/) is a list of immunological tools developed at the La Jolla Institute for Immunology, CA that helps in fine-tuning protein immune properties. It does so by using a manually curated database of numerous experimental protein structures unlike other databases, which are comprised of both experimental and computationally predicted structures. Several tools can be used to fine-tune protein properties; however, the one used in this dissertation is the protein De-immunization[107] tool to humanize a protein sequence or reduce its immunogenicity.

Finally, RosettaFold[108] and AlphaFold[67] were used to predict the re-designed structures. AlphaFold2 has completely revolutionized the field of protein folding and the newly modeled or re-designed structures in this dissertation were predicted using the two as previously mentioned. However, Sept 2023 Google Deepmind's AlphaFold released a new algorithm called AlphaMissense[109] in Science that can accurately predict if a mutation in the protein structure has the likelihood of being pathogenic. Unfortunately, due to time constraints, this tool was not used in this dissertation but any updated new algorithms related to this field are worthy of being

mentioned in this dissertation as a source of useful information. There are other tools as well but these tools were majorly used in the projects of this dissertation.

Many of these tools were used to fine-tune the proteins and some of them were also used to shortlist the most promising designs for experiments. The data generated from the DARPin MD simulation results exhibited clear distinction of the designs that should be shortlisted, and the ones that should be discarded. The final structure prediction of the helical bundles using RosettaFold also showed the structures that folded to its native state and the ones that did not. Based on this, the starting point for the next iterative rounds of fine-tuning and shortlisting was the structure that folded. Therefore, just as it is important to generate more predictions to increase the probability of experimental success, it is equally important to make decisions based on scientific rationale and apply sophisticated down-selection process to reduce the experimental sample space. This helps establish a faster and cost-effective technique as opposed to when conducting experiments on all the predictions.

1.10. Brief overview of research projects

The entirety of this doctoral dissertation will mirror the Abstract at the beginning. This chapter is heavily focused on the motivation behind the work that has been done i.e., acknowledging the excellent qualities of therapeutic antibodies, addressing their limitations, introducing *in-silico* methods to predict antigen binding proteins to establish a different technique to bind target antigens, and using scientific rationale and research-based data to apply scientific protocols to shortlist the most promising designs that are most worthy of being experimentally tested. It is practically unfeasible in terms of time, resources and cost expenses to experiment with thousands of predictions unless the sample space is narrowed down to the most promising fewer designs.

This step connects all the dissertation chapters and these protocols will be applied to design and shortlist alternative scaffolds to bind their targets.

Monobodies are engineered 10th fibronectin III domains, which have binding loops that are homologous to Ab CDRs[93], [95], [110]. Chapter 2 will discuss designing thousands of monobodies to serve as transmembrane receptors to bind 3 distinct target epitope peptides HA[111], MYC[112], and FLAG[113] using PETEI. As mentioned previously, the molecular mechanics forcefields CHARMM and Rosetta were used frequently in this dissertation projects. Similarly, they were used to energy minimize designs with fewer than four positive residues in the CDRs to relax the structures. The Rosetta Interface Analyzer calculated and analyzed their binding energy per buried surface area (BE/BSA). This step was repeated for a database of 231 experimental, non-redundant Ab-proteins and the trends of BE vs BSA was found similar to the designs. The p values were the best for the FLAG, followed by the HA and then MYC. The designs were rank-ordered from best to worst based on the BE/BSA trend and the top ranked 30 were selected for experiments. 10 out of 30 designs showed experimental success (33.33%) and it has been submitted for publication as of July 2024.

Chapter 3 will elucidate on using MutDock to design thousands of DARPin – α-Cobratoxin (α-Cbtx) complexes and apply different scientific protocols to down-select the most promising designs for experiments. DARPins are small alternative protein binders with high thermostability, solubility, and lack of Cys residues, easy to express, do not require PTM and are easy to engineer to form drug conjugates for controlled drug delivery[53], [114], [115], [116], [117]. The antigen, α-Cbtx is a lethal component of snake venom found in the *Naja sp.* that can cause neurological failures and local tissue damage due to snakebite envenoming[118], [119], [120]. Once 2145 MutDock designs were generated, they were down-selected to top 39 designs using the same series

of scientific protocols mentioned for Chapter 2, and affinity matured using RosettaDesign to improve their binding affinity. The 3900 designs were further shortlisted to top 73 using both the protocols mentioned previously and visualizing them in UCSF Chimera to find good structural features in the Rosetta affinity matured designs. Short 5 ns MD simulations were conducted to assess the structural integrity of the 73 designs over time and their conformational BE/BSA trajectory were plotted against time. The calculated standard deviations of this metric and the plot characteristics assisted in down-selecting the most promising 51 designs for experiments.

The most important criterion for any protein is its stability. The higher the threshold to withstand mutations, high temperatures (because proteins denature at higher temperatures), possess high solubility, the tendency to refold and other such factors contribute to the stability of a protein. Typically, such a stable structure of a protein sequence that has been scientifically predicted is a helical bundle with minimal loops[121], [122], [123], [124]. Also termed 'ideal' protein structures [121], these structures are a bundle of α-helices that form a close unit linked together by loops or sheets. This gives rise to the solid hydrophobic protein core that massively contributes to the structural stability. If a protein can be designed with such characteristics and additionally, the loops joining these helices be used as protein binding regions then a highly stable protein design can be arrived at that mimics an antibody and is highly stable.

Chapter 4 use tools like CamSol, HoTMuSiC, SCooP, IEDB De-immunization, and RosettaDesign to fine-tune the physicochemical properties of 5 α-helical bundles to arrive at the desired 'ideal' binding protein. Such experimental helical bundles were searched on the RCSB PDB and top 5 were shortlisted from a list of 75 helical bundles. Several factors such as Cys residues, *de novo* designs, extra end loop, and the position of an end loop to cause steric hindrance were considered before discarding these proteins. This was followed by re-designing their

sequences using RosettaDesign to improve solubility. Moreover, properties such as solubility, thermostability, immunogenicity, etc. were thoroughly fine-tuned and optimized in such a way that the final molecule exhibited the desired properties. In between these processes, wherever it was necessary, a choice was made to work on a specific scaffold based on their solubility and stability scores. Since this project dealt with several mutated structures and a single tool was unable to optimize all the properties altogether, decisions had to be made in between the steps before working on a selected design. Once all the properties were fine-tuned, RosettaFold predicted all the engineered designs, out of which only 1 folded with good solubility and thermostability values and a humanized sequence. This was the starting point for the next iterative rounds of fine-tuning to generate 16 different 'ideal' binding proteins. Therefore, observably, the level of complexity of the development and applied scientific protocols implemented in the subsequent chapters keeps increasing progressively, and that leads to a more sophisticated down-selection process for designs worthy of experiments.

Some more projects were included in the fifth chapter. The commonality in all the chapters is that they all deal with single domain protein binders and they connect through a series of complex scientific protocols, employed to shortlist the most promising designs from a large set of thousands. The concluding chapter will focus on the future directions of the current projects.

References

[1] P. S. Huang, S. E. Boyken, and D. Baker, “The coming of age of de novo protein design,” *Nature*, vol. 537, no. 7620, pp. 320–327, 2016, doi: 10.1038/nature19946.

[2] P. Chames, M. Van Regenmortel, E. Weiss, and D. Baty, “Therapeutic antibodies: Successes, limitations and hopes for the future,” *Br J Pharmacol*, vol. 157, no. 2, pp. 220–233, 2009, doi: 10.1111/j.1476-5381.2009.00190.x.

[3] et al. Janeway CA Jr, Travers P, Walport M, “The structure of a typical antibody molecule.,” *Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001.*

[4] H. W. Schroeder and L. Cavacini, “Structure and function of immunoglobulins,” *Journal of Allergy and Clinical Immunology*, vol. 125, no. 2 SUPPL. 2, pp. S41--S52, Feb. 2010, doi: 10.1016/j.jaci.2009.09.046.

[5] C. Rosales and E. Uribe-Querol, “Fc receptors: Cell activators of antibody functions,” *Advances in Bioscience and Biotechnology*, vol. 04, no. 04, pp. 21–33, 2013, doi: 10.4236/abb.2013.44a004.

[6] J. M. Hayes, M. R. Wormald, P. M. Rudd, and G. P. Davey, “Fc gamma receptors: Glycobiology and therapeutic prospects,” *J Inflamm Res*, vol. 9, pp. 209–219, 2016, doi: 10.2147/JIR.S121233.

[7] J. V. R. Kinet and Jean-Pierre, “Fc Receptors,” *Adv Exp Med Biol*, vol. 640, pp. 22–34, 2008, doi: 10.1007/978-0-387-09789-3_3.

[8] V. Schirrmacher, P. Halloran, and C. S. David, “Interactions of Fc receptors with antibodies against Ia antigens and other cell surface components,” *Journal of Experimental Medicine*, vol. 141, no. 5, pp. 1201–1209, 1975, doi: 10.1084/jem.141.5.1201.

[9] F. Nimmerjahn and J. V Ravetch, “Fc$γ$ receptors as regulators of immune responses,” *Nat Rev Immunol*, vol. 8, no. 1, pp. 34–47, 2008, doi: 10.1038/nri2206.

[10] K. B. Yancey and Z. Lazarova, “Complement system,” *Clinical and Basic Immunodermatology*, pp. 105–118, 2008, doi: 10.1007/978-1-84800-165-7_7.

[11] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik, “The complement system and innate immunity - antibodies,” *The complement system and innate immunity*, 2001, [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK27100/figure/A162/?report=objectonly

[12] J. R. Dunkelberger and W. C. Song, “Complement and its role in innate and adaptive immune responses,” *Cell Res*, vol. 20, no. 1, pp. 34–50, 2010, doi: 10.1038/cr.2009.139.

[13] H. Rus, C. Cudrici, and F. Niculescu, “The role of the complement system in innate immunity,” *Immunol Res*, vol. 33, no. 2, pp. 103–112, 2005, doi: 10.1385/IR:33:2:103.

[14] H. P. Peng, K. H. Lee, J. W. Jian, and A. S. Yang, “Origins of specificity and affinity in antibody-protein interactions,” *Proc Natl Acad Sci U S A*, vol. 111, no. 26, 2014, doi: 10.1073/pnas.1401131111.

[15] G. A. Dalkas, F. Teheux, J. M. Kwasigroch, and M. Rooman, “Cation-$π$, amino-$π$, $π$-$π$, and H-bond interactions stabilize antigen-antibody interfaces,” *Proteins: Structure, Function and Bioinformatics*, vol. 82, no. 9, pp. 1734–1746, 2014, doi: 10.1002/prot.24527.

[16] E. De Genst *et al.*, “Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies,” *Proc Natl Acad Sci U S A*, vol. 103, no. 12, pp. 4586–4591, 2006, doi: 10.1073/pnas.0505379103.

[17] D. R. Burton, “Antibody: the flexible adaptor molecule,” *Trends in Biochemical Sciences*. 1990. doi: 10.1016/0968-0004(90)90178-E.

[18] A. Bates and C. A. Power, “David vs. Goliath: The Structure, Function, and Clinical Prospects of Antibody Fragments,” *Antibodies*, vol. 8, no. 2, p. 28, 2019, doi: 10.3390/antib8020028.

[19] Z. A. Ahmad, S. K. Yeap, A. M. Ali, W. Y. Ho, N. B. M. Alitheen, and M. Hamid, “ScFv antibody: Principles and clinical application,” *Clin Dev Immunol*, vol. 2012, 2012, doi: 10.1155/2012/980250.

[20] S. Muyldermans, “Nanobodies: Natural Single-Domain Antibodies,” *Annu Rev Biochem*, vol. 82, no. 1, pp. 775–797, 2013, doi: 10.1146/annurev-biochem-063011-092449.

[21] 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.

[22] J. M. Hayes *et al.*, “Glycosylation and Fc Receptors BT - Fc Receptors,” M. Daeron and F. Nimmerjahn, Eds., Cham: Springer International Publishing, 2014, pp. 165–199. doi: 10.1007/978-3-319-07911-0_8.

[23] J. N. Arnold, M. R. Wormald, R. B. Sim, P. M. Rudd, and R. A. Dwek, “The Impact of Glycosylation on the Biological Function and Structure of Human Immunoglobulins,” *Annu Rev Immunol*, vol. 25, no. 1, pp. 21–50, 2007, doi: 10.1146/annurev.immunol.25.022106.141702.

[24] S. Krapp, Y. Mimura, R. Jefferis, R. Huber, and P. Sondermann, “Structural analysis of human IgG-Fc glycoforms reveals a correlation between glycosylation and structural integrity,” *J Mol Biol*, vol. 325, no. 5, pp. 979–989, 2003, doi: 10.1016/S0022-2836(02)01250-0.

[25] Y. El Abd, A. Tabll, R. Smolic, and M. Smolic, “Mini-review: The market growth of diagnostic and therapeutic monoclonal antibodies – SARS CoV-2 as an example,” *Hum Antibodies*, vol. 30, pp. 15–24, 2022, doi: 10.3233/HAB-211513.

[26] X. Lyu *et al.*, “The global landscape of approved antibody therapies,” *Antib Ther*, vol. 5, no. 4, pp. 233–257, Oct. 2022, doi: 10.1093/abt/tbac021.

[27] R. M. Lu *et al.*, “Development of therapeutic antibodies for the treatment of diseases,” *J Biomed Sci*, vol. 27, no. 1, pp. 1–30, 2020, doi: 10.1186/s12929-019-0592-z.

[28] N. Ayati, P. Saiyarsarai, and S. Nikfar, “Short and long term impacts of COVID-19 on the pharmaceutical sector,” *DARU, Journal of Pharmaceutical Sciences*, vol. 28, no. 2, pp. 799–805, 2020, doi: 10.1007/s40199-020-00358-5.

[29] M. Alessandra Vigano *et al.*, “Correction: DARPins recognizing mTFP1 as novel reagents for in vitro and in vivo protein manipulations (Biology Open (2018) 11 (bio036749) DOI: 10.1242/bio.036749),” *Biol Open*, vol. 7, no. 12, 2018, doi: 10.1242/bio.040832.

[30] M. W. Traxlmayr *et al.*, “Directed evolution of Her2/neu-binding IgG1-Fc for improved stability and resistance to aggregation by using yeast surface display,” *Protein Engineering, Design and Selection*, vol. 26, no. 4, pp. 255–265, 2013, doi: 10.1093/protein/gzs102.

[31] D. Könning and H. Kolmar, “Beyond antibody engineering: Directed evolution of alternative binding scaffolds and enzymes using yeast surface display,” *Microb Cell Fact*, vol. 17, no. 1, pp. 1–17, 2018, doi: 10.1186/s12934-018-0881-3.

[32] U. T. Bornscheuer, B. Hauer, K. E. Jaeger, and U. Schwaneberg, “Directed Evolution Empowered Redesign of Natural Proteins for the Sustainable Production of Chemicals and Pharmaceuticals,” *Angewandte Chemie - International Edition*, vol. 58, no. 1, pp. 36–40, 2019, doi: 10.1002/anie.201812717.

[33] S. Lapetina and H. Gil-Henn, “A guide to simple, direct, and quantitative in vitro binding assays,” *J Biol Methods*, vol. 4, no. 1, p. 62, 2017, doi: 10.14440/jbm.2017.161.

[34] J. A. SPIERS, “Goldberg’s theory of antigen-antibody reactions in vitro.,” *Immunology*, vol. 1, no. 2, pp. 89–102, 1958.

[35] 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.

[36] M. S. Packer and D. R. Liu, “Methods for the directed evolution of proteins,” *Nat Rev Genet*, vol. 16, no. 7, pp. 379–394, 2015, doi: 10.1038/nrg3927.

[37] R. J. Flanagan and A. L. Jones, “Fab antibody fragments: Some applications in clinical toxicology,” *Drug Saf*, vol. 27, no. 14, pp. 1115–1133, 2004, doi: 10.2165/00002018-200427140-00004.

[38] P. Holliger and P. J. Hudson, “Engineered antibody fragments and the rise of single domains,” *Nat Biotechnol*, vol. 23, no. 9, pp. 1126–1136, 2005, doi: 10.1038/nbt1142.

[39] S. Singh *et al.*, “Monoclonal Antibodies: A Review,” pp. 85–99, 2018, doi: 10.2174/1574884712666170809124728.

[40] L. M. Weiner, R. Surana, and S. Wang, “Monoclonal antibodies: Versatile platforms for cancer immunotherapy,” *Nat Rev Immunol*, vol. 10, no. 5, pp. 317–327, 2010, doi: 10.1038/nri2744.

[41] A. Frenzel, M. Hust, and T. Schirrmann, “Expression of recombinant antibodies,” *Front Immunol*, vol. 4, no. JUL, pp. 1–20, 2013, doi: 10.3389/fimmu.2013.00217.

[42] M. P. Deonarain *et al.*, “Small-format drug conjugates: A viable alternative to adcs for solid tumours?,” *Antibodies*, vol. 7, no. 2, pp. 1–19, 2018, doi: 10.3390/antib7020016.

[43] M. Pindrus *et al.*, “Solubility Challenges in High Concentration Monoclonal Antibody Formulations: Relationship with Amino Acid Sequence and Intermolecular Interactions,” *Mol Pharm*, vol. 12, no. 11, pp. 3896–3907, 2015, doi: 10.1021/acs.molpharmaceut.5b00336.

[44] M. Gil-Garcia *et al.*, “Combining Structural Aggregation Propensity and Stability Predictions to Redesign Protein Solubility,” *Mol Pharm*, vol. 15, no. 9, pp. 3846–3859, 2018, doi: 10.1021/acs.molpharmaceut.8b00341.

[45] P. Garidel, A. B. Kuhn, L. V. Schäfer, A. R. Karow-Zwick, and M. Blech, “High-concentration protein formulations: How high is high?,” *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 119, pp. 353–360, Oct. 2017, doi: 10.1016/j.ejpb.2017.06.029.

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

[47] S. Panowski, S. Bhakta, H. Raab, P. Polakis, and J. R. Junutula, “Review 2014-Site-specific Ab drug conjugates for cancer tx,” *MAbs*, vol. 6, no. 1, pp. 34–45, 2014, doi: 10.4161/mabs.27022.

[48] P. Agarwal and C. R. Bertozzi, “Site-specific antibody-drug conjugates: The nexus of bioorthogonal chemistry, protein engineering, and drug development,” *Bioconjug Chem*, vol. 26, no. 2, pp. 176–192, 2015, doi: 10.1021/bc5004982.

[49] A. Wagh, H. Song, M. Zeng, L. Tao, and T. K. Das, “Challenges and new frontiers in analytical characterization of antibody-drug conjugates,” *MAbs*, vol. 10, no. 2, pp. 222–243, 2018, doi: 10.1080/19420862.2017.1412025.

[50] J. Löfblom, F. Y. Frejd, and S. Ståhl, “Non-immunoglobulin based protein scaffolds,” *Current Opinion in Biotechnology*, vol. 22, no. 6. pp. 843–848, Dec. 2011. doi: 10.1016/j.copbio.2011.06.002.

[51] K. Škrlec, B. Štrukelj, and A. Berlec, “Non-immunoglobulin scaffolds: A focus on their targets,” *Trends Biotechnol*, vol. 33, no. 7, pp. 408–418, 2015, doi: 10.1016/j.tibtech.2015.03.012.

[52] X. Yu, Y.-P. Yang, E. Dikici, S. K. Deo, and S. Daunert, “Beyond Antibodies as Binding Partners: The Role of Antibody Mimetics in Bioanalysis”, doi: 10.1146/annurev.

[53] J. Schilling *et al.*, “Thermostable designed ankyrin repeat proteins (DARPins) as building blocks for innovative drugs,” *Journal of Biological Chemistry*, vol. 298, no. 1, p. 101403, 2022, doi: https://doi.org/10.1016/j.jbc.2021.101403.

[54] H. K. Binz, M. T. Stumpp, P. Forrer, P. Amstutz, and A. Plückthun, “Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins,” *J Mol Biol*, vol. 332, no. 2, pp. 489–503, 2003, doi: https://doi.org/10.1016/S0022-2836(03)00896-9.

[55] S. Hober, S. Lindbo, and J. Nilvebrant, “Bispecific applications of non-immunoglobulin scaffold binders,” *Methods*, vol. 154, no. October 2018, pp. 143–152, 2019, doi: 10.1016/j.ymeth.2018.09.010.

[56] F. M. Veronese and A. Mero, “The Impact of PEGylation on Biological Therapies,” *BioDrugs*, vol. 22, no. 5, pp. 315–329, 2008, doi: 10.2165/00063030-200822050-00004.

[57] N. Tao *et al.*, “Thymosin α1 and Its Role in Viral Infectious Diseases: The Mechanism and Clinical Application.,” *Molecules*, vol. 28, no. 8, Apr. 2023, doi: 10.3390/molecules28083539.

[58] M. D. Shoulders and R. T. Raines, “Collagen structure and stability,” *Annu Rev Biochem*, vol. 78, pp. 929–958, 2009, doi: 10.1146/annurev.biochem.77.032207.120833.

[59] J. W. Adamson and C. A. Finch, “Hemoglobin function, oxygen affinity, and erythropoietin.,” *Annu Rev Physiol*, vol. 37, pp. 351–369, 1975, doi: 10.1146/annurev.ph.37.030175.002031.

[60] C. B. Anfinsen, “Principles that govern the folding of protein chains.,” *Science*, vol. 181, no. 4096, pp. 223–230, Jul. 1973, doi: 10.1126/science.181.4096.223.

[61] B. A. Baldo, “Chimeric Fusion Proteins Used for Therapy: Indications, Mechanisms, and Safety,” *Drug Saf*, vol. 38, no. 5, pp. 455–479, 2015, doi: 10.1007/s40264-015-0285-9.

[62] R. K. Campbell, E. R. Bergert, Y. Wang, J. C. Morris, and W. R. Moyle, “Chimeric proteins can exceed the sum of their parts: Implications for evolution and protein design,” *Nat Biotechnol*, vol. 15, no. 5, pp. 439–443, 1997, doi: 10.1038/nbt0597-439.

[63] H. W. Hellinga, “Rational protein design: Combining theory and experiment,” *Proc Natl Acad Sci U S A*, vol. 94, no. 19, pp. 10015–10017, 1997, doi: 10.1073/pnas.94.19.10015.

[64] P. Sormanni, F. A. Aprile, M. Vendruscolo, and P. M. Tessier, “Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins,” *Proc Natl Acad Sci U S A*, vol. 112, no. 32, pp. 9902–9907, 2015, doi: 10.1073/pnas.1422401112.

[65] P. Sormanni, F. A. Aprile, and M. Vendruscolo, “Third generation antibody discovery methods:: In silico rational design,” *Chem Soc Rev*, vol. 47, no. 24, pp. 9137–9157, 2018, doi: 10.1039/c8cs00523k.

[66] C. M. Yu *et al.*, “Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface,” *PLoS One*, vol. 7, no. 3, pp. 1–15, 2012, doi: 10.1371/journal.pone.0033340.

[67] 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.

[68] M. Varadi *et al.*, “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences,” *Nucleic Acids Res*, vol. 52, no. D1, pp. D368–D375, Jan. 2024, doi: 10.1093/nar/gkad1011.

[69] 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.

[70] J. L. Watson *et al.*, “De novo design of protein structure and function with RFdiffusion,” *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023, doi: 10.1038/s41586-023-06415-8.

[71] B. Kuhlman, “Designing protein structures and complexes with the molecular modeling program Rosetta,” *Journal of Biological Chemistry*, vol. 294, no. 50, pp. 19436–19443, 2019, doi: https://doi.org/10.1074/jbc.AW119.008144.

[72] S. Lyskov and J. J. Gray, “The RosettaDock server for local protein-protein docking.,” *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. 233–238, 2008, doi: 10.1093/nar/gkn216.

[73] 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, 2003, doi: 10.1021/ja026939x.

[74] G. C. P. Van Zundert *et al.*, “The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes,” *J Mol Biol*, vol. 428, no. 4, pp. 720–725, 2016, doi: 10.1016/j.jmb.2015.09.014.

[75] 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.

[76] P. Sormanni and M. Vendruscolo, “Protein solubility predictions using the camsol method in the study of protein homeostasis,” *Cold Spring Harb Perspect Biol*, vol. 11, no. 12, Dec. 2019, doi: 10.1101/cshperspect.a033845.

[77] F. Pucci, J. M. Kwasigroch, and M. Rooman, “Protein Thermal Stability Engineering Using HoTMuSiC.,” *Methods Mol Biol*, vol. 2112, pp. 59–73, 2020, doi: 10.1007/978-1-0716-0270-6_5.

[78] Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman, “PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality,” *BMC Bioinformatics*, vol. 12, no. 1, p. 151, 2011, doi: 10.1186/1471-2105-12-151.

[79] F. Pucci, J. M. Kwasigroch, and M. Rooman, “SCooP: an accurate and fast predictor of protein stability curves as a function of temperature.,” *Bioinformatics*, vol. 33, no. 21, pp. 3415–3422, Nov. 2017, doi: 10.1093/bioinformatics/btx417.

[80] F. Pucci and M. Rooman, “Towards an accurate prediction of the thermal stability of homologous proteins.,” *J Biomol Struct Dyn*, vol. 34, no. 5, pp. 1132–1142, May 2016, doi: 10.1080/07391102.2015.1073631.

[81] F. Pucci and M. Rooman, “Stability Curve Prediction of Homologous Proteins Using Temperature-Dependent Statistical Potentials,” *PLoS Comput Biol*, vol. 10, no. 7, p. e1003689, Jul. 2014, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003689

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

[83] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general Amber force field,” *J Comput Chem*, vol. 25, no. 9, pp. 1157–1174, 2004, doi: 10.1002/jcc.20035.

[84] J. Huang *et al.*, “CHARMM36m: an improved force field for folded and intrinsically disordered proteins.,” *Nat Methods*, vol. 14, no. 1, pp. 71–73, Jan. 2017, doi: 10.1038/nmeth.4067.

[85] N. Schmid *et al.*, “Definition and testing of the GROMOS force-field versions 54A7 and 54B7,” *European Biophysics Journal*, vol. 40, no. 7, pp. 843–856, 2011, doi: 10.1007/s00249-011-0700-9.

[86] Y. Fang, “Protein Folding: The Gibbs Free Energy,” pp. 1–18, 2012, [Online]. Available: http://arxiv.org/abs/1202.1358

[87] S. K. Burley *et al.*, “RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D464--D474, 2019, doi: 10.1093/nar/gky1004.

[88] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[89] 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, 2011, doi: 10.1016/j.jmb.2011.09.001.

[90] E. Cukuroglu, H. B. Engin, A. Gursoy, and O. Keskin, “Hot spots in protein–protein interfaces: Towards drug discovery,” *Prog Biophys Mol Biol*, vol. 116, no. 2, pp. 165–173, 2014, doi: https://doi.org/10.1016/j.pbiomolbio.2014.06.003.

[91] T. Kortemme and D. Baker, “A simple physical model for binding energy hot spots in protein-protein complexes,” *Proc Natl Acad Sci U S A*, vol. 99, no. 22, pp. 14116–14121, 2002, doi: 10.1073/pnas.202485799.

[92] B. S. Zerbe, D. R. Hall, S. Vajda, A. Whitty, and D. Kozakov, “Relationship between hot spot residues and ligand binding hot spots in protein-protein interfaces.,” *J Chem Inf Model*, vol. 52, no. 8, pp. 2236–2244, Aug. 2012, doi: 10.1021/ci300175u.

[93] P. G. Chandler and A. M. Buckle, “Development and Differentiation in Monobodies Based on the Fibronectin Type 3 Domain,” *Cells*, vol. 9, no. 3, p. 610, 2020, doi: 10.3390/cells9030610.

[94] P. D. B. Ortholog, P. R. List, I. A. Fn, S. E. Omim, and I. Genecards, “Fibronectin”.

[95] R. Pankov and K. M. Yamada, “Fibronectin at a glance,” *J Cell Sci*, vol. 115, no. 20, pp. 3861–3863, 2002, doi: 10.1242/jcs.00059.

[96] A. Manuscript, “CHARMM.pdf,” vol. 31, no. 4, pp. 671–690, 2011, doi: 10.1002/jcc.21367.CHARMM.

[97] V. M. Chauhan and R. J. Pantazes, “MutDock: A computational docking approach for fixed-backbone protein scaffold design,” *Front Mol Biosci*, vol. 9, no. August, pp. 1–15, 2022, doi: 10.3389/fmolb.2022.933400.

[98] R. F. Alford *et al.*, “HHS Public Access,” vol. 13, no. 6, pp. 3031–3048, 2018, doi: 10.1021/acs.jctc.7b00125.The.

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

[100] A. D. Mackerell, M. Feig, and C. L. Brooks, “Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation,” *J Comput Chem*, vol. 25, no. 11, pp. 1400–1415, 2004, doi: 10.1002/jcc.20065.

[101] W. Humphrey, A. Dalke, and K. Schulten, “VMD: visual molecular dynamics.,” *J Mol Graph*, vol. 14, no. 1, pp. 27-28,33-38, Feb. 1996, doi: 10.1016/0263-7855(96)00018-5.

[102] J. C. Phillips *et al.*, “Scalable molecular dynamics with NAMD,” *J Comput Chem*, vol. 26, no. 16, pp. 1781–1802, Dec. 2005, doi: 10.1002/jcc.20289.

[103] A. Rosace *et al.*, “Automated optimisation of solubility and conformational stability of antibodies and proteins,” *Nat Commun*, vol. 14, no. 1, p. 1937, 2023, doi: 10.1038/s41467-023-37668-6.

[104] P. Sormanni, F. A. Aprile, and M. Vendruscolo, “The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility,” *J Mol Biol*, vol. 427, no. 2, pp. 478–490, 2015, doi: https://doi.org/10.1016/j.jmb.2014.09.026.

[105] F. Pucci, R. Bourgeas, and M. Rooman, “Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC,” *Sci Rep*, vol. 6, no. 1, p. 23257, 2016, doi: 10.1038/srep23257.

[106] R. Vita *et al.*, “The Immune Epitope Database (IEDB): 2018 update.,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D339–D343, Jan. 2019, doi: 10.1093/nar/gky1006.

[107] S. K. Dhanda *et al.*, “Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity,” *Immunology*, vol. 153, no. 1, pp. 118–132, Jan. 2018, doi: 10.1111/imm.12816.

[108] R. Krishna *et al.*, “Generalized biomolecular modeling and design with RoseTTAFold All-Atom.,” *Science*, vol. 384, no. 6693, p. eadl2528, Apr. 2024, doi: 10.1126/science.adl2528.

[109] J. Cheng *et al.*, “Accurate proteome-wide missense variant effect prediction with AlphaMissense,” *Science (1979)*, vol. 381, no. 6664, p. eadg7492, Jun. 2024, doi: 10.1126/science.adg7492.

[110] A. Koide, C. W. Bailey, X. Huang, and S. Koide, “The fibronectin type III domain as a scaffold for novel binding proteins.,” *J Mol Biol*, vol. 284, no. 4, pp. 1141–1151, Dec. 1998, doi: 10.1006/jmbi.1998.2238.

[111] H. Yang, P. J. Carney, J. C. Chang, Z. Guo, J. M. Villanueva, and J. Stevens, “Structure and receptor binding preferences of recombinant human A(H3N2) virus hemagglutinins,” *Virology*, vol. 477, pp. 18–31, 2015, doi: https://doi.org/10.1016/j.virol.2014.12.024.

[112] P. Lavigne, M. P. Crump, S. M. Gagné, R. S. Hodges, C. M. Kay, and B. D. Sykes, “Insights into the mechanism of heterodimerization from the 1H-NMR solution structure of the c-Myc-Max heterodimeric leucine zipper11Edited by P. E. Wright,” *J Mol Biol*, vol. 281, no. 1, pp. 165–181, 1998, doi: https://doi.org/10.1006/jmbi.1998.1914.

[113] D. M. Thal *et al.*, “Crystal structures of the M1 and M4 muscarinic acetylcholine receptors,” *Nature*, vol. 531, no. 7594, pp. 335–340, 2016, doi: 10.1038/nature17188.

[114] R. Tamaskovic, M. Simon, N. Stefan, M. Schwill, and A. Plückthun, *Designed ankyrin repeat proteins (DARPins): From research to therapy*, vol. 503. 2012. doi: 10.1016/B978-0-12-396962-0.00005-7.

[115] A. Plückthun, “Designed ankyrin repeat proteins (DARPins): Binding proteins for research, diagnostics, and therapy,” *Annu Rev Pharmacol Toxicol*, vol. 55, pp. 489–511, 2015, doi: 10.1146/annurev-pharmtox-010611-134654.

[116] C. Zahnd *et al.*, “Efficient tumor targeting with high-affinity designed ankyrin repeat proteins: Effects of affinity and molecular size,” *Cancer Res*, vol. 70, no. 4, pp. 1595–1605, 2010, doi: 10.1158/0008-5472.CAN-09-2724.

[117] 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.

[118] K. T. Yee, M. Pitts, P. Tongyoo, P. Rojnuckarin, and M. C. Wilkinson, “Snake venom metalloproteinases and their peptide inhibitors from Myanmar Russell’s viper venom,” *Toxins (Basel)*, vol. 9, no. 1, 2017, doi: 10.3390/toxins9010015.

[119] H. F. Williams *et al.*, “The urgent need to develop novel strategies for the diagnosis and treatment of snakebites,” *Toxins (Basel)*, vol. 11, no. 6, pp. 1–29, 2019, doi: 10.3390/toxins11060363.

[120] K. Suryamohan *et al.*, “The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins,” *Nat Genet*, vol. 52, no. 1, pp. 106–117, 2020, doi: 10.1038/s41588-019-0559-8.

[121] N. Koga *et al.*, “Principles for designing ideal protein structures,” *Nature*, vol. 491, no. 7423, pp. 222–227, 2012, doi: 10.1038/nature11600.

[122] R. Koga *et al.*, “Robust folding of a de novo designed ideal protein even with most of the core mutated to valine,” *Proc Natl Acad Sci U S A*, vol. 117, no. 49, pp. 31149–31156, 2020, doi: 10.1073/pnas.2002120117.

[123] R. B. Hill, D. P. Raleigh, A. Lombardi, and W. F. Degrado, “De novo design of helical bundles as models for understanding protein folding and function,” *Acc Chem Res*, vol. 33, no. 11, pp. 745–754, 2000, doi: 10.1021/ar970004h.

[124] J. Ferrando and L. A. Solomon, “Recent progress using de novo design to study protein structure, design and binding interactions,” *Life*, vol. 11, no. 3. MDPI AG, pp. 1–23, Mar. 01, 2021. doi: 10.3390/life11030225.

2. *In-silico* designing of monobodies to serve as transmembrane receptors to bind epitope tag peptides

This chapter has been adapted with permission from 'Extracellular Peptide-Ligand Dimerization Actuator Receptor Design for Reversible and Spatially Dosed 3D Cell-Material Communication', by Matthias Recktenwald*, Ritankar Bhattacharya*, Mohammed Mehdi Benmassaoud, James MacAulay, Varun M. Chauhan, Leah Davis, Evan Hutt, Peter A. Galie, Mary M. Staehle, Nichole M. Daringer, Robert J. Pantazes, Sebastián L. Vega. Journal: ACS Synthetic Biology

* These authors contributed equally to the publication

2.1. *Background information*

Transmembrane receptors acts as transporters and enable eukaryotic cellular machinery to transmit signal across the membrane into the cell.[1] Biomaterials research innovations uses relevant chemical and physical reactions to investigate cellular responses and assist in the development of several therapeutic interventions.[2], [3] Biomaterials research has made tremendous advancements in functionalizing biologically relevant materials with small peptide ligands that can initiate biological pathways.[4], [5] However, till today, there has been no development of peptide material-receptor interactions that can act independently of such biological pathways.[6], [7] Conversely, advancements in synthetic biology have displayed the potential of incorporating eukaryotic cells with pre-engineered transmembrane receptors which activate non-native pathways.[8] These receptors have been designed to respond to soluble and cellular-like ligands; however, limited progress has been made in designing synthetic receptors that can respond to ligands.[9], [10] The field synthetic biology and materials science have advanced to a great extent independently, although a popular experts believe that together they can address difficult challenges in biotechnology and health.[7] The development of peptide-ligands to activate such pathways in engineered cells may help in the formation of an alliance in both fields.

Over the past three decades, synthetic biologists successfully developed transmembrane receptors that are able to sense extracellular signals and convert them into cellular responses.[11],

[12] A well-known example of this is the development of CAR (chimeric antigen receptor) T-cell therapy which enables immune cells to identify and fight cancer tissues.[13], [14] Since their development, a plethora of synthetic transmembrane receptors have been developed that can detect extracellular inputs. These receptors are used to target cell bound antigens, which are able to initiate these responses. However, receptors which dimerizes to relay an external signal such as MESA (Modular Extracellular Sensor Architecture)[15], [16], [17] and GEMS (Generalized Extracellular Molecule Sensor)[17] do not require such cues and are better suited for responding to soluble signals such as cytokines. To our knowledge, till date, fluorescent proteins are the only ligands used within biomaterials for activating engineered transmembrane receptors.[18] This is excellent for systems like synNotch to be adapted into biomaterials and biological research.[19], [20] However, these fluorescent ligands can be immunogenic,[21] thereby limiting its full potential of being used as a biomaterial to manipulate cellular machinery programmed with synthetic receptors. Recent advancements in peptide synthesis technology have allowed for relatively cheaper production at scale,[22], [23] and can be easily incorporated to soft and rigid materials with spatiotemporal control via thiol-norbornene click chemistry reactions.[24], [25], [26]

With the goal of creating a transmembrane receptor that can sense and respond to such molecules, epitope tag peptides were selected as ligands. Epitope tags are short, residue sequences used to isolate proteins for manufacturing, or for proteins localization studies.[27] The three distinct peptides being used are the FLAG, MYC, and HA epitope tags.[28] Peptides which include these orthogonal tags can be designed to include a cystine which allows for their use in thiol-norbornene click chemistry reactions. Since endogenous receptors are unable to sense epitope tag peptides, novel receptors were needed to be developed. Advancements in *in-silico* protein design enable us to design a receptor specific to these peptide ligands from scratch, without relying

entirely on the parts found in nature.[29] To direct the search for peptide responsive receptors, PETEI (Proteins Designed with Targeted Exceptional Interactions) was developed and used to design monobodies (based on the 10th fibronectin type III domain),[30] which bind to HA, FLAG, and MYC epitope tag peptide sequences. To evaluate the utility of these peptide ligands for use in transmembrane receptor systems, a push-pull post translational method was developed,[31] and a novel peptide-sensing transmembrane receptor termed EPDA (Extracellular Peptide-Ligand Dimerization Actuator) was created. Upon extracellular activation, intracellularly a reconstitution scheme is adapted that phosphorylates a substrate to bring two fluorescent protein halves together.[32] The system can be deactivated by removing the phosphoryl group, thereby reversible the phosphorylation process.

Once candidate monobodies, based on binding affinity to either HA, FLAG, or MYC epitopes were determined *in silico*, these monobodies were incorporated into the EPDA receptor designs which were tested experimentally. Stimulatory EPDA receptors consist of an HA-specific monobody and an intracellular zipper domain, and a FLAG-specific monobody which includes an intracellular kinase. For the Inhibitory EPDA receptor design, the FLAG-specific monobody and intracellular kinase domain were substituted with a MYC-specific monobody and intracellular phosphatase domain, which removes the phosphoryl group from the activated substrate.

2.2. Methods and materials

2.2.1. PETEI Database Generation

PETEI was used to design monobody binding proteins for HA, FLAG, and MYC epitopes via two main steps: database generation and binder design. The database generation step begins with selecting a scaffold protein. The binding loops that PETEI designs attach to this scaffold and collectively they form a complete protein domain. Once the scaffold was selected all structures in the PDB[33] were searched for loops that geometrically match the attachment points, namely $C\alpha_i$-$C\alpha_f$, $C\beta_i$-$C\beta_f$, $C\alpha_i$-$C\beta_f$, $C\beta_i$-$C\alpha_f$, N_i-N_f, O_i-O_f, and C_i-C_f distances within 0.225 Å of those in the scaffold protein, where 'i' and 'f' refer to the initial and final attachment point residues. The threshold of 0.225 Å was identified through trial and error as a value capable of identifying many loops while still having them appear to align with the attachment points upon visual inspection.

After loops are identified from the PDB, they are scanned for structural compatibility with the scaffold they attach to as well as with the loops for other attachment points. Relatively few monobody structures have been experimentally determined, the monobody domain is homologous to the variable domains of antibodies, and the binding loops of monobodies are structurally equivalent to the Complementarity Determining Regions (CDRs) of antibodies. Therefore, analysis of the CDRs in a non-redundant database of antibody-protein complexes[34] was used to identify quantitative cutoffs for structural compatibility.

The worst 10% of antibody CDRs have 16 heavy atom (i.e., non-hydrogen) steric clashes (i.e., two atoms closer than the sum of their van der Waals (vdW) radii) and an interaction energy of 237.5 kcal/mol with the scaffold of the antibody, as calculated using the electrostatics and vdW components of the CHARMM force field.[69,70] The PETEI-identified loops are limited to no worse than these thresholds to ensure they are comparable to CDRs. Similarly, the worst 10% of CDR

pairs were found to have three heavy atom steric clashes and interaction energies of 55.6 kcal/mol with one another. PETEI-designed loops are considered compatible with one another if they do not exceed these thresholds. Finally, PETEI ensures that every loop is compatible with at least one loop from each of the other attachment points.

After a database of binding loops has been identified, PETEI positions the scaffold and the database of loops to facilitate the later design steps. Specifically, the scaffold is positioned such that the centroid of the loop attachment point is at the origin and the binding loops point up in the positive z-direction. The binding loops are then searched for the spatial coordinates where exceptionally strong binding interactions can occur around them, the feature that inspired PETEI's acronym.

PETEI considers a variety of protein-protein interactions, including hydrogen bonds, salt bridges, π-π stacking, and π-cation interactions. For each of these interactions, there is a thermodynamically-optimal geometry. For example, hydrogen bonds are strongest when there is a straight line from the hydrogen bond donor through the hydrogen to the hydrogen bond acceptor, when that line points at a lone pair of electrons on the acceptor, and when the donor and acceptor are a particular distance apart that depends on the chemical groups of the bond. **Table 1** lists the interactions and corresponding ideal distances that PETEI searches for, where the distances were identified either through analyses of literature data sets or directly from prior literature reports. In addition to the distances, every type of interaction requires at least one additional geometric constraint (e.g., the vector from the hydrogen bond donor through the hydrogen for a hydrogen bond, the plane of an aromatic ring for π interactions, etc.), but the distances were the key parameters that required literature review to determine. Using the Dunbrack rotamer library[35] to consider alternative amino acid side chain positions, the positions of all possible optimal

interactions around each loop in the database are calculated and stored. Note that PETEI does not allow for mutations in the proteins it designs to ensure that the structures of the designed binding loops match their experimentally observed geometries, so the optimal interaction locations are only those for the native amino acid at each position in a loop.

Table 1. Interactions and ideal distances in PETEI search

Interaction	Distance (Å)
A hydrogen bond between an N-H (donor) and O=C (acceptor)[34]	2.89
A hydrogen bond between an N-H (donor) and a HO-C (acceptor) [34]	3.00
A hydrogen bond between an O-H (donor) and a O=C (acceptor) [34]	2.79
A hydrogen bond between an O-H (donor) and a HO-C (acceptor) [34]	2.86
The distances from C to C in an ARG to ASN/GLN salt bridge [34]	3.96
The distance from N to C in a LYS to ASN/GLN bridge [34]	3.18
The distance from the central C in ARG to the center of an aromatic ring in an ARG-π interaction[36]	3.75
The distance from the N in LYS to the center of an aromatic ring in a LYS – π interaction [36]	4.20
The vertical displacements from center of ring to center of ring in a parallel displaced π-π stacking interaction[37]	3.75
The horizontal displacements from center of ring to center of ring in a parallel displaced π-π stacking interaction[38]	1.50
The distance from center of ring to center of ring in a T-shaped π-π interaction[39]	5.00
The distance from the center of histidine ring to the center of benzene ring in a parallel histidine-π interaction [36]	3.50
The distance from the center of histidine ring to the center of benzene ring in a T-shaped histidine-π interaction [36]	4.25

2.2.2. Binder design with PETEI

The second main step of PETEI is designing binding proteins that are predicted to have exceptionally good interactions with no obvious detrimental features. Just as the generation of the database of binding loops began with the user specifying the scaffold protein the loops would attach to, binder design begins with the user specifying the structure of the target protein and identifying the specific amino acids that PETEI should target (i.e., the epitope). The target protein is then oriented so that the epitope is centered above the origin in the x-y plane and points in the negative z-direction (i.e., towards the database of binding loops).

PETEI searches for positions of the target protein and monobody loop combinations that have a sufficient number of strong interactions. The target protein has six degrees of freedom (i.e., three rotational and three translational). PETEI does not consider rotations around the x- and y-axes, as large rotations around these axes would point the target epitope away from the binding pocket of the designed protein. The remaining rotational degree of freedom is addressed in one-degree increments around the z-axis while the translational degrees of freedom are addressed simultaneously. For each rotational increment, the spatial positions where thermodynamically optimal interactions can occur around the epitope are calculated. PETEI then searches for translations of the target protein that match a user-defined minimum number (e.g., four) of complementary interactions (e.g., hydrogen bond donors with acceptors) within a tolerance of 0.33 Å. The tolerance was identified through trial and error as an appropriate balance of finding many combinations of possible interactions while remaining close to the thermodynamic optimum positions. The resulting selected interactions are not thermodynamically optimal but should be exceptionally good versions of the interactions because they are close to that optimum.

Identified monobodies with a sufficient number of exceptional interactions are subsequently screened to eliminate designs with any of three types of detrimental features: steric clashes, charge-charge clashes, and excess positively charged amino acids. Monobody designs with two or more steric clashes involving backbone atoms or atoms in the side chains of the residues forming the exceptional interactions are removed from consideration. For this purpose, a steric clash is defined as two heavy atoms being within 80% of the sum of their vdW radii. Note that PETEI does allow steric clashes among side chains that are not forming the targeted exceptional interactions under the assumption that those side chains would repack to eliminate the clashes. PETEI eliminates designs with any charge-charge clashes, where such clashes are defined as two atoms with the same charge (i.e., positively charged nitrogen atoms of lysine and arginine and the negatively charged oxygen atoms of aspartate and glutamate) within 4.2 Å of one another. PETEI assumes binding is occurring at physiological pH and therefore does not consider histidine as a possible charged amino acid. Finally, PETEI does not consider monobodies with more than three positively charged residues in the binding loops. A preliminary version of PETEI that did not include this constraint was prone to generating designs with large numbers of positively charged residues that appeared qualitatively dissimilar to antibody-protein complexes. While it is interesting to hypothesize that those proteins are still experimentally viable binders, the current version of PETEI aims to generate designs with binding modes that are qualitatively assimilate natural binding interactions.

Successful PETEI designs are considered those with a sufficient number of targeted exceptional interactions and no detrimental features. In the event that a successful design does not have a selection for every binding loop, the database is searched for components that are compatible with the selected loops and do not introduce any detrimental features. PETEI also does

not retain designs with greater than 60% similarity at the level of loop ID with previously identified successful designs. Here, this constraint means that each design has at most one loop that is the same as any other design.

2.2.3. Computational Evaluation of Designs

PETEI generated nearly 2,000 monobody designs, which were more than could be experimentally evaluated in this study. To determine which designs to test, further computational analysis was conducted. All complexes were first minimized using the CHARMM force field[40], [41] with fixed backbone atoms to repack clashing side chains. The complexes were subsequently all-atom minimized using the Rosetta force field and the Rosetta Interface Analyzer was used to calculate predicted binding energies and buried surface areas. CHARMM was used first because it was found to be able to successfully resolve the sometimes substantial side chain steric clashes that PETEI introduces much more frequently than Rosetta. Once CHARMM had resolved those clashes, Rosetta was used to minimize the entire complexes and calculate the properties because it is the most commonly used protein force field. In addition to the PETEI-generated designs, the non-redundant antibody-protein database was prepared using the same protocol to provide a comparison for the calculated values

2.3. Results

2.3.1. PETEI Algorithm

In order to design the receptors specific to these peptide ligands from scratch, we developed and used PETEI (Proteins Designed with Targeted Exceptional Interactions), a novel algorithm capable of identifying candidate monobodies based on epitope binding affinities. PETEI is motivated by the observation that protein binding interactions are commonly understood to have hotspots,[42] where a hotspot is an amino acid that contributes a disproportionately large

percentage of the overall binding energy. We hypothesized that monobodies designed to have strong hotspot interactions without large steric clashes or too many charged residues have a high likelihood of binding to target peptides when experimentally tested.

Monobodies are based on the 10th type III fibronectin type III domain, which is structurally similar to the antigen-binding variable domains of antibodies. Like antibodies, monobodies bind to targets through modular loops that can be exchanged to alter binding affinities. The three binding loops of monobodies are attached to a scaffold that comprises the rest of the monobody domain and is unaltered between different designs. PETEI designs monobodies that feature multiple, thermodynamic binding interactions by generating a database of modular parts followed by selecting loops from those parts that bind to peptides of interest.

Chain A from Protein Data Bank[33] (PDB) file 1TTG[38] was selected as the scaffold for the monobody designs. The three binding loops designed by PETEI replace residues 22-30, 51-55, and 76-87 in file 1TTG. The generated database had 707 structures for loop 1, 708 for loop 2, and 958 for loop 3, for a total combinatorial diversity of approximately 480 million possible monobodies. This database of loops was used to generate monobody designs for HA, FLAG, and MYC peptides.

Residues 58-65 of chain D from PDB file 4WE7 were used as the structure of the HA peptide,[43] all seven residues of chain C from PDB file 5CXV were used as the FLAG peptide,[44] and residue 3-31 of chain A from PDB file 1A93 were used as the MYC peptide.[45] PETEI predicted 372 HA-binding monobodies, 1110 FLAG-binding monobodies, and 582 MYC-binding monobodies. PDB files of the predicted complexes are available in the Supplemental Information. These complexes were then minimized, and their predicted binding energies and

buried surface areas were calculated. The candidate monobodies were ranked for experimental testing using their predicted binding energy per buried surface area.

The distributions of binding energy per buried surface area for PETEI-predicted HA (n = 372), FLAG (n = 1110), and MYC (n = 582) binding monobodies are shown in **Figure 1**. For comparison, the same data for a non-redundant database of antibody-protein complexes (n = 231) is included.[46] Antibodies are used as a comparison because they are the archetypical binding protein and there are far more experimentally determined antibody-protein complexes than those of monobodies. Collectively, FLAG-binding monobodies have significantly better values ($p = 0.0053$), HA-binding monobodies have comparable values ($p = 0.1735$), and MYC-binding monobodies have significantly lower values ($p = 4.7 \times 10^{-11}$) than the naturally occurring antibodies, where significance was assessed using two-tailed t-tests assuming unequal variances. From these, the 10 highest ranking monobody designs for each peptide were selected for experimental testing

To provide a perspective into the binding metrics data that was used as a standard, a total of 231 non-redundant antibody-protein complexes were used. Their binding energy and buried surface areas were calculated and plotted as a curve as shown in Figure 2. The trend clearly showed an increasing trend with a negative slope. This suggests that with an increase in buried surface area, the binding energy of the complexes got stronger. Since a correlation was found between these two metrics, instead of relying on them individually, the two were merged together into a more reliable metric. Additionally, these metrics were obtained from experimentally determined structures. As a result of this, the PETEI designed and minimized complexes were rank ordered based on this dataset standard.

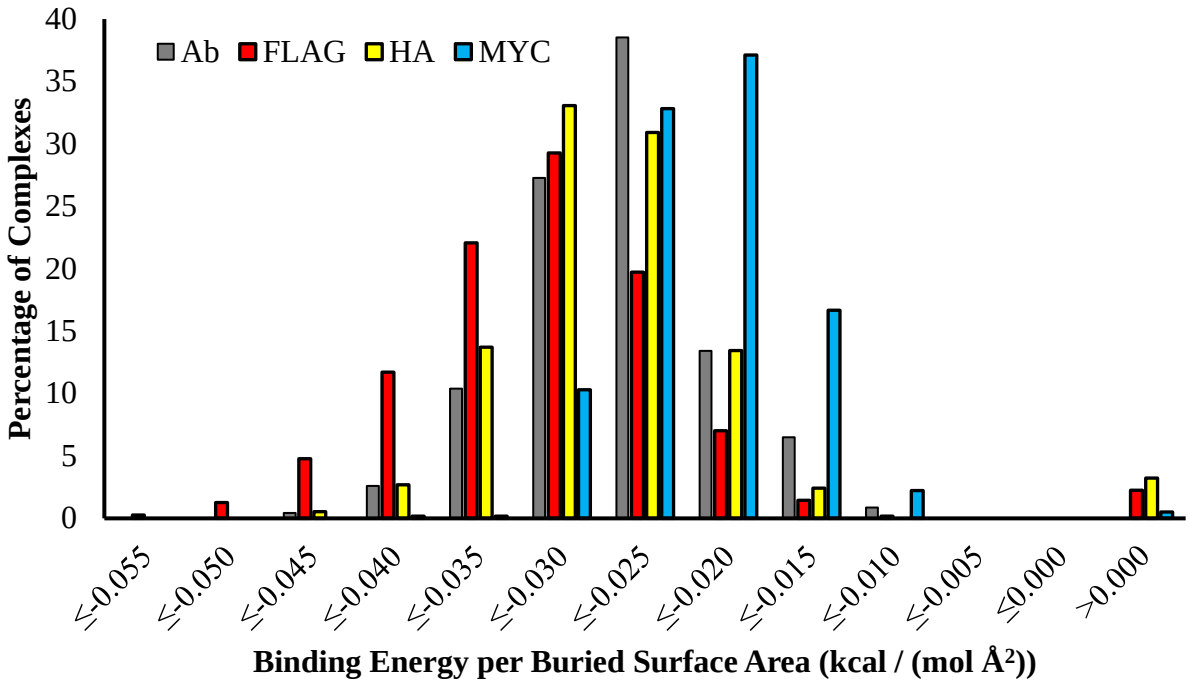

Figure 1. **Binding energies per buried surface area for predicted monobodies.** Histogram of the binding energies per buried surface area for PETEI-predicted HA (yellow), FLAG (red), and MYC (blue) binding monobodies, and naturally occurring antibodies from a non-redundant antibody-protein database (Ab, gray). The FLAG-binding monobodies have significantly better values ($p = 0.0053$), the HA-binding monobodies have comparable values ($p = 0.1735$), and the MYC-binding monobodies have significantly worse values ($p = 4.7 \times 10^{-11}$) than naturally occurring antibodies, where significance was assessed using two-tailed t-tests assuming unequal variances.

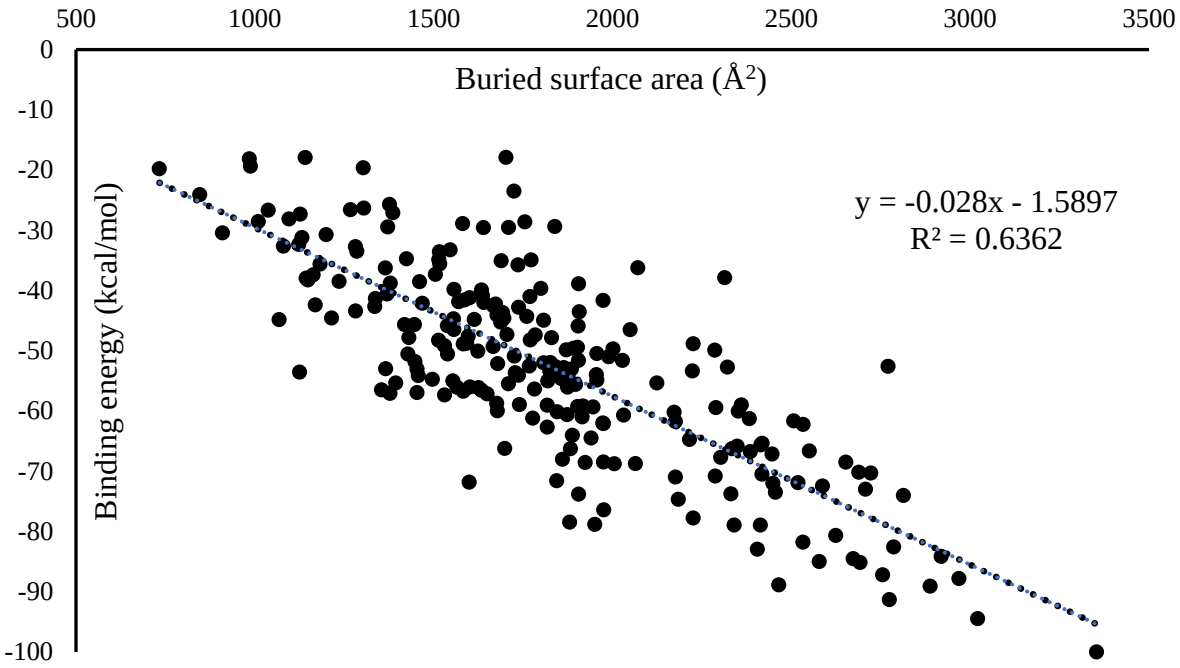

Figure 2. **Predicted Binding Energy versus Buried Surface Area for a Non-redundant Database of Antibody-Protein Complexes**. The data demonstrates that predicted binding energies strengthen (i.e., become more negative) as buried surface area increases. This correlation caused us to rank order designs based on their binding energies per buried surface area for experimental testing.

Figure 3 shows the same binding energy versus buried surface area plot but for the 3 monobody-peptide designed complexes. These curves were kept separate from the antibody-protein database plot because there would have been too many data points in a single plot and the database would not have remained recognizable due to the presence of more than 2000 datapoints of the monobody-peptide complexes. It can be clearly seen that the trends are similar for the designed complexes curves as well with the FLAG-monobodies showing the highest increase in binding energy with an increase in the buried surface area, followed by HA and finally MYC. Even experimentally, FLAG binding results were the best, which was followed by HA and finally MYC. This suggests the significance of designing protein binders based on their structural features.

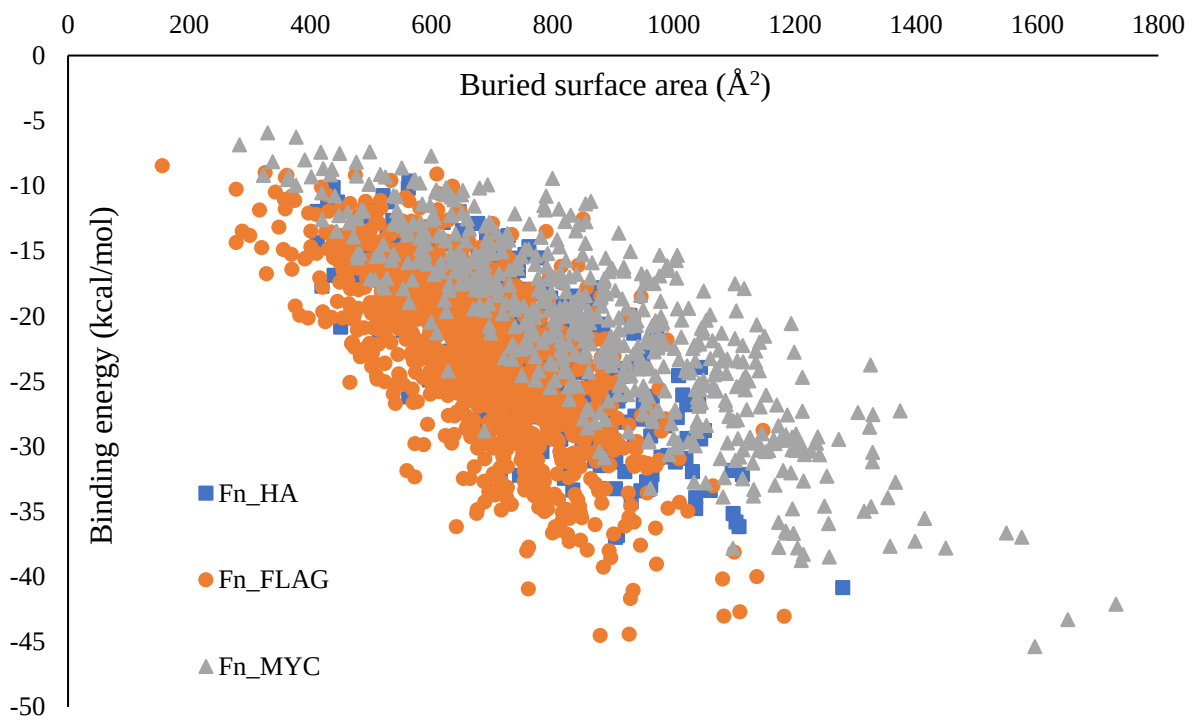

Figure 3. **Monobody-peptide binding metrics plot**. This image shows the trend in binding characteristics of the Monobody-peptide designed complexes. It is similar to the previous plot, which shows that an increase in the BSA also strengthens the BE of the designs. Since, there are nearly 2000 data points in this plot belonging to 3 distinct protein-ligand complexes, it was kept in colored fonts. The font color and shape have been kept different for the better understanding of the readers

The monobodies were setup for fluorescence experiment, with a push-pull translational system as mentioned in the previous section. Once the setup was accomplished, all group showed an increase in fluorescence activity. However, when the 20-residue linker was attached to both the zipper and kinase halves, the cells displayed the maximum fold change. Therefore, this was the first screening on the monobody-peptide receptors. All 10 designs of each monobody-peptide complex were tested. 40% i.e. 4/10 HA monobodies displayed a significant increase in background fluorescence and HA3 design showed 1.92-fold increase in the activity. Similarly, for FLAG monobodies, 50% or 5/10 showed a significant increase in fluorescence and in particular FLAG5 showed 2.61-fold increase in activity. Finally, for MYC monobodies, only MYC7 receptor

displayed increase in fluorescence, with an increase in 34% activity. If the computational results are observed, this is consistent with the computational results where the best binding metrics was exhibited by the FLAG designs, followed by the HA designs and finally the MYC designs.

2.4. Discussion

This study focused on developing a novel protein design algorithm called PETEI, and coming up with protocols to shortlist designs for experiments, although it was a part of a bigger research. Peptide ligands were designed using HA, FLAG, and MYC epitopes, and to identify ligand-receptor pairs, over 2,000 monobodies were built in silico using a novel PETEI algorithm. The designs were minimized using molecular mechanics forcefields CHARMM and Rosetta and their binding metrics were calculated, and rank ordered from best to worst. Monobodies found to have the highest theoretical peptide ligand affinity were tested experimentally, resulting in the identification of three peptide-monobody (HA-HA3, FLAG-FLAG5, MYC-MYC7) pairs. This suggests that although machine learning has progressed to a significant extent; however, rational protein design using a combination of good structural features of the binder and considering the effect of hotspot residues are equally important in the field of computational protein engineering. This further paves the way to design better proteins by selecting appropriate binders, an epitope with bulkier and stabilized residues and always considering the effect of hot-spot residues. The design was followed by experiments were Stimulatory and Inhibitory EPDA receptors that enable cells to respond to ligand peptides were designed. Extracellular linkers were then optimized, to develop final designs for receptors that included an intracellular zipper (H3ZA), kinase (F5KB), and phosphatase (M7PB) domain. This design-build-test-learn cycle culminated in the development of Stimulatory and Inhibitory EPDA receptor pairs, that dimerize in the presence of HA-c-FLAG and HA-c-MYC peptides, respectively. Intracellularly, Stimulatory EPDA receptors

phosphorylate a substrate that merges two fluorescent protein halves, whereas Inhibitory EPDA receptors revert split fluorescent proteins back to their unmerged, inactive state via substrate dephosphorylation.

The algorithm PETEI used here was also a part of Dr. Varun M. Chauhan's dissertation, Class of 2022, Pantazes Research Group (back then) by the name of AUBIE. AUBIE the Tiger is the name of Auburn University's Mascot. As a result of which, the name was changed to PETEI to avoid probable conflicts in the future. The tool was developed in the Pantazes Lab before the advent of machine learning, so it was during that time Ultra-rapid. However, ever since the advent of artificial intelligence and machine learning, such rational design tools are no longer faster than them. This was also another reason to change the name. This acknowledgement is to avoid any future conflicts about the name of the algorithm being different in the publication and in the dissertation.

References

[1] M. Recktenwald *et al.*, “Engineering transcriptional regulation for cell-based therapies,” *SLAS Technol*, p. 100121, Feb. 2024, doi: 10.1016/j.slast.2024.100121.

[2] K. A. Gultian, R. Gandhi, T. W. B. Kim, and S. L. Vega, “Self-Forming Norbornene-Tetrazine Hydrogels with Independently Tunable Properties,” *Macromol Biosci*, vol. 23, no. 3, Mar. 2023, doi: 10.1002/mabi.202200425.

[3] C. Yang, M. W. Tibbitt, L. Basta, and K. S. Anseth, “Mechanical memory and dosing influence stem cell fate,” *Nat Mater*, vol. 13, no. 6, pp. 645–652, 2014, doi: 10.1038/nmat3889.

[4] C. A. Deforest and D. A. Tirrell, “A photoreversible protein-patterning approach for guiding stem cell fate in three-dimensional gels,” *Nat Mater*, vol. 14, no. 5, pp. 523–531, May 2015, doi: 10.1038/nmat4219.

[5] C. Ligorio and A. Mata, “Synthetic extracellular matrices with function-encoding peptides,” *Nature Reviews Bioengineering*, vol. 1, no. 7, pp. 518–536, Apr. 2023, doi: 10.1038/s44222-023-00055-3.

[6] E. M. Bressler, S. Adams, R. Liu, Y. L. Colson, W. W. Wong, and M. W. Grinstaff, “Boolean logic in synthetic biology and biomaterials: Towards living materials in mammalian cell therapeutics,” *Clin Transl Med*, vol. 13, no. 7, Jul. 2023, doi: 10.1002/ctm2.1244.

[7] A. P. Liu *et al.*, “The living interface between synthetic biology and biomaterial design,” *Nat Mater*, vol. 21, no. 4, pp. 390–397, Apr. 2022, doi: 10.1038/s41563-022-01231-3.

[8] C. Zhang, J. Liu, J. F. Zhong, and X. Zhang, “Engineering CAR-T cells,” *Biomarker Research*, vol. 5, no. 1. BioMed Central Ltd., 2017. doi: 10.1186/s40364-017-0102-y.

[9] C. J. Bashor, I. B. Hilton, H. Bandukwala, D. M. Smith, and O. Veiseh, “Engineering the next generation of cell-based therapeutics,” *Nature Reviews Drug Discovery*, vol. 21, no. 9. Nature Research, pp. 655–675, Sep. 01, 2022. doi: 10.1038/s41573-022-00476-6.

[10] E. Zah *et al.*, “Systematically optimized BCMA/CS1 bispecific CAR-T cells robustly control heterogeneous multiple myeloma,” *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-16160-5.

[11] F. Meng and T. Ellis, “The second decade of synthetic biology: 2010–2020,” *Nature Communications*, vol. 11, no. 1. Nature Research, Dec. 01, 2020. doi: 10.1038/s41467-020-19092-2.

[12] H. Abken, “Building on Synthetic Immunology and T Cell Engineering: A Brief Journey through the History of Chimeric Antigen Receptors,” *Hum Gene Ther*, vol. 32, no. 19–20, pp. 1011–1028, Oct. 2021, doi: 10.1089/hum.2021.165.

[13] R. M. Young, N. W. Engel, U. Uslu, N. Wellhausen, and C. H. June, “Next-Generation CAR T-cell Therapies,” *Cancer Discov*, vol. 12, no. 7, pp. 1625–1633, Jul. 2022, doi: 10.1158/2159-8290.CD-21-1683.

[14] S. Gill and C. H. June, “Going viral: chimeric antigen receptor T-cell therapy for hematological malignancies,” 2014.

[15] K. A. Schwarz, N. M. Daringer, T. B. Dolberg, and J. N. Leonard, “Rewiring human cellular input-output using modular extracellular sensors,” *Nat Chem Biol*, vol. 13, no. 2, pp. 202–209, Feb. 2017, doi: 10.1038/nchembio.2253.

[16] N. M. Daringer, R. M. Dudek, K. A. Schwarz, and J. N. Leonard, “Modular Extracellular sensor architecture for engineering mammalian cell-based devices,” *ACS Synth Biol*, vol. 3, no. 12, pp. 892–902, Dec. 2014, doi: 10.1021/sb400128g.

[17] L. Scheller, T. Strittmatter, D. Fuchs, D. Bojar, and M. Fussenegger, “Generalized extracellular molecule sensor platform for programming cellular behavior article,” *Nat Chem Biol*, vol. 14, no. 7, pp. 723–729, Jul. 2018, doi: 10.1038/s41589-018-0046-z.

[18] M. Garibyan *et al.*, “Engineering Programmable Material-To-Cell Pathways Via Synthetic Notch Receptors To Spatially Control Cellular Phenotypes In Multi-Cellular Constructs”, doi: 10.1101/2023.05.19.541497.

[19] L. Morsut *et al.*, “Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors,” *Cell*, vol. 164, no. 4, pp. 780–791, Feb. 2016, doi: 10.1016/j.cell.2016.01.012.

[20] K. T. Roybal *et al.*, “Engineering T Cells with Customized Therapeutic Response Programs Using Synthetic Notch Receptors,” *Cell*, vol. 167, no. 2, pp. 419-432.e16, Oct. 2016, doi: 10.1016/j.cell.2016.09.011.

[21] R. Stripecke, C. Villacres, D. C. Skelton, N. Satake, S. Halene, and D. B. Kohn, “Immune response to green fluorescent protein: implications for gene therapy,” 1999. [Online]. Available: http://www.stockton-press.co.uk/gt

[22] R. Behrendt, P. White, and J. Offer, “Advances in Fmoc solid-phase peptide synthesis,” *Journal of Peptide Science*, vol. 22, no. 1. John Wiley and Sons Ltd, pp. 4–27, Jan. 01, 2016. doi: 10.1002/psc.2836.

[23] H. Masui and S. Fuse, “Recent Advances in the Solid-and Solution-Phase Synthesis of Peptides and Proteins Using Microflow Technology,” *Organic Process Research and Development*, vol. 26, no. 6. American Chemical Society, pp. 1751–1765, Jun. 17, 2022. doi: 10.1021/acs.oprd.2c00074.

[24] W. M. Gramlich, I. L. Kim, and J. A. Burdick, “Synthesis and orthogonal photopatterning of hyaluronic acid hydrogels with thiol-norbornene chemistry,” *Biomaterials*, vol. 34, no. 38, pp. 9803–9811, Dec. 2013, doi: 10.1016/j.biomaterials.2013.08.089.

[25] S. L. Vega *et al.*, “Combinatorial hydrogels with biochemical gradients for screening 3D cellular microenvironments,” *Nat Commun*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/s41467-018-03021-5.

[26] M. Recktenwald *et al.*, “Facile method for covalently binding peptides onto polycaprolactone films and nanofibers,” *Mater Lett*, vol. 365, Jun. 2024, doi: 10.1016/j.matlet.2024.136432.

[27] I. Segal, D. Nachmias, A. Konig, A. Alon, E. Arbely, and N. Elia, “A straightforward approach for bioorthogonal labeling of proteins and organelles in live mammalian cells, using a short peptide tag,” *BMC Biol*, vol. 18, no. 1, Jan. 2020, doi: 10.1186/s12915-019-0708-7.

[28] D. W. Wood, “New trends and affinity tag designs for recombinant protein purification,” *Current Opinion in Structural Biology*, vol. 26, no. 1. Elsevier Ltd, pp. 54–61, 2014. doi: 10.1016/j.sbi.2014.04.006.

[29] N. Ferruz, M. Heinzinger, M. Akdel, A. Goncearenco, L. Naef, and C. Dallago, “From sequence to function through structure: Deep learning for protein design,” *Computational and Structural Biotechnology Journal*, vol. 21. Elsevier B.V., pp. 238–250, Jan. 01, 2023. doi: 10.1016/j.csbj.2022.11.014.

[30] A. Koide, C. W. Bailey, X. Huang, and S. Koide, “The Fibronectin Type III Domain as a Scaffold for Novel Binding Proteins.”

[31] X. Yang *et al.*, “Engineering synthetic phosphorylation signaling networks in human cells”, doi: 10.1101/2023.09.11.557100.

[32] Leah Davis *et al.*, “Synthetic phosphorylation networks with fluorescence and luminescence expansion for Theranostic applications,” *Submitted to ACS Synthetic Biology*, 2024.

[33] H. M. Berman *et al.*, “The Protein Data Bank,” *Acta Crystallogr D Biol Crystallogr*, vol. 58, no. 6, pp. 899–907, Jun. 2002, doi: 10.1107/S0907444902003451.

[34] V. M. Chauhan and R. J. Pantazes, “Analysis of conformational stability of interacting residues in protein binding interfaces,” *Protein Engineering, Design and Selection*, vol. 36, Jan. 2023, doi: 10.1093/protein/gzad016.

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

[36] 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,” *Chem Sci*, vol. 9, no. 10, pp. 2655–2665, 2018, doi: 10.1039/C7SC04905F.

[37] C. Chipot, R. Jaffe, B. Maigret, D. A. Pearlman, and P. A. Kollman, “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,” *J Am Chem Soc*, vol. 118, no. 45, pp. 11217–11224, Nov. 1996, doi: 10.1021/ja961379l.

[38] A. L. Main, T. S. Harvey, M. Baron, J. Boyd, and I. D. Campbell, “The three-dimensional structure of the tenth type III module of fibronectin: An insight into RGD-mediated interactions,” *Cell*, vol. 71, no. 4, pp. 671–678, Nov. 1992, doi: 10.1016/0092-8674(92)90600-H.

[39] M. O. Sinnokrot and C. D. Sherrill, “Highly Accurate Coupled Cluster Potential Energy Curves for the Benzene Dimer: Sandwich, T-Shaped, and Parallel-Displaced Configurations,” *J Phys Chem A*, vol. 108, no. 46, pp. 10200–10207, Nov. 2004, doi: 10.1021/jp0469517.

[40] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, “A program for macromolecular energy, minimization, and dynamics calculations,” *J Comput Chem*, vol. 4, no. 2, pp. 187–217, Jun. 1983, doi: 10.1002/jcc.540040211.

[41] 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.

[42] A. A. Bogan and K. S. Thorn, “Anatomy of Hot Spots in Protein Interfaces.” [Online]. Available: http://motorhead.

[43] H. Yang, P. J. Carney, J. C. Chang, Z. Guo, J. M. Villanueva, and J. Stevens, “Structure and receptor binding preferences of recombinant human A(H3N2) virus hemagglutinins,” *Virology*, vol. 477, pp. 18–31, Mar. 2015, doi: 10.1016/j.virol.2014.12.024.

[44] D. M. Thal *et al.*, “Crystal structures of the M1 and M4 muscarinic acetylcholine receptors,” *Nature*, vol. 531, no. 7594, pp. 335–340, Mar. 2016, doi: 10.1038/nature17188.

[45] P. Lavigne, M. P. Crump, S. M. Gagné, R. S. Hodges, C. M. Kay, and B. D. Sykes, “Insights into the mechanism of heterodimerization from the 1H-NMR solution structure of the c-Myc-Max heterodimeric leucine zipper,” *J Mol Biol*, vol. 281, no. 1, pp. 165–181, Aug. 1998, doi: 10.1006/jmbi.1998.1914.

[46] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” 2018, pp. 2113–2118. doi: 10.1016/B978-0-444-64241-7.50347-5.

3. Application of *in-silico* tools to design DARPin – α-Cobratoxin complexes

3.1 Background information

Countless diseases have plagued human society since time immemorial, which required constant attention and remedies to be treated. Human beings have always found scientific breakthroughs to develop a variety of medical therapies to treat ailments such as the Bubonic plague, Spanish Flu, SARS, COVID-19, and HIV/AIDS, and prevent them from gaining severity. However, amidst the list of ailments, venomous snakebite is often overlooked. Although venomous snakes are omnipresent across the world and pose significant health hazards, they are seldom considered to be a health risk. The toxins can create a concoction of physical and neurological disruptions in humans and animals, resulting in severe disorders, which may also be fatal at times. Unfortunately, the impact of snakebite envenoming on human health is often underestimated and for this reason, quick and effective treatment remains vastly underexplored. For these reasons, snakebites remain one of the leading causes of death in several nations of the world[1], [2], [3], [4], [5].

The α-Cobratoxin (α-Cbtx) is an extremely potent neurotoxin found in the venom of several cobra species (*Naja naja/kaouthia/oxiana/siamensis)*, which has a high affinity for nicotinic acetylcholine receptors (nAChRs) and subtypes[6], [7], [8], [9] as shown in Figure 1. These receptors are critical for the transmission of neuromuscular signals. The α-Cbtx binds to these molecules with high affinity, and blocks neural transmission between the nerves and muscles, thereby causing respiratory and renal malfunctioning, partial or complete paralysis, local tissue death, organ failure, blurred vision, and a plethora of other critical symptoms[10], [11]. It belongs to the family of three-fingered toxins (3Ftx), which are relatively flat molecules with a small hydrophobic core that gives rise to ‘three fingers’[12]. The arrangement of important amino acids

to form of folds and loops is specifically responsible for interacting with the nicotinic proteins and subtypes.

Currently, the production of antivenom is focused on animal immunization, where horses or sheep are injected with small quantities of the venom over months to gradually develop immunity by generating antivenom antibodies (Abs)[1]. The Abs generated must match the snake species toxin to temporarily bind them in the blood and buy the victim some time for the appropriate medical care. There may arise several challenges in treating venomous snakebites e.g., Abs have low diffusability to penetrate localized tissues and bind the toxin[13], [14], and if not administered quickly, the antivenom is rendered ineffective due to the rapid progression of the toxin that causes tissue destruction. Unfortunately, the entire system is rigged with an inefficient, slow, and laborious cocktail challenges[1], [11].

Although there have been some recent developments in using peptide inhibitors[15], [16], [17], small molecule therapies (SMT) such as using single chain variable fragments (scFvs)[18], nanobodies (Nbs)[19] or sdAbs (single domain antibodies), they are still not as easily available in the market as Ab, since a lot more research is needed for their approval. There is a paucity of effective therapies against a malady, which has been and is still being neglected in many nations across the globe. Hence, there is an urgent call to find highly effective, quick, and alternative strategies to treat snakebite envenoming.

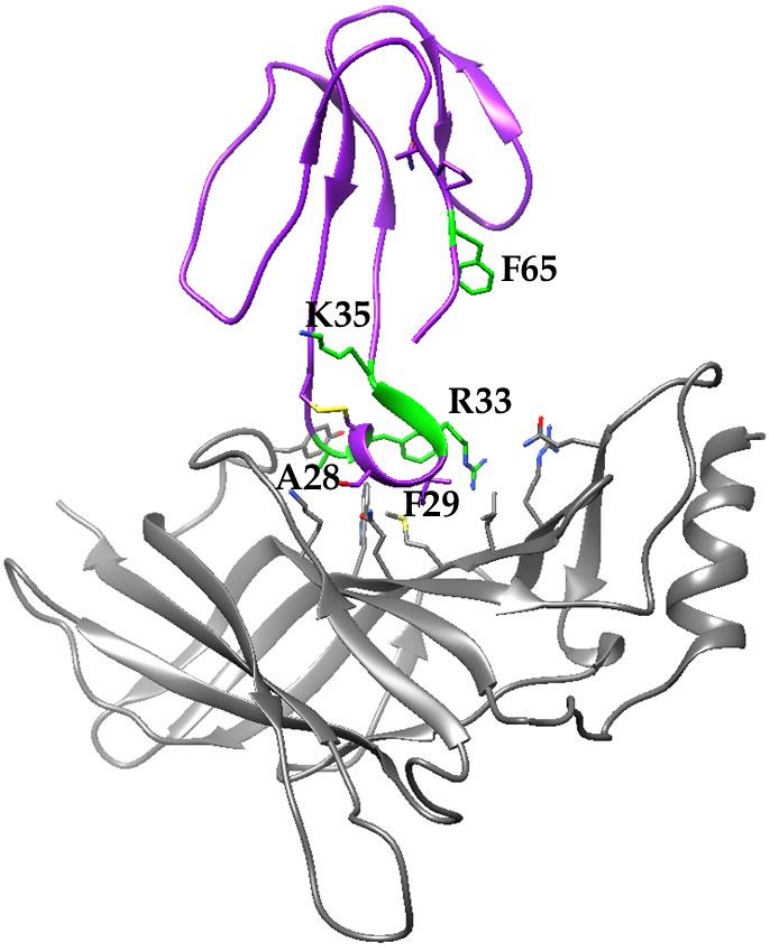

Figure 1. **α-Cobratoxin in complex with AChBP**. This complex was extracted from PDB 1YI5 and observed in UCSF Chimera. The purple molecule is α-Cbtx, with some of its key binding residues highlighted in green. It is interacting with an acetylcholine binding protein (AChBP), which is shaded in light gray.

Designed Ankyrin Repeat Proteins (DARPins) as shown in Figure 2 is a class of engineered alternative protein binders that are gaining prominence due to their properties and therapeutic promise [20], [21], [22]. These molecules can specifically bind antigens with high affinity although structurally dissimilar to Abs[23]. For this reason, they are also called Ab mimetics[13]. Typically, a DARPin is about 1/10th the size of a conventional Ab or 14-15 kDa, generally consisting of 3-4 repeat modules or motifs with its hydrophobic core between its N and C terminus[20], [23], [24], [25]. Unlike Abs which uses modular binding loops called complementarity determining regions (CDRs)[26], [27], [28], to bind Ags, DARPins use fixed backbone binding surfaces to bind

proteins[29]. A key difference in improve their binding affinity is that antibody CDRs can be wholly grafted whereas, a common method for DARPins is to dock them to target Ags followed by mutating their residue sidechains[29] to accomplish the same. Such Ab CDR-like modularity in their binding surfaces is not a common observation nor is it a strategy to improve their properties.

Innately by themselves, DARPins are highly soluble and thermostable molecules, can be easily expressed in prokaryotic cells like *E. coli*, and do not require post-translational modifications (PTM) like Abs[30], [31]. This significantly reduces production costs [20], [32]. Due to their small size, they have high tumor penetrability and are potentially effective therapeutics for cancer treatment[20], [21], [31]. The presence of Cysteine residue serves as a site for drug conjugation to assist in drug dosage[33]. Often, the presence of multiple surface exposed Cys residues can lead to uncontrolled drug dosage. However, DARPins have the advantage of not having any surface exposed Cys residue and therefore they can be easily engineered to have a singular drug conjugation site[20], [31]. They can further be PEGylated[22], [34] to increase their renal thresholds without altering their potency. Such properties make these molecules high-potential therapeutic binders. For these reasons, DARPins were the choice for protein binders to target α-Cbtx.

The advent of computational protein engineering and design (CPE and CPD) has revolutionized the concept of *de novo* Ab and other therapeutic binder design[14], [35], [36], [37], [38], [39], [40], [41]. The PROTEIN PANT(z) Group at Auburn University developed another novel algorithm called MutDock[42], which as its name suggests is a dock, and mutate tool to find good binding positions for such alternative binders. Thus, MutDock was used to design several DARPins in complex with α-Cbtx in the early phase of this project. This was followed by Rosetta

affinity maturation (AffMat) to improve the binding affinity of the DARPins further. CHARMM forcefield[43], [44], and Rosetta forcefield energy minimizations [45] were performed to relax the existing strains in the designs, and Rosetta Interface Analyzer (RIA)[46] was used to calculate the binding metrics needed to quantify the designs. Visual Molecular Dynamics (VMD)[47] and Nanoscale Molecular Dynamics (NAMD) software[48] MD simulations were performed to further assess structural stability over time and shortlist the good designs from the bad ones. UCSF Chimera[49] is a molecular visualization tool, which was used to observe the designs and identify characteristics that are detrimental to complex stability and discard them.

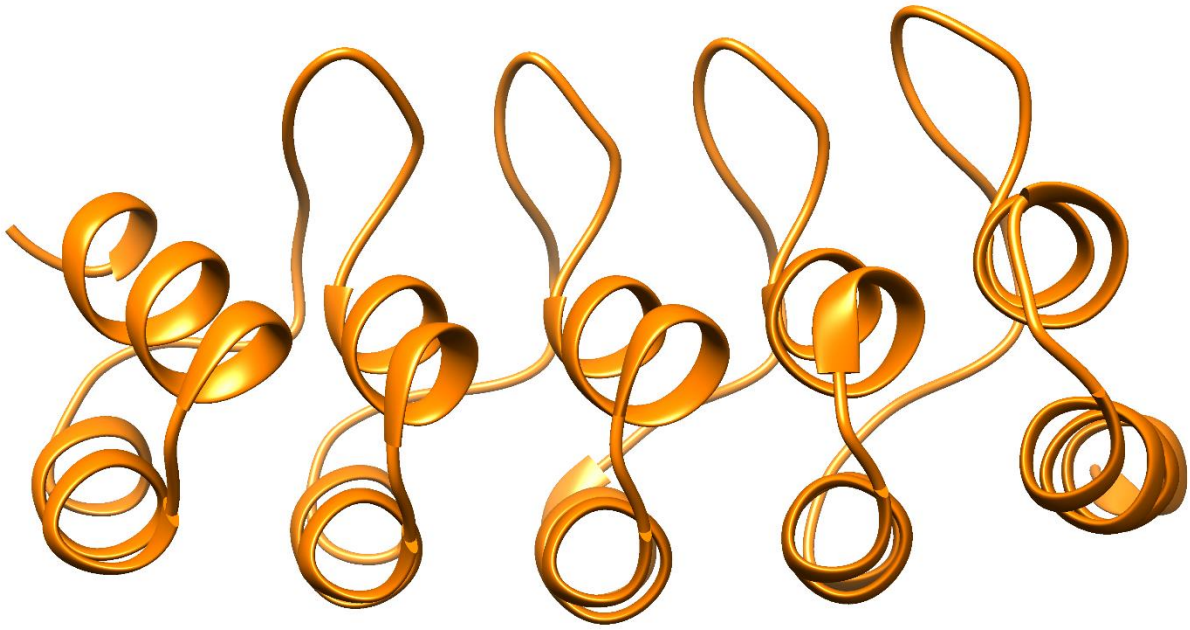

Figure 6. **Full crystallized DARPin 6FPA**. This image is a pictorial representation of a crystallized structure of PDB 6FPA, Designed Ankyrin Repeat Protein (DARPin) molecule as observed in UCSF Chimera.

3.2 Methods

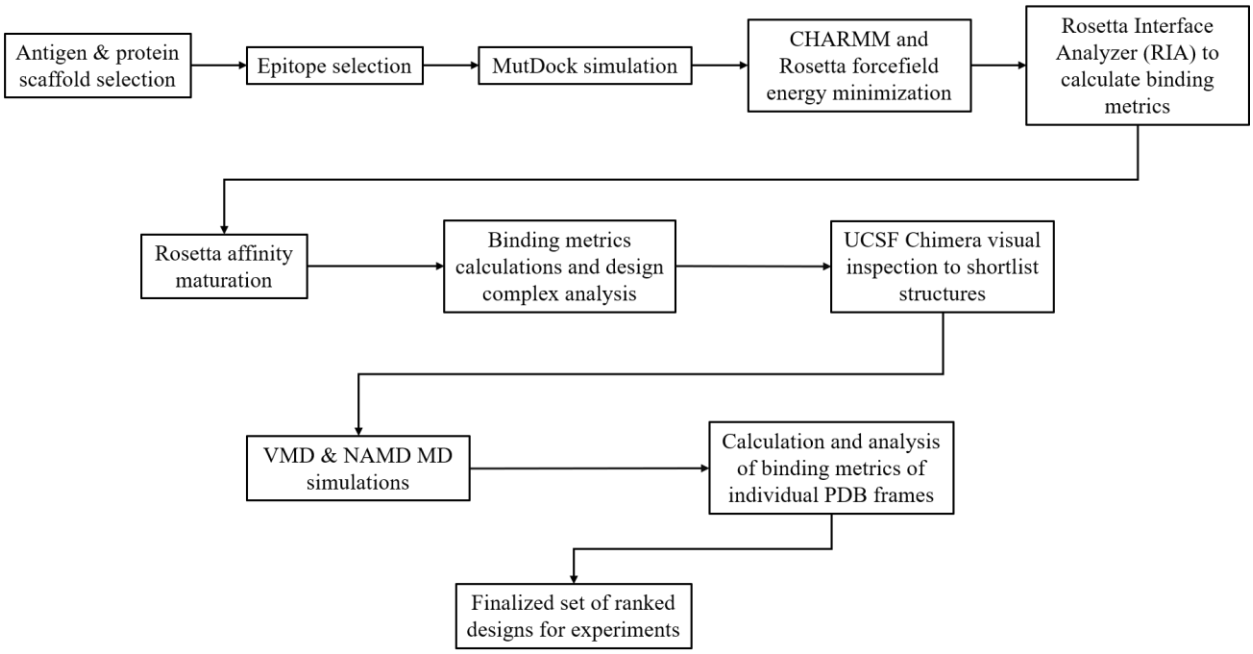

Figure 3. **Hierarchical flow of information of the entire methods section**. This image shows the chronology of the method section for better comprehension.

3.2.1. Finding the appropriate protein scaffold and target antigen

A protein binder and a target antigen of choice were needed as inputs for the design simulation. The well-known RCSB PDB[50] was explored for DARPins, and PDB 6FPA[51] was selected as the protein scaffold. It was a non-mutated, 1.58Å structure, which was of the highest quality back then and it had no missing residues. For the target antigen α-Cbtx, it was necessary to find a structure, which was in complex with a binder (protein/peptide) to locate the target epitope. Other factors considered were sequence identity, resolution, mutated residues, cobra species, etc. The structure PDB 6ZFM was selected, which was Cbtx bound to peptide inhibitors. It was a 1.90Å structure derived from *Naja kaouthia*, with no mutations or missing residues.

3.2.2. MutDock simulation

A novel tool called MutDock[42] was developed in the PROTEIN PANT(z) Lab to design good binding poses. As its name suggests it comprises of two parts: 1. Docking or finding favorable

binding positions between the binder and target antigen; 2. Mutate the binder residues that do not qualify as a good binding position with the antigen. Some of the key features of this tool are: It uses known structural features such as hydrogen bonds (H-bonds) to drive favorable binding pose identification and docking; It does not use forcefields during the design process but they are used to evaluate the designs based on certain binding metrics; There is no concept of the binder translation or rotation around the target because it uses pairwise distance matching and alignment between several eligible H-bond forming residues of both the epitope and paratope; This allows structurally favorable poses to form multiple H-bonds in a single step, which accounts for strong binding interactions; In case of steric clashes, the residue finds other poses and a maximum of five structural distinct rotamers (Dunbrack library)[52] are used to generate a design pose. If the binding pose remains of poor quality, then the residue is mutated and there is a priority list of residue mutability, which can be found in the MutDock article[42] in great detail. This entire design simulation steps are repeated for all the selected epitope residues, which were used as inputs, and the final designs are generated. All simulation runs were submitted in the Auburn University Easley Cluster as individual jobs.

3.2.3. CHARMM & Rosetta forcefields energy minimization

Although forcefields were not used for energy calculations during the MutDock design run, they were used to relax the final poses and evaluate them. Two energy minimization forcefields were used to relax the designs i.e., CHARMM[44] fixed backbone forcefield energy minimization followed by Rosetta REF2015[53] all-atom forcefield energy minimization calculations. If energy minimization calculations are not used during a design/prediction process, there will always remain some minor steric clashes, which can be relaxed using these forcefields. Since the primary motive of the study was to design protein complexes with good binding poses, the most important

metric was binding energy (B.E.), which is the difference in free energy between the bound and unbound state of the complex[54]. The other metric scores such as buried surface area (BSA)[55], shape complementarity (Sc)[56], [57], [58] were also calculated using RIA. The structures were sorted according to these scores and the best designs were selected.

3.2.4. Rosetta affinity maturation to improve binding

To further improve the binding affinity of the DARPins, the best MutDock generated and Rosetta minimized design complexes were selected for the Rosetta affinity maturation process. The general fixed backbone design application (fixbb)[59], [60] was used along with a *resfile*. The entire path to the database and the designed complex PDB files were used as input, the packing corresponded to the instructions that were provided in the *resfile*, the number of iterations was kept as 100 structures per minimized design, the latest weights function i.e., ref2015 was used. The *resfile* has specific syntax and conventions that are necessary to be incorporated in the correct way e.g., ALLAA and ALLAwc mean to allow all amino acids including cysteines, ALLAAxc means to allow all amino acids except cysteine, POLAR, and APOLAR mean to allow canonical polar and canonical non-polar amino acids respectively, and several other commands that are necessary to be understood clearly and then used for the purpose. It is the instructions of this file that dictate the interface residue mutations and are therefore responsible for improving the binding affinity of the protein binder. The *resfile* and the remaining commands are all included in Appendix 1. The generated affinity matured files were minimized and RIA was used to calculate the new binding metric scores. These scores were compared to the ones, which were generated using MutDock. At the end of every major step of this project, the best designs were selected and filtered to enter the next phase of the process. However, post-screening there were still plenty of structures, which were further subjected to visual inspection using UCSF Chimera[61].

3.2.5. UCSF Chimera was used to visually inspect and shortlist the designs

UCSF Chimera[61] was used to observe the selected affinity-matured designs and study their characteristic features. They were visually inspected to check whether they have good structural conformation e.g., the complete burial of the complex core, salt bridges, hydrophobic interactions, hydrogen bonds, etc. After this step, the designs were further narrowed down for molecular dynamics simulation runs.

3.2.6. Molecular dynamics simulations to assess structural stability over time

The affinity-matured designs were further filtered and prepared for 5ns MD simulation (MDs) runs. Visual Molecular Dynamics (VMD)[47] was used to prepare the QwikMD files. All non-standard molecules were removed from the PDBs and they were 'cleaned' before creating a qwikmd file for each affinity matured design. Apart from a qwikmd file, a run and setup folder, and an INFOMD file were also made. The INFOMD file had detailed information for running the MDs. A few parameters were adjusted i.e., nsteps was set to 500 in the Annealing.conf file and the dcdfreq, XSTfreq, and restartfreq were all set to 25000 in the MD.conf file. All of these generated files were uploaded to the Auburn Easley Cluster. Since MDs runs have a high duration, entire nodes (48 cores) were occupied for running individual complex simulations using NAMD[48], installed in the Auburn Easley Cluster.

There is a sequence of steps in an MD simulation run: energy minimization to relax and release excess strain from the system; this is followed by heating the system (annealing) to the temperature (300ºC in this case) at which the simulation is to be carried out; this is followed by the equilibration step to equilibrate the system before the simulation and measuring the properties. This is followed by conducting the simulation runs. After the files were generated post MDs, the PSF and DCD files were downloaded to the PC. VMD[47] was used again to load these files and generate the full

500-frame PDB. This PDB had all the individual frames of the affinity matured designed complex i.e., all 500 conformations throughout the 5ns MD were present in this one big PDB file. Each frame from this PDB file was extracted to form 500 PDB files. This entire process was repeated for all the designs selected for MDs runs. The RIA was used to calculate the binding metric scores for each of the 500 frames for all the structures and their data was analyzed.

3.2.7. Finalizing the design rankings for experiments

The main objective of this project was to design DARPins that bind α-Cbtx. For this reason, while evaluating the structures from the beginning, the B.E has been the focal point although other binding metrics were also considered. Later while observing the trend of plots, a certain correlation was arrived at, which led to the merging of two individual metrics to form a more reliable one. This reasoning has been explained in detail in Chapter 2. As it is well established that designs with such metrics have been successfully experimentally tested over the years[57], [62], [63], [64], the final rankings were based on them. However, since there were 500 different frames, which suggests that there were 500 different PDB structures (different conformation) per MDs run, the characteristics had to be considered for all of them. Therefore, the standard deviations (SD) of these metrics were calculated for each of the 73 MDs runs. It's worth mentioning that although the 500 frames B.E/BSA plots were made using smoothened data with a moving average of 5 data points, the SD were made without using any average i.e., using raw data from the MDs. The primary reason to consider this criterion (SD) is that it provides a characteristic measure of the entire trajectory of the designs over time, which cannot be found in the singular forcefield binding metric scores. Finally, designs with the lowest SD values were the top-ranked structures; however, a few mid-ranked designs displayed sub-par structural characteristics and were discarded and, in this way, the best 51 structures were selected from the list of 73.

3.3. Results

3.3.1. Choice of protein scaffold and antigen

Figures 4 and 5 below show the protein binder of choice PDB 6FPA and the target antigen of interest PDB 6ZFM respectively as observed in UCSF Chimera. The DARPin in Figure 4 is orange with its paratope residues highlighted in blue. The α-Cbtx, highlighted in purple (Figure 5) was extracted from PDB 6ZFM, which had multiple chains of the toxin with peptide bound to them. There are two epitopes, I and II. Some of the residues in I (highlighted in green) were obtained from literature surveys[65], [66], which binds to nAChRs and subtypes, and the residues of epitope II (highlighted in blue) were selected by identifying the presence of ample pre-stabilized residues. A pre-stabilized residue has fewer degrees of freedom (DOF) of motion when it is in its unbound state and in its bound state, it further loses those degrees of motion.

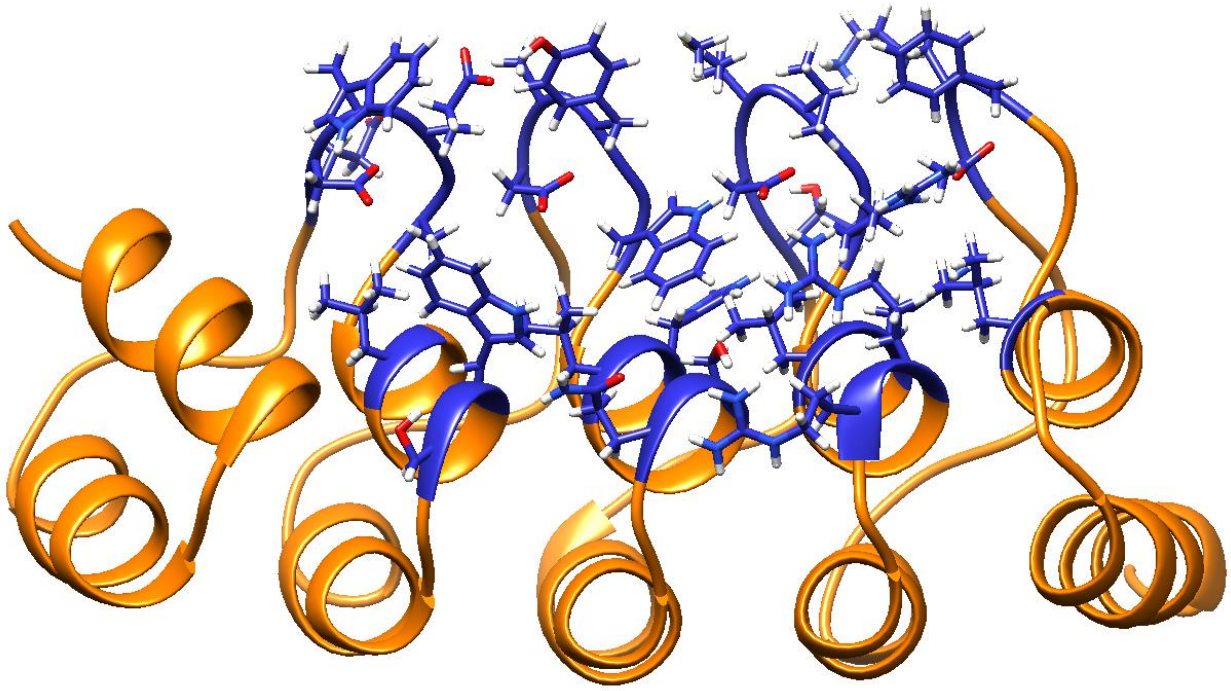

Figure 4. **DARPin paratope residue identification**. PDB 6FPA is an X-ray crystallography 1.58 Å resolution DARPin, expressed in *E. coli* cells, and has no missing residues. The highlighted blue residues represent the identified paratope binding region (PBR), which will be used by MutDock to find favorable H-bond forming region with the targeted epitope residues.

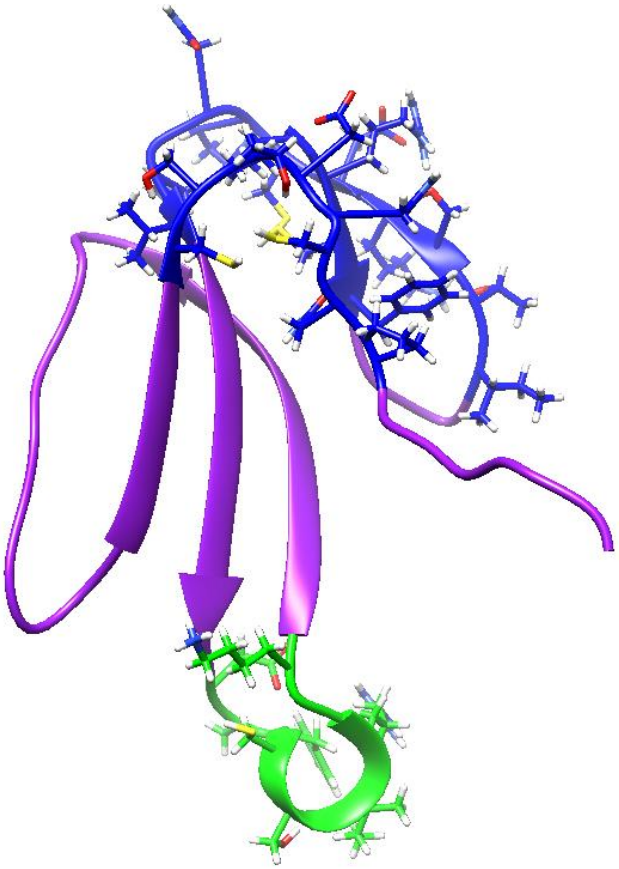

Figure 5. **α-Cobratoxin PDB 6ZFM epitope identification**. It is an X-ray crystallography 1.90 Å cobratoxin obtained from *Naja kaouthia*. The highlighted green residues represent epitope I whereas the blue residues represent the identified epitope II consisting of pre-stabilized residues.

3.3.2. Epitope identification

Figure 6 shows the key residues of α-Cbtx interacting with an AChP (PDB 1YI5). These residues were a portion of epitope I. Other residues were also selected to increase the binding probability by strong bond formations. Some of these residues are W25, D27, F29, R33, K35, R36, and F65. The Trp and Phe residues are responsible for forming the hydrophobic π-π interactions owing to the presence of their big aromatic rings, while interacting with similar receptor residues[67][68]. The charged residues i.e., Arg, and Lys form salt bridges, and cation-π interactions with several neighboring receptor residues as they have long sidechains. The Asp in position 27 is a highly conserved residue, which plays a key role in binding to *Torpedo* nAChR [6][69]. Prior studies were conducted by mutating this residue to check the effectivity in binding,

which showed significant changes in the binding affinity[69], [70]. From Figure 6, F65 may look like it is placed far away from a potential binding site but it establishes interaction with AChBP loop residues[66]. Research has been conducted to establish the intermolecular interactions between these neurotoxins and nicotinic receptor proteins. These interactions have been well tabulated in the literature by Bourne Y, et al.[66]. Hence, after a thorough literature survey, the choice of residues for epitope I were K23, W25, C26, D27, A28, F29, C30, R33, K35, R36, K45, and F65.

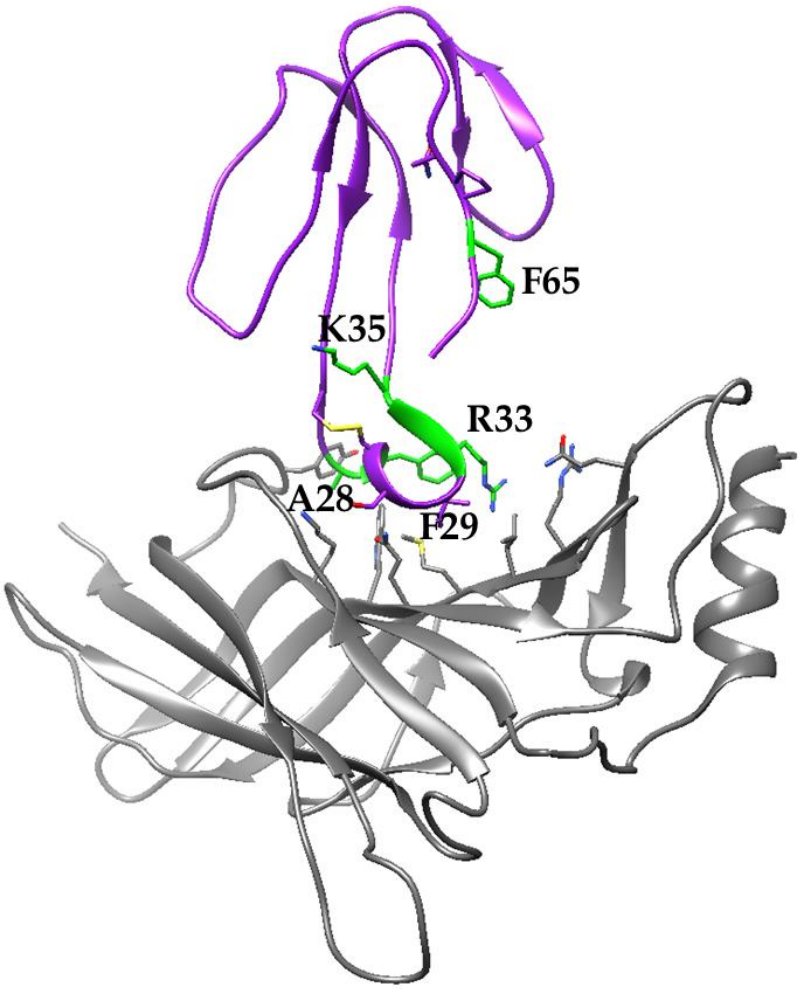

Figure 6. **α-Cobratoxin in complex highlighting its key binding residues**. The figure was extracted from PDB 1YI5 and it was visualized in UCSF Chimera. The molecule in light gray is an acetylcholine binding protein, which is interacting with the purple α-Cbtx. The green highlighted atoms represent some of the key binding residues, which have been established in several studies to interact with these receptor proteins.

From Figure 7, it is observed that there are both polar and non-polar residues, which are significantly close to each other such that they are limited in terms of their movement. For example, F4 has a big aromatic ring, which is stabilized by P64, N61, N63. Similarly, I9 is partially interacting with F65, which is not visible here since it was included in epitope I but that does not deny its presence around these residues and stabilizing them. Near the protruding loop with N16, H18 is being stabilized by P15, I1, C14, A42 and A43. C3 and C14 are forming a disulfide bond, which further stabilizes them. Additionally, R2 is forming a salt bridge with D13 and interacting with polar residues S1, and N61. The third loop near S58 has two more C57 and C62 which helps stabilize that region. Thus, the concoction of numerous stabilized residues and the convex shape of this region led to the identification of epitope II.

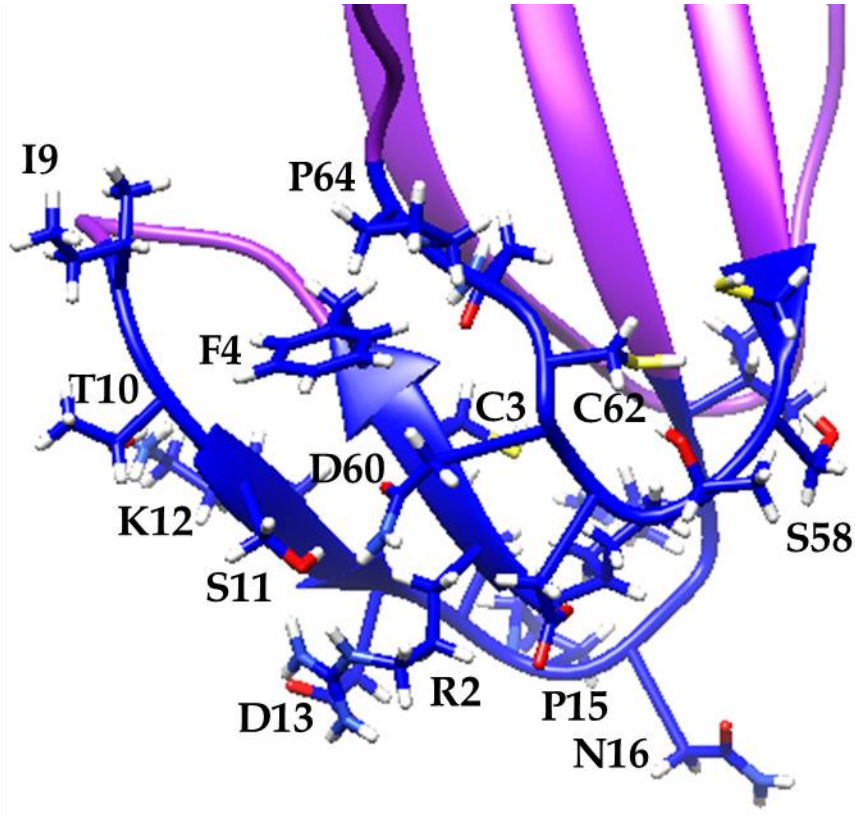

Figure 7. **α-Cobratoxin's key binding residues of alternate epitope**. The blue residues represent epitope II, which was selected on the basis of pre-stabilization. MutDock was developed in a way such that it designs more structures in presence of higher pre-stabilized residues.

3.3.3. MutDock results analysis

After finalizing the epitopes, the paratope residues were selected and there were 2 separate MutDock design simulation runs. A total of 2145 designs were generated out of which 375 were from epitope I and 1770 were from epitope II. All the designs were minimized using CHARMM and Rosetta forcefield energy minimization, and their binding metric scores were calculated using the RIA module as mentioned earlier. While CHARMM fixed backbone energy minimization was consistently able to relax all the generated structures, Rosetta all-atom energy minimizer was used since it is the most commonly used forcefield and conveniently the finalized calculated binding metric scores can be easily compared and understood.

Figure 8 shows the results of MutDock generated structures' characteristic plot of B.E. (y-axis) vs. BSA (x-axis). The curve font colors were retained based on the color of the respective epitopes. Epitope I is green while epitope II is blue. The triangular data points, both green and blue are the ones, were selected based on the binding metric threshold values as shown in Table 1. The trend shows that with an increase in the BSA of the complexes, their B.E. got stronger. A reason for this is that initially in the unbound state, there was a higher value of solvent accessible surface area (SASA) as there were ample interactions with the solvent molecules. As the molecules interacted and eventually formed a complex with a strong hydrophobic core, there was a gradual decrease in SASA, which indicates an increase in BSA. This is only possible in the case of a complex with strong B.E, where there is almost no solvent interaction with the complex core. Instead of evaluating the complexes using B.E. and BSA as separate metrics, the two were combined for more reliability and made 100-fold for ease of calculation. In a previous study conducted in our lab, a database consisting of 231 experimentally determined non-redundant antibody-protein structures had an average B.E/BSA score of -2.8 ± 0.6 kcal/mol. $Å^2$. The actual score is -0.028 ±

0.006 kcal/mol. However, for convenience in calculating and analyzing these figures, the score was multiplied by -100 as shown in Table 1 (4^{th} column, 1^{st} row). The threshold value for selecting the affinity matured structures was within the average database range. As previously mentioned, an Sc value of 0.60 has been consistent in generating good quality designs, the Sc value was kept at 0.60. Since the structure is way smaller than an Ab, the BSA, and B.E./BSA were kept at 600$Å^2$ and 2.50 kcal/mol. $Å^2$ respectively. These complexes are significantly smaller (10% the size) than conventional Ig, so it is expected that they will not have comparable values of BSA. However, since Sc is a mathematical function or a structural geometric fit between the two molecules, they are likely to have standard values as other complexes as well. An Sc value of around 0.60 is considered to be good for such complex formations[40], [57], [58]. Considering these criteria altogether, there were a total of 39 (23 epitope I and 16 epitope II) designs, which were further selected for Rosetta affinity maturation.

Table 1. RIA calculated binding metric thresholds for shortlisting promising MutDock design

Complex	**Sc**	**BSA**	**B.E/BSA (-100x)**
Epitope I	≥0.60	≥600 $Å^2$	>2.50 kcal/mol. $Å^2$
Epitope II	≥0.60	≥600 $Å^2$	>2.50 kcal/mol. $Å^2$

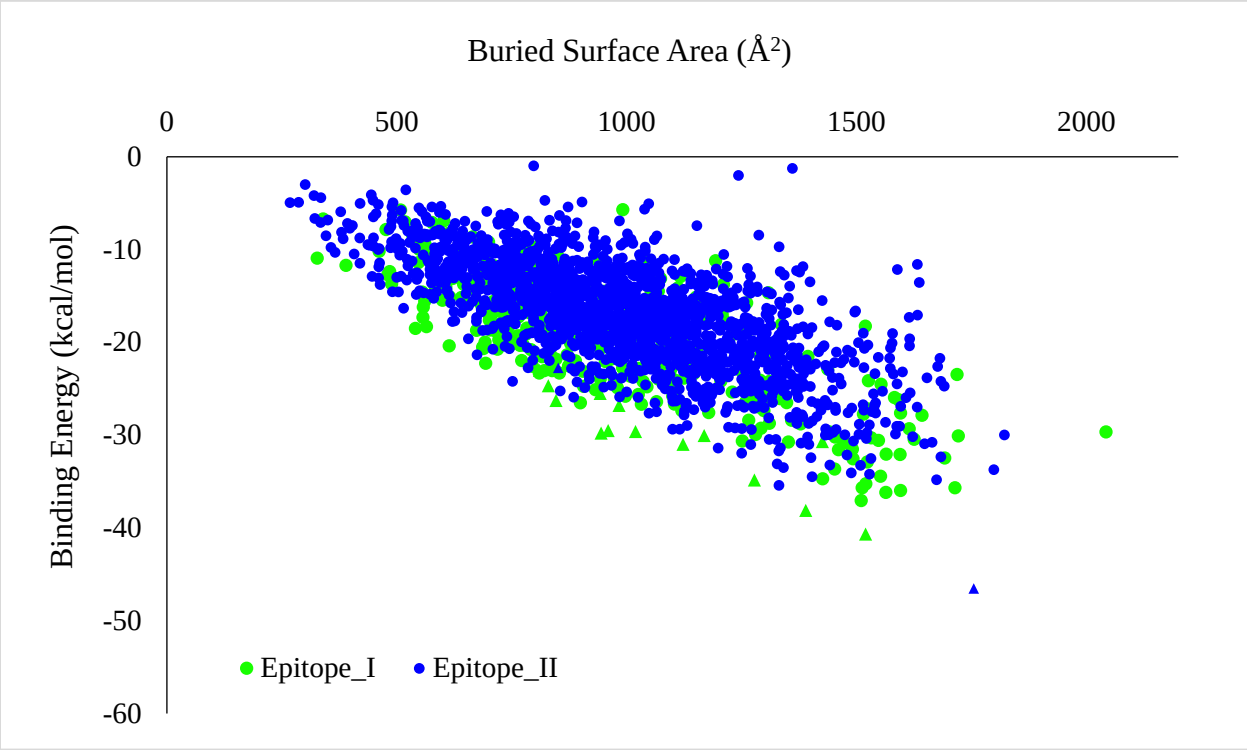

Figure 8. **Comparative analysis of MutDock generated epitopes I & II designs**. Epitope I is highlighted in green while epitope II is highlighted in blue. It is evident from the curves that they have a negative linear slope, which suggests that with increase in the BSA, the B.E got stronger. The data points marked as triangles in both the font colors are the ones, which were selected for Rosetta affinity maturation to further improve binding affinity. One thing to consider in this plot is that since there are more epitope II designs compared to epitope I, they are overshadowing the epitope I data points by a great extent. Here, the epitope II marker size has been reduced to allow more visibility for the epitope I points. Hence, the variability in marker size.

3.3.4. Rosetta affinity maturation results analysis

3900 structures were obtained from the affinity maturation designs i.e., 2300 and 1600 for epitopes I and II respectively. Figure 9 displays the RIA calculated B.E. vs. BSA curve for the affinity matured designs. Observably on the same scale, the affinity matured curves have a higher magnitude of negative slope compared to the one in Figure 8. Additionally, from the plot characteristics, it can be inferred that along with an increase in BSA, the B.E became significantly stronger than the MutDock plot. Since, the BSA gradually increases, the only way the curves can slide down along the negative y-axis is if there is an increment in the negative magnitude of the

numerator, which is possible only if the B.E gets stronger. This suggests that the affinity-matured complexes had better binding characteristics than the ones obtained directly from MutDock.

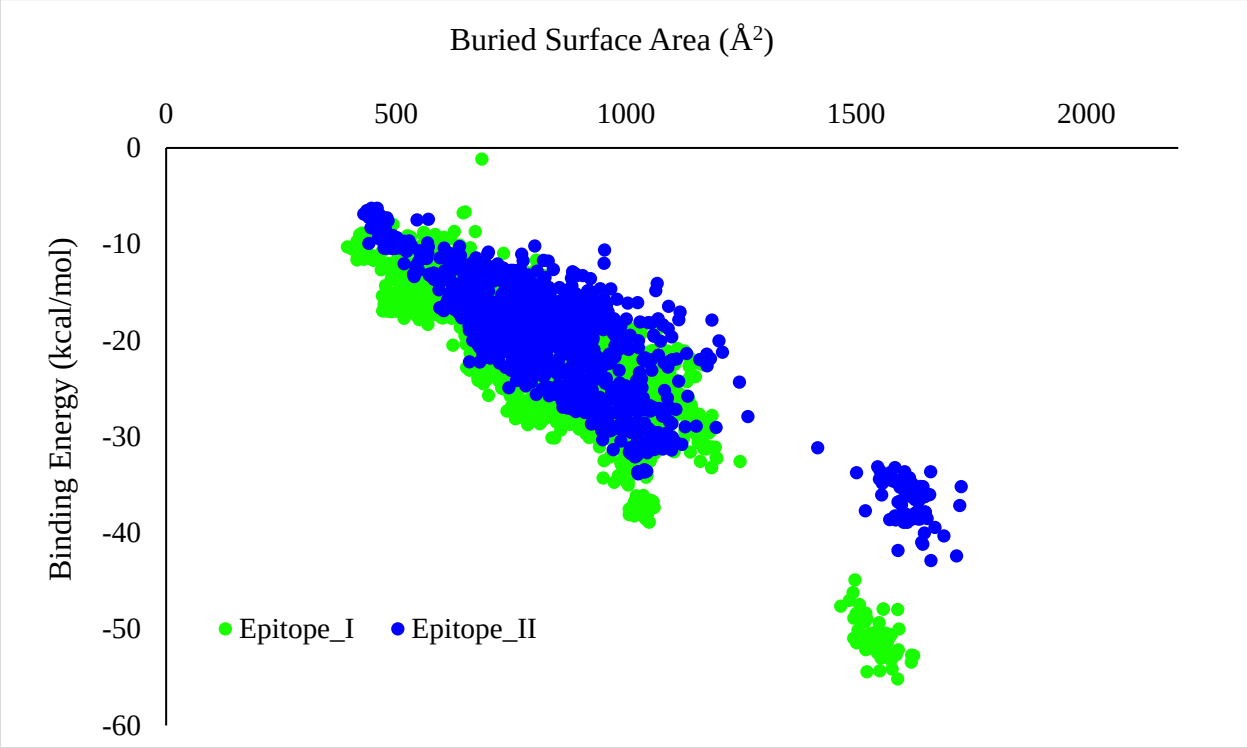

Figure 9. **Comparative analysis of epitopes I & II Rosetta affinity maturation results**. Both the plots have the same scale because of easier comparison. The plot shows an increase in the negative slope, suggesting that there was an increment in the negative magnitude of B.E. Post affinity maturation, with increase in BSA, the B.E. got even stronger.

3.3.5. Visual inspection using UCSF Chimera

The molecular mechanics (MM) forcefields are not the absolute benchmarks for assessing protein structures. Often, they cannot capture the characteristics that molecular dynamics (MD) simulations can. For assessing the structural stability of the designs over time, MDs was used. As earlier stated, standard thresholds were considered to filter the designs for affinity maturation and on a similar basis, Table 2 shows the thresholds for further shortlisting the best designs. Since the binding feature of the designs improved with affinity maturation, close to 200 structures. The structures were visually inspected using UCSF Chimera qualitatively, to check for promising binding features. Some of the characteristics considered while visually inspecting these designs

are salt bridge formation, since this is the strongest kind of electrostatic interaction; complete burial of the core residues, and the absence of voids/gaps around it. The presence of voids around the peripheral area of the core packing will lead to the entry of solvent molecules. This is detrimental to the complex stability. Generally, CPE has been more efficient in designing complexes with strong hydrophobic core interactions rather than polar ones. Post visualization in UCSF Chimera, 73 complexes were selected matching all the qualitative and quantitative thresholds as shown in Table 2. The reason for choosing these values is that there was more affinity matured complexes with better binding metrics than the Mutdock generated designs, which it is supposed to do. Therefore, a stricter threshold value was considered for this affinity matured designs and even then, they had to be narrowed down to the top 73 for MDs.

Table 2. Binding metric thresholds for shortlisting promising Rosetta affinity matured designs

Complex	**Sc**	**BSA**	**BE/BSA (-100x)**
Epitope I	≥0.65	≥700 $Å^2$	>3.00 kcal/mol. $Å^2$
Epitope II	≥0.65	≥700 $Å^2$	>3.00 kcal/mol.$Å^2$

3.3.6. Molecular dynamics simulation results analysis

The binding metrics of the individual 500 frames were calculated using RIA and plotted against time (5ns). The results have been shown in 12 different plots, each having the trajectories of 6 distinct designed complexes. To plot smoother curves, a moving average of 5 data points was considered and applied to all the 500 data points for each designed complex. This indicates that the curves will be aesthetically appealing compared to their actual value plots. The plots were B.E/BSA vs. time (5ns) and their corresponding BSA vs. time (5ns). They were qualitatively and quantitatively analyzed and then further cross-verified to match the results.

The qualitative aspect was that the curves should have a near zero slope. To elaborate, if the 1st frame of a design started at a point (x, y) and went on till frame 499th following a certain trajectory. Then mathematically, the 500th frame should be at a point (x`, y`) where (x`, y`) tends to (x, y) such that $\Delta y/\Delta x$ is as minuscule as possible. This is to prove that the designed complex ends at that exact conformation or near to it from where it started i.e., there should not be any conformational changes at the end. If the trajectory ends in an upward trend, it represents dissociation since the binding energy weakens with time. If the trajectory ends in a downward trend, it exhibits a better conformation than the one where it started or simply the current designed complex is not good enough. As described above, a curve can have one of the three outcomes.

Another method of qualitative analysis is to seek designs with a lower range of frequency i.e., a relatively tighter or less fluctuating trajectory. This suggests that the intermolecular bonds are strong such that they help the complex remain in their designed conformation with minimal deviations throughout the period. On the other hand, highly, fluctuating trajectories would indicate that the ligand is detaching from the protein to an extent at some point, the frames of which will possess the weakest B.E and low B.E./BSA values as well. It is therefore highly desirable for the designs to have the least fluctuating trajectory over time. The range of fluctuations can be quantified by calculating the standard deviation (SD) of each data point and then ranking them with the least to most variability. The ones with high SDs are ranked last compared to the ones with low SDs, which suggests the relatively tight nature of the complex trajectories of the latter. Once the designed complexes have been ranked using the nature of the plots (qualitatively) and the SD (quantitatively), the two distinct ranking lists can be matched to conclude if the top and the bottom rankings are similar. They might not be the same but there should not be a case where a qualitatively top-ranked structure is found at the bottom in terms of its quantitative ranking.

After this step, the other score rankings can similarly be matched to check if the top and bottom-ranked structures are close to each other. Mathematically, the characteristics of a curve observed in the plots should represent the quality of its SD as well. These criteria helped evaluate the structures and rank them from top to bottom with the currently available computational capacity.

All the plots have an equal range of axes values to ensure fair comparison e.g., all values start from 0 for both the axes. For the plots on the left-hand side i.e., B.E/BSA vs. time, the range is 0.0 to -4.0 Kcal/mol. $Å^2$. Similarly, for the ones on the right, the range is 0 to 1800 $Å^2$. This is to maintain consistency throughout the plots for better comparison. Since the basis for evaluating these structures has been previously explained in detail, they will not be discussed in greater detail. Instead, a few unique examples will be used to understand the key takeaway from these plots. In Figure 10, the complexes 213_81 and 245_45 are considered in the first plot. Both of them have similar trends i.e., they start with relatively tightness and fewer fluctuations, but they keep increasing, to a point where they cross the 0.0^{th} mark for the B.E/BSA. From a thermodynamic perspective, a positive free energy represents non-spontaneity, and that represents instability. In the MD animation, the complex starts dissociating slightly after the halfway time, which corresponds to the characteristics of these two curves. From a quantitative perspective, 213_81 ranks 73^{rd} and 67^{th} out of 73 complexes, whereas 245_45 ranks 72^{nd} and 36^{th} in the SD calculations for B.E/BSA and BSA respectively. For 245_45, the B.E had a stronger effect on the dissociation tendency compared to BSA. These findings summarize that such structures, which cross the 0.0^{th} range can be marked as bad designs. Moreover, a weak complex can also be identified qualitatively from the plots e.g., 1025_36 has a wide range of fluctuation throughout its trajectory on the B.E/BSA plot, which signifies that the complex is not attached strongly. To match it up, we can

observe that its BSA has significant variations on the right-side plot corresponding to its left. The BSA values nearly halved almost halfway into the MDs. To further justify, complex 1025_36 ranks 46th and 66th out of 73 complexes on the SD calculations for B.E/BSA and BSA respectively. It has a more or less good rank than the previous complexes. Hence, it was considered to be weak rather than being treated as a bad design.

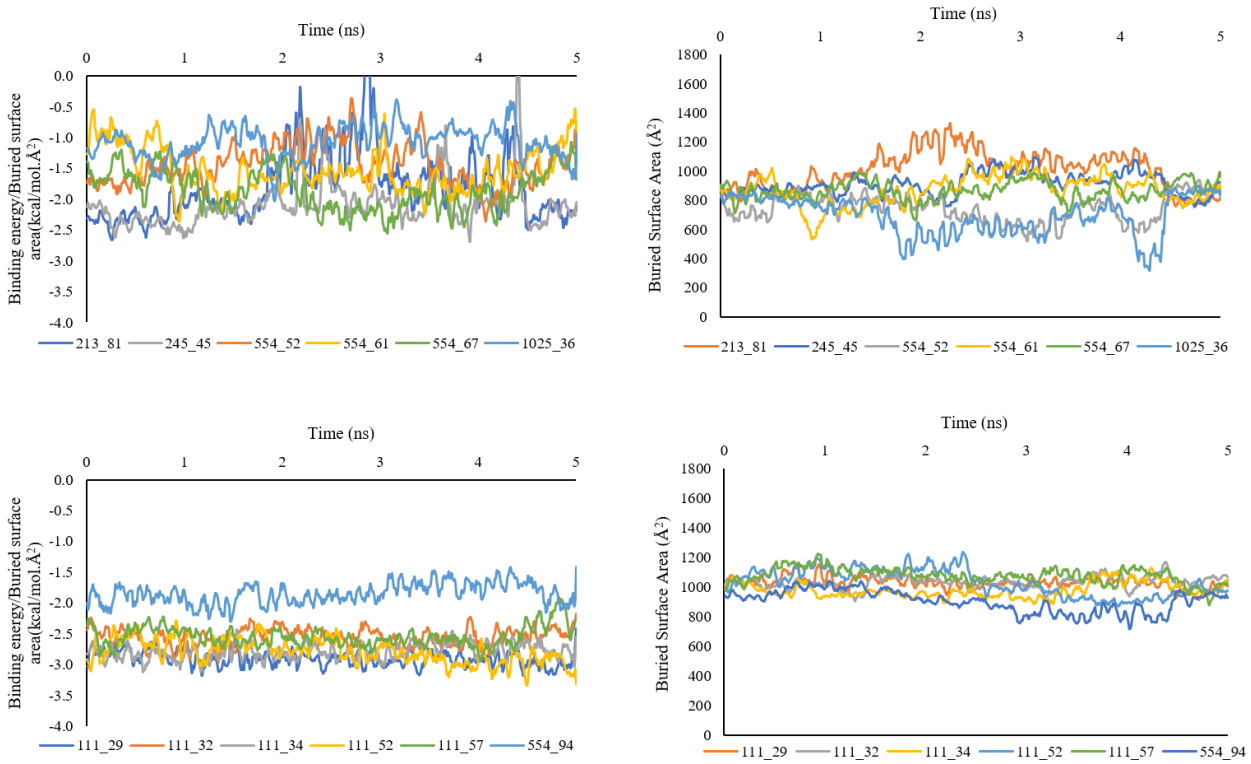

Figure 10. **Molecular dynamics simulation trajectory of affinity matured designs**. It shows the characteristics of 12 designs. There are 4 distinct plots, 2 with B.E/BSA vs. time on the left and 2 with BSA vs. time on the right. Both the left and right plots belong to the same designs. A detailed explanation has been provided in the previous sections to understand these plots.

We have already observed and identified the characteristic plots of bottom-ranked designs, so our focus should shift to the good ones. In Figure 11, all the complexes except 111_58, 111_59, and 111_62 have relatively stable trajectories. They are characterized by tighter or less fluctuating trajectories, they exhibit a relatively straighter curve with a near zero slope, which suggests that throughout the simulation, they held onto their structural conformation and did not exhibit a

tendency to dissociate or move much due to the absence of strong bonds. These are all the characteristics, which should be present in a design to render it stable. Additionally, the ranks of the complexes 111_63 and 111_78 are 13th and 10th in the SD calculations of B.E/BSA respectively, and 4th and 31st for BSA SD calculations respectively. The same categorical ranks for the B.E/BSA and BSA SD calculations of complexes: 111_85 – 9th and 19th; 111_89 – 17th and 7th; 111_91 – 20th and 60th; 111_93 – 15th and 22nd; 111_94 – 7th and 12th; 111_97 – 8th and 11th. Therefore, judging by the quantitative results and rankings of these complexes, it can be claimed that with the appropriate observation and data analysis a complex can be identified to be good or bad based on these metrics and the quality of the plot as well. The three curves, which were excluded were not 'bad', but they did not lie in the top tier rankings, which was quite evident from their nature itself.

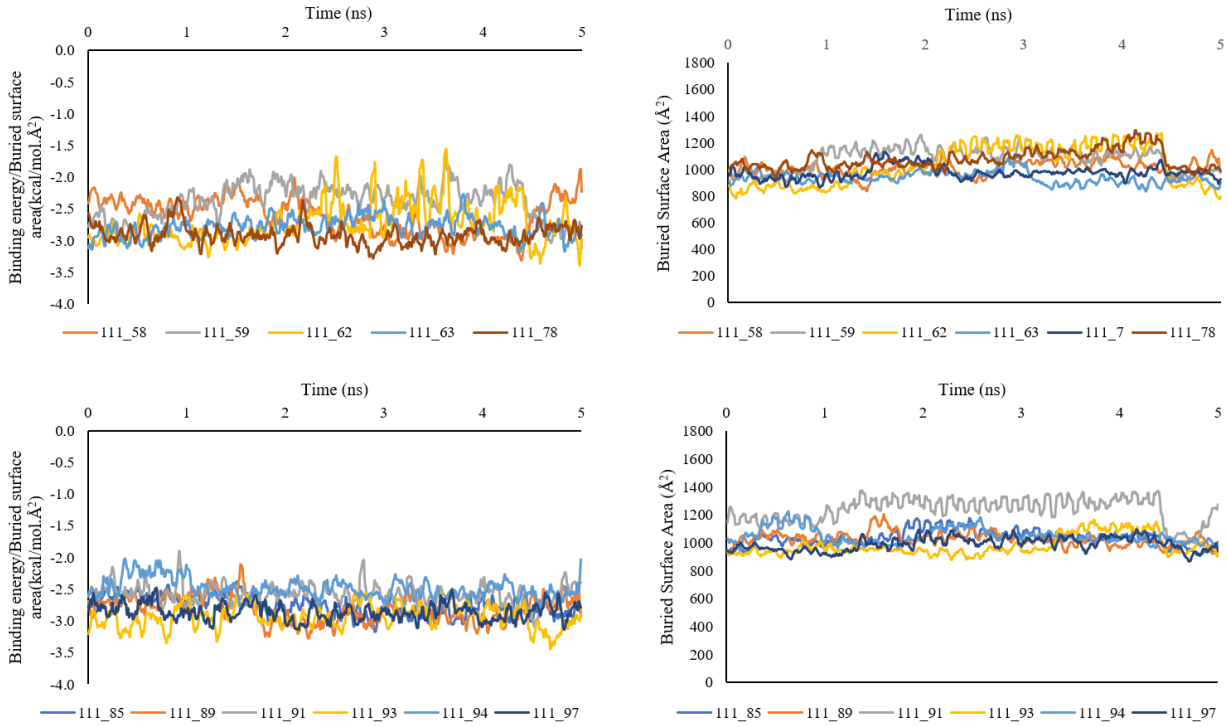

Figure 11. **Molecular dynamics simulation trajectory of affinity matured designs**. This figure shows the MDs characteristics of the designed complexes. This plot was used to identify the behavior of 'good' designs over time.

Although it was not previously mentioned, the second set of plots presented in Figure 10 (B.E/BSA and BSA) only had good structures. All their trajectorial characteristics matched the descriptions of 'good' designs as previously explained. In terms of SD calculation rankings for B.E/BSA and BSA respectively they were: 111_29 – 2nd and 2nd; 111_32 – 3rd and 1st; 111_34 – 4th and 9th; 111_52 - 24th and 39th; 111_57 – 5th and 18th; 554_94 – 23rd and 26th. Hence, it is quite clear that they occupy the top-ranked positions qualitatively and quantitatively.

The full set of rankings based on the individual SD can be found in the Appendix. The overall ranking of the complexes can be found in Table 5. The designs in this project were consistently selected based on several filters e.g., binding metrics, structural visualization, MDs trajectories, MDs animation visualization, etc. Similarly, the final rankings were obtained based on further observing the MDs' animations and viewing at their interfacial stability, plot characteristics, and SD rankings. On this basis, only the top and middle rankings will be experimentally tested in the laboratory. The bottom rankings will only be tested if the selected designs fail.

From the plot characteristics, it can be observed that MM forcefields are unable to capture the behavior of designed complexes over time, which can change to a great extent as it was shown previously in certain cases. This is verifiable by using MDs for such designed complexes. Therefore, it is necessary to conduct MDs to check the behavior of such complexes over time because MDs can capture what the MM forcefields fail e.g., binding complexes can have excellent metrics but they can prove to be either dissociating or have weak bond strength or display unstable nature of complex formation over time. The following 6 complexes in Table 3 and Table 4 have good or top ranked binding metrics, which were shortlisted for MDs runs:

Table 3. Complexes with top tier binding metrics but with bottom order SD rankings

Complex	**Sc**	**BSA ($Å^2$)**	**B.E/BSA (Kcal/mol)**
32_37	0.773	786.268	-3.657

111_62	0.719	1017.16	-3.625
32_42	0.756	783.44	-3.614
111_42	0.723	1019.202	-3.599
32_28	0.751	844.924	-3.568
213_11	0.705	974.102	-3.566

These structures have high binding metric values but the following are their B.E/BSA and BSA SD calculation rankings from the MDs plots:

Table 4. Complexes with binding metrics in Table 3 with their respective SD rankings

Complex	**Sc (Rank)**	**BSA (Rank)**	**B.E/BSA (Rank)**
32_37	54th	72nd	68th
111_62	37th	69th	67th
32_42	60th	48th	64th
111_42	3rd	33th	51st
32_28	68th	73rd	71st
213_11	62nd	71st	69th

Except for one rank in Sc of a single complex, none of the other complexes showed any ranks even within the top 25, whereas the MM forcefield analyzed metric values were all in the top 25 positions. This suggests the importance of MDs in such studies. Based on the nature of the plots and their quantification such fallacies can be identified and further examined to select the best designs and test them experimentally in the laboratory.

Table 5. Table showing the final design rankings and the ones in red are the discarded designs.

Rank	*PDB ID*	*BE/BSA (kcal/mol.Å^2)*	*SD (BE/BSA)*	*Rank*	*PDB ID*	*BE/BSA (kcal/mol.Å^2)*	*SD (BE/BSA)*
1	364_79	-3.246	0.190	37	554_36	-3.105	0.341
2	111_29	-3.532	0.228	38	554_100	-3.104	0.346
3	111_32	-3.021	0.233	39	554_83	-3.103	0.347
4	111_34	-3.463	0.239	40	32_83	-3.584	0.347
5	111_57	-3.278	0.242	41	32_45	-3.302	0.347
6	364_84	-3.213	0.243	42	111_50	-3.522	0.348
7	111_94	-3.257	0.246	43	554_48	-3.053	0.352

8	111_97	-3.452	0.247	44	554_57	-3.059	0.353
9	111_85	-3.020	0.252	45	554_35	-3.275	0.354
10	111_78	-3.785	0.259	46	1025_36	-3.226	0.357
11	111_1	-3.699	0.269	47	111_59	-3.482	0.364
12	364_5	-3.296	0.271	48	554_44	-3.013	0.368
13	111_63	-3.703	0.274	49	32_66	-3.486	0.368
14	111_41	-3.692	0.274	50	111_2	-3.657	0.377
15	111_93	-3.703	0.278	51	111_42	-3.599	0.378
16	111_46	-3.149	0.279	52	390_73	-3.371	0.380
17	111_89	-3.758	0.282	53	554_90	-3.057	0.384
18	111_49	-3.700	0.288	54	554_27	-3.014	0.385
19	111_28	-3.577	0.288	55	111_5	-3.466	0.414
20	111_91	-3.727	0.289	56	1025_62	-3.081	0.426
21	111_44	-3.692	0.291	57	554_52	-3.215	0.430
22	554_49	-3.032	0.291	58	554_67	-3.215	0.430
23	554_94	-3.264	0.299	59	111_67	-3.534	0.435
24	111_52	-3.537	0.299	60	111_96	-3.478	0.441
25	32_87	-3.577	0.300	61	554_32	-3.052	0.441
26	554_84	-3.059	0.306	62	554_21	-3.041	0.452
27	554_69	-3.256	0.311	63	554_61	-3.080	0.461
28	32_50	-3.707	0.313	64	32_42	-3.614	0.498
29	554_42	-3.059	0.315	65	554_68	-3.067	0.503
30	111_7	-3.319	0.318	66	554_38	-3.029	0.518
31	111_17	-3.645	0.318	67	111_62	-3.625	0.519
32	111_26	-3.487	0.321	68	32_37	-3.657	0.520
33	554_37	-3.264	0.326	69	213_11	-3.566	0.522
34	554_65	-3.048	0.327	70	281_57	-3.029	0.527
35	554_18	-3.237	0.339	71	32_28	-3.568	0.539
36	111_58	-3.723	0.340	72	245_45	-3.149	0.545
				73	213_81	-3.016	0.768

3.4. Discussion

This chapter discussed using computational tools to design and analyze the DARPins - α-Cbtx complex. The protein scaffold and antigen were selected from the RCSB PDB after careful inspection of several criteria. The two were the inputs in the MutDock instruction file to run a design simulation. MutDock, as mentioned before is a computational docking algorithm that

simultaneously finds good binding poses and mutates residues. Two different epitopes were selected I and II based on their interaction with nicotinic binding proteins and subtypes, and the presence of pre-stabilized residues. MutDock generated 2145 designs, which were minimized using CHARMM36 and Rosetta energy minimization forcefields, and their binding scores were calculated using RIA for design evaluation. The top 39 designs were shortlisted for Rosetta affinity maturation. 3900 designs were generated as affinity matured complexes and evaluated using the forcefields. 200 affinity matured designs were shortlisted. These 200 designs were further subjected to UCSF Chimera visual inspection and the best 73 designs were used in NAMD[48] MDs runs. Each frame was evaluated using RIA and analyzed qualitatively and quantitatively. The designs were then ranked from the best to worst and the top 51 were selected for experiments.

The final results were evaluated without any identifiable bias i.e., the plots were observed and their characteristics were studied. Once they were analyzed and ranked, their SD were calculated. The complex with the least SD values represented less fluctuation, stronger binding affinity, and more structural stability. Thus, they were ranked the best and the ones with the highest SD scores were ranked at the bottom. Once, the structures were quantified, the structures were matched based on the qualitative and quantitative rankings. Although the ranks were not the same, they matched being on the top/mid or bottom tier of the list. This suggests that the structures were well identified and correctly analyzed for further experimental testing in the laboratory.

this

This project used computational tools to generate designs, analyze them, and predict the ones that will potentially work in wet laboratory experiments. For the next chapter, methods were developed to design '*ideal*' binding proteins from scratch. These proteins are highly soluble, thermostable, humanized binders that can bind antigens and interact with immune receptors.

References

[1] B. Y. C. Willyard, “A Fanged Crisis,” vol. 621, no. September, pp. 40–47, 2023.

[2] J. P. Chippaux, *Incidence and mortality due to snakebite in the Americas*, vol. 11, no. 6. 2017. doi: 10.1371/journal.pntd.0005662.

[3] I. Bolon *et al.*, “Snakebite in domestic animals: First global scoping review,” *Prev Vet Med*, vol. 170, no. June, p. 104729, 2019, doi: 10.1016/j.prevetmed.2019.104729.

[4] W. Suraweera *et al.*, “Trends in snakebite deaths in India from 2000 to 2019 in a nationally representative mortality study,” *Elife*, vol. 9, pp. 1–37, 2020, doi: 10.7554/eLife.54076.

[5] N. L. S. Roberts *et al.*, “Global mortality of snakebite envenoming between 1990 and 2019,” *Nat Commun*, vol. 13, no. 1, pp. 1–11, 2022, doi: 10.1038/s41467-022-33627-9.

[6] C. M. Barber, G. K. Isbister, and W. C. Hodgson, “Alpha neurotoxins,” *Toxicon*, vol. 66, pp. 47–58, 2013, doi: 10.1016/j.toxicon.2013.01.019.

[7] D. C. I. Koh, A. Armugam, and K. Jeyaseelan, “Snake venom components and their applications in biomedicine,” *Cellular and Molecular Life Sciences*, vol. 63, no. 24, pp. 3030–3041, 2006, doi: 10.1007/s00018-006-6315-0.

[8] C. M. Modahl, A. K. Mukherjee, and S. P. Mackessy, “An analysis of venom ontogeny and prey-specific toxicity in the Monocled Cobra (Naja kaouthia),” *Toxicon*, vol. 119, pp. 8–20, 2016, doi: 10.1016/j.toxicon.2016.04.049.

[9] K. Suryamohan *et al.*, “The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins,” *Nat Genet*, vol. 52, no. 1, pp. 106–117, 2020, doi: 10.1038/s41588-019-0559-8.

[10] O. H. Del Brutto and V. J. Del Brutto, “Neurological complications of venomous snake bites: A review,” *Acta Neurol Scand*, vol. 125, no. 6, pp. 363–372, 2012, doi: 10.1111/j.1600-0404.2011.01593.x.

[11] J. M. Gutiérrez, J. J. Calvete, A. G. Habib, R. A. Harrison, D. J. Williams, and D. A. Warrell, “Snakebite envenoming,” *Nat Rev Dis Primers*, vol. 3, no. 1, 2017, doi: 10.1038/NRDP.2017.63.

[12] S. Nirthanan and M. C. E. Gwee, “Three-Finger α-Neurotoxins and the Nicotinic Acetylcholine Receptor, Forty Years On,” *J Pharmacol Sci*, vol. 94, no. 1, pp. 1–17, 2004, doi: 10.1254/jphs.94.1.

[13] X. Yu, Y.-P. Yang, E. Dikici, S. K. Deo, and S. Daunert, “Beyond Antibodies as Binding Partners: The Role of Antibody Mimetics in Bioanalysis”, doi: 10.1146/annurev.

[14] S. Goswami, W. Wang, T. Arakawa, and S. Ohtake, “Developments and Challenges for mAb-Based Therapeutics,” *Antibodies*, vol. 2, no. 4, pp. 452–500, 2013, doi: 10.3390/antib2030452.

[15] C. F. Komives *et al.*, “Opossum peptide that can neutralize rattlesnake venom is expressed in Escherichia coli,” *Biotechnol Prog*, vol. 33, no. 1, pp. 81–86, Jan. 2017, doi: https://doi.org/10.1002/btpr.2386.

[16] L. O. Albulescu *et al.*, “A therapeutic combination of two small molecule toxin inhibitors provides broad preclinical efficacy against viper snakebite,” *Nat Commun*, vol. 11, no. 1, pp. 1–14, 2020, doi: 10.1038/s41467-020-19981-6.

[17] K. T. Yee, M. Pitts, P. Tongyoo, P. Rojnuckarin, and M. C. Wilkinson, “Snake venom metalloproteinases and their peptide inhibitors from Myanmar Russell’s viper venom,” *Toxins (Basel)*, vol. 9, no. 1, 2017, doi: 10.3390/toxins9010015.

[18] Z. A. Ahmad, S. K. Yeap, A. M. Ali, W. Y. Ho, N. B. M. Alitheen, and M. Hamid, “ScFv antibody: Principles and clinical application,” *Clin Dev Immunol*, vol. 2012, 2012, doi: 10.1155/2012/980250.

[19] P. Bannas, J. Hambach, and F. Koch-Nolte, “Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics,” *Front Immunol*, vol. 8, no. NOV, pp. 1–13, 2017, doi: 10.3389/fimmu.2017.01603.

[20] M. T. Stumpp, H. K. Binz, and P. Amstutz, “DARPins: A new generation of protein therapeutics,” *Drug Discov Today*, vol. 13, no. 15–16, pp. 695–701, 2008, doi: 10.1016/j.drudis.2008.04.013.

[21] R. Tamaskovic, M. Simon, N. Stefan, M. Schwill, and A. Plückthun, *Designed ankyrin repeat proteins (DARPins): From research to therapy*, vol. 503. 2012. doi: 10.1016/B978-0-12-396962-0.00005-7.

[22] C. Zahnd *et al.*, “Efficient tumor targeting with high-affinity designed ankyrin repeat proteins: Effects of affinity and molecular size,” *Cancer Res*, vol. 70, no. 4, pp. 1595–1605, 2010, doi: 10.1158/0008-5472.CAN-09-2724.

[23] H. K. Binz *et al.*, “High-affinity binders selected from designed ankyrin repeat protein libraries,” *Nat Biotechnol*, vol. 22, no. 5, pp. 575–582, 2004, doi: 10.1038/nbt962.

[24] M. Gebauer and A. Skerra, “Engineered protein scaffolds as next-generation antibody therapeutics,” *Curr Opin Chem Biol*, vol. 13, no. 3, pp. 245–255, 2009, doi: 10.1016/j.cbpa.2009.04.627.

[25] 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.

[26] A. Bates and C. A. Power, “David vs. Goliath: The Structure, Function, and Clinical Prospects of Antibody Fragments,” *Antibodies*, vol. 8, no. 2, p. 28, 2019, doi: 10.3390/antib8020028.

[27] X. Liu *et al.*, “Computational design of an epitope-specific Keap1 binding antibody using hotspot residues grafting and CDR loop swapping,” *Sci Rep*, vol. 7, no. January, pp. 1–11, 2017, doi: 10.1038/srep41306.

[28] H. W. Schroeder and L. Cavacini, “Structure and function of immunoglobulins,” *Journal of Allergy and Clinical Immunology*, vol. 125, no. 2 SUPPL. 2, pp. S41--S52, Feb. 2010, doi: 10.1016/j.jaci.2009.09.046.

[29] F. Gaudreault *et al.*, “Exploring rigid-backbone protein docking in biologics discovery: a test using the DARPin scaffold,” *Front Mol Biosci*, vol. 10, no. August, pp. 1–13, 2023, doi: 10.3389/fmolb.2023.1253689.

[30] J. Schilling *et al.*, “Thermostable designed ankyrin repeat proteins (DARPins) as building blocks for innovative drugs,” *Journal of Biological Chemistry*, vol. 298, no. 1, p. 101403, 2022, doi: https://doi.org/10.1016/j.jbc.2021.101403.

[31] A. Plückthun, “Designed ankyrin repeat proteins (DARPins): Binding proteins for research, diagnostics, and therapy,” *Annu Rev Pharmacol Toxicol*, vol. 55, pp. 489–511, 2015, doi: 10.1146/annurev-pharmtox-010611-134654.

[32] H. K. Binz, M. T. Stumpp, P. Forrer, P. Amstutz, and A. Plückthun, “Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins,” *J Mol Biol*, vol. 332, no. 2, pp. 489–503, 2003, doi: https://doi.org/10.1016/S0022-2836(03)00896-9.

[33] D. Shinmi *et al.*, “One-Step Conjugation Method for Site-Specific Antibody-Drug Conjugates through Reactive Cysteine-Engineered Antibodies,” *Bioconjug Chem*, vol. 27, no. 5, pp. 1324–1331, May 2016, doi: 10.1021/acs.bioconjchem.6b00133.

[34] F. M. Veronese and A. Mero, “The Impact of PEGylation on Biological Therapies,” *BioDrugs*, vol. 22, no. 5, pp. 315–329, 2008, doi: 10.2165/00063030-200822050-00004.

[35] M. Linder, “Computational enzyme design: Advances, hurdles and possible ways forward,” *Comput Struct Biotechnol J*, vol. 2, no. 3, p. e201209009, 2012, doi: 10.5936/csbj.201209009.

[36] K. Gulati and K. M. Poluri, “An Overview of Computational and Experimental Methods for Designing Novel Proteins,” *Recent Pat Biotechnol*, vol. 10, no. 3, pp. 235–263, 2016, doi: 10.2174/1872208310666161013152249.

[37] S. Fischman and Y. Ofran, “Computational design of antibodies,” *Curr Opin Struct Biol*, vol. 51, pp. 156–162, 2018, doi: 10.1016/j.sbi.2018.04.007.

[38] D. Baran *et al.*, “Principles for computational design of binding antibodies,” *Proc Natl Acad Sci U S A*, vol. 114, no. 41, pp. 10900–10905, 2017, doi: 10.1073/pnas.1707171114.

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

[40] C. E. Tinberg *et al.*, “Computational design of ligand-binding proteins with high affinity and selectivity,” *Nature*, vol. 501, no. 7466, pp. 212–216, 2013, doi: 10.1038/nature12443.

[41] P. S. Huang, S. E. Boyken, and D. Baker, “The coming of age of de novo protein design,” *Nature*, vol. 537, no. 7620, pp. 320–327, 2016, doi: 10.1038/nature19946.

[42] V. M. Chauhan and R. J. Pantazes, “MutDock: A computational docking approach for fixed-backbone protein scaffold design,” *Front Mol Biosci*, vol. 9, no. August, pp. 1–15, 2022, doi: 10.3389/fmolb.2022.933400.

[43] A. Manuscript, “CHARMM.pdf,” vol. 31, no. 4, pp. 671–690, 2011, doi: 10.1002/jcc.21367.CHARMM.

[44] J. Huang *et al.*, “CHARMM36m: an improved force field for folded and intrinsically disordered proteins.,” *Nat Methods*, vol. 14, no. 1, pp. 71–73, Jan. 2017, doi: 10.1038/nmeth.4067.

[45] R. F. Alford *et al.*, “HHS Public Access,” vol. 13, no. 6, pp. 3031–3048, 2018, doi: 10.1021/acs.jctc.7b00125.The.

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

[47] W. Humphrey, A. Dalke, and K. Schulten, “VMD: visual molecular dynamics.,” *J Mol Graph*, vol. 14, no. 1, pp. 27-28,33-38, Feb. 1996, doi: 10.1016/0263-7855(96)00018-5.

[48] J. C. Phillips *et al.*, “Scalable molecular dynamics with NAMD,” *J Comput Chem*, vol. 26, no. 16, pp. 1781–1802, Dec. 2005, doi: 10.1002/jcc.20289.

[49] E. F. Pettersen *et al.*, “UCSF Chimera—A visualization system for exploratory research and analysis,” *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: https://doi.org/10.1002/jcc.20084.

[50] S. K. Burley *et al.*, “RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D464--D474, 2019, doi: 10.1093/nar/gky1004.

[51] M. Alessandra Vigano *et al.*, “Correction: DARPins recognizing mTFP1 as novel reagents for in vitro and in vivo protein manipulations (Biology Open (2018) 11 (bio036749) DOI: 10.1242/bio.036749),” *Biol Open*, vol. 7, no. 12, 2018, doi: 10.1242/bio.040832.

[52] F. E. Cohen, “Bayesian statistical analysis of protein side-chain rotamer preferences,” pp. 1661–1681, 1997.

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

[54] H. Hata, D. P. Tran, M. M. Sobeh, A. Kitao, and M. M. Sobeh, “Binding free energy of protein/ligand complexes calculated using dissociation Parallel Cascade Selection Molecular Dynamics and Markov state model,” *Biophys Physicobiol*, vol. 18, pp. 305–316, 2021, doi: 10.2142/biophysico.bppb-v18.037.

[55] D. Chakravarty, M. Guharoy, C. H. Robert, P. Chakrabarti, and J. Janin, “Reassessing buried surface areas in protein-protein complexes,” *Protein Science*, vol. 22, no. 10, pp. 1453–1457, 2013, doi: 10.1002/pro.2330.

[56] M. C. Lawrence, “S c : measuring shape complementarity at protein- protein interfaces,” no. Figure 1, pp. 1–5, 1999.

[57] 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*, vol. 32, no. 16, pp. 2451–2456, 2016, doi: 10.1093/bioinformatics/btw197.

[58] R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov, “Shape complementarity at protein–protein interfaces,” *Biopolymers*, vol. 34, no. 7. pp. 933–940, 1994. doi: 10.1002/bip.360340711.

[59] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, “Design of a Novel Globular Protein Fold with Atomic-Level Accuracy Downloaded from,” 2003. [Online]. Available: http://science.sciencemag.org/

[60] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker, “A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins,” *J Mol Biol*, vol. 332, no. 2, pp. 449–460, 2003, doi: 10.1016/S0022-2836(03)00888-X.

[61] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[62] J. Chen, N. Sawyer, and L. Regan, “Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area,” *Protein Science*, vol. 22, no. 4, pp. 510–515, 2013, doi: 10.1002/pro.2230.

[63] T. Steinbrecher *et al.*, “Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability—Large-Scale Validation of MD-Based Relative Free Energy Calculations,” *J Mol Biol*, vol. 429, no. 7, pp. 948–963, 2017, doi: 10.1016/j.jmb.2016.12.007.

[64] J. A. Marsh, “Buried and accessible surface area control intrinsic protein flexibility,” *J Mol Biol*, vol. 425, no. 17, pp. 3250–3263, 2013, doi: 10.1016/j.jmb.2013.06.019.

[65] T. Lynagh *et al.*, “Peptide Inhibitors of the α-Cobratoxin–Nicotinic Acetylcholine Receptor Interaction,” *J Med Chem*, vol. 63, no. 22, pp. 13709–13718, Nov. 2020, doi: 10.1021/acs.jmedchem.0c01202.

[66] Y. Bourne, T. T. Talley, S. B. Hansen, P. Taylor, and P. Marchot, “Crystal structure of a Cbtx-AChBP complex reveals essential interactions between snake α-neurotoxins and nicotinic receptors,” *EMBO Journal*, vol. 24, no. 8, pp. 1512–1522, 2005, doi: 10.1038/sj.emboj.7600620.

[67] T. Lynagh *et al.*, “Peptide Inhibitors of the α-Cobratoxin-Nicotinic Acetylcholine Receptor Interaction,” *J Med Chem*, vol. 63, no. 22, pp. 13709–13718, 2020, doi: 10.1021/acs.jmedchem.0c01202.

[68] Y. Bourne, T. T. Talley, S. B. Hansen, P. Taylor, and P. Marchot, “Crystal structure of a Cbtx – AChBP complex reveals a -neurotoxins and nicotinic receptors,” vol. 24, no. 8, pp. 1512–1522, 2005, doi: 10.1038/sj.emboj.7600620.

[69] S. Antil, D. Servent, and A. Ménez, “Variability among the sites by which curaremimetic toxins bind to Torpedo acetylcholine receptor, as revealed by identification of the functional residues of α-cobratoxin,” *Journal of Biological Chemistry*, vol. 274, no. 49, pp. 34851–34858, 1999, doi: 10.1074/jbc.274.49.34851.

[70] O. Trémeau *et al.*, “Genetic engineering of snake toxins: The functional site of erabutoxin A, as delineated by site-directed mutagenesis, includes variant residues,” *Journal of Biological Chemistry*, vol. 270, no. 16, pp. 9362–9369, 1995, doi: 10.1074/jbc.270.16.9362.

4. Application of *in-silico* tools to design 'ideal' binding proteins

4.1. Background information

Antibodies (Abs) are excellent therapeutic binders that are used in today's world to treat diseases and chronic infections. However, as previously discussed in the previous chapters, they have certain limitations. Therefore, it is important to move beyond Abs and arrive at a solution to design a binder, which might be comparable to an Ab in terms of its effectiveness and physicochemical properties.

It can be concluded from the previous chapters that the most important criterion of a good binding protein is its stability i.e., its ability to withstand high temperatures, undergo mutations and still be able to fold to its native structure, and able to withstand sheer stress and tendency to refold due to its withdrawal. For a protein to be stable, it needs to have a strong core that has resulted from the hydrophobic collapse during its folding process. Protein structures exist in several forms in nature. However, according to research, there is one kind of structure in particular that is considered ideally to be one of the most stable structures called the alpha-helical bundle with minimal loops[1], [2], [3], [4]. According to a Rosetta simulations study, it was found that the folding ability of a tertiary protein motif majorly depends on the secondary structure lengths and the connecting loops, and not necessarily on the residue sequence details[5]. These findings were written down as rules to generate a stable ideal protein structure[1], [5]. These α-helices bundles form a close unit linked together by loops or sheets, which gives rise to the solid hydrophobic protein core that massively contributes to the structural stability.

If such a protein is designed with well-engineered characteristics such as high solubility and thermostability, low immunogenicity, and additional beneficial characteristics of a binding protein with the loops joining these helices to be used as antigen binding paratopes then we can potentially

arrive at a highly stable protein that mimics the function of an antibody (Ab). Although several protein structures in the PDB resemble alpha-helical cylinders, for them to mimic an Ab, they must remain connected with coils on both ends (Figure 1) so that they can be grafted to fit appropriate loops to bind to a target protein as can be accomplished by an Ab. However, to accomplish all of the above tasks to engineer any protein successfully several mutations need to be incorporated throughout its stable framework such that these mutations are not detrimental to the final folding of the protein to its native structure. Therefore, it must be ensured that these sequence mutations are collectively responsible for the protein's native folding state to outweigh all its non-native folding states. For this reason, it is also necessary to emphasize the stability of these point mutations. However, it is nearly impossible to find a single mutation for an amino acid, which will positively affect its thermostability, solubility as well as reduction in immunogenicity because one optimization will always come at the cost of another for the same amino acid. Thus, the thermodynamic stability of these point mutations must be very carefully considered before mutating them, be it on the protein surface or the protein core because an incorrect mutation can prove to be detrimental to the folding of the entire protein.

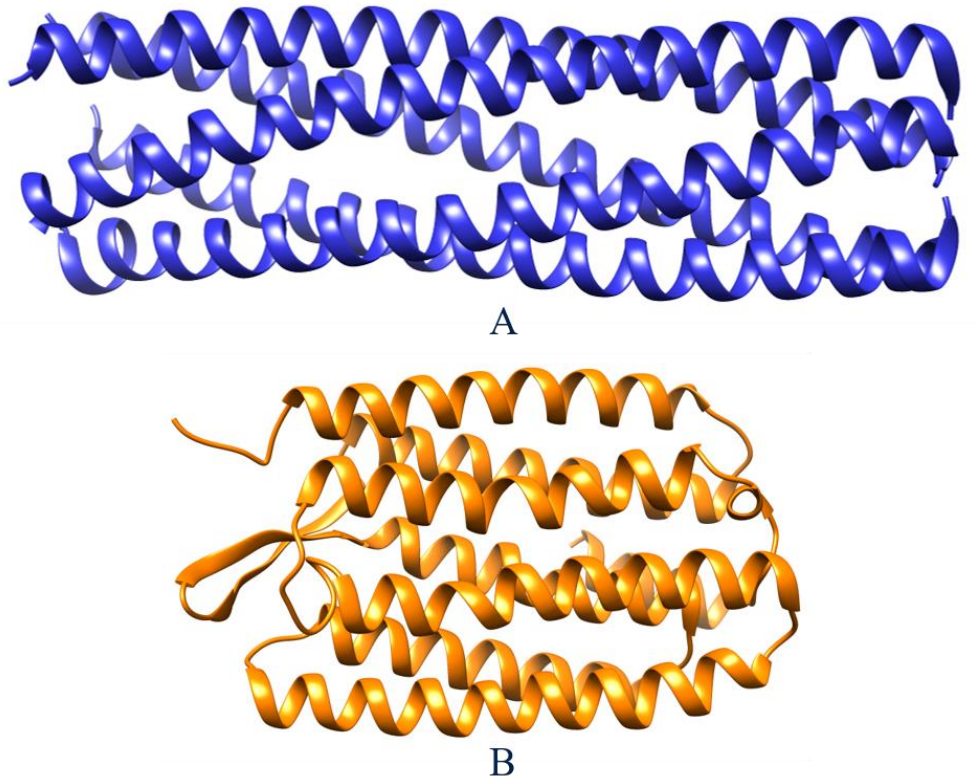

Figure 1. **Alpha helical bundle and coiled coil difference**. The structures represent helical coils obtained from the RCSB PDB. Figure A represents a coiled coil (PDB 2GUV), which cannot be appropriate to engineer for this project because it lacks the end terminal loops/coils that connects these helices. Figure B represents an engineered 'ideal' protein (Scaffold PDB 3HAS), where the helices are connected with minimal loops at the terminals.

4.2. Methods

4.2.1. Target protein scaffold selection

The RCSB PDB[6] (https://www.rcsb.org) was extensively searched for the suitable protein scaffold of choice i.e., 6-7 alpha helical bundles. The search was conducted using the names that these scaffolds are commonly referred to as e.g., 6/7 helical bundle(s), alpha helical bundles, alpha helices, helical coils, and coiled coils. This was done to prevent potentially favorable structures from escaping the search space. While conducting this search, it was understood that very few structures would satisfy the search criteria. Finally, 75 PDB structures obtained were alpha helical bundles. The remaining structures were discarded, as even though their framework could be engineered, they lacked the terminal loops/coils which joined the helices together. That means they had individual chains or helices, which did not fit the desired scaffold description.

4.2.2. Measuring solubility of the PDB files

After the selection of the appropriate scaffolds, their solubility values were calculated using the CamSol server's Intrinsic solubility profile calculator[7], [8] and SoluProt solubility servers[9]. The reason for using both tools is because these were predictions and it was unlikely that both would produce the same value of solubility. The scale was different for both the tools e.g., CamSol had scores (-1,1) whereas SoluProt ranged from 0-5. It was also known that the profiles produced would be highly insoluble since the proteins were membrane proteins and transport proteins. It was still necessary to calculate their solubility profiles to characterize them eventually. Later, all calculations and fine-tuning of solubility were accomplished using CamSol. The University of Cambridge developed a variety of tools dealing with protein solubility (https://www-cohsoftware.ch.cam.ac.uk/) i.e., their rapid and accurate prediction in varying pH, improving the protein solubility profiles using point mutations that also consider the stability of the protein

structure[10], and more. According to their article [7] by Sormanni et. al, the intrinsic solubility profile was established by using the secondary structure propensities whose values were all calculated from the PDB itself using the structures at its 50% sequence identity and a hydrophobicity scale, which was adapted from the Wimley–White scale[11]. This helped in removing additional biases towards calculating solubilities of amyloid-like protein aggregation and also in the case of residues such as proline and glycine which disfavor the formation of beta sheets and thus against the formation of amyloids. Also, residues with a score less than -1 are considered to be poorly soluble and hurt the overall protein solubility while scores larger than 1 represent high solubility, which positively impacts the overall protein solubility. To calculate the solubility, the protein sequence or the FASTA file was needed at the specific pH (7.0 in general) and the results were displayed within a minute.

4.2.3. Measuring thermostability of the PDB files

The PDB files were individually subjected to predicting solubility and thermostability at the beginning to help narrow down the list based on the value of these parameters. After the protein solubility profiles were calculated using the CamSol server, the individual PDB files were also subjected to their thermostability prediction using the well-known SCooP server[12], [13], [14], which is a fast and accurate protein thermostability prediction tool. It is based on the famous thermodynamic Gibbs-Helmholtz equation (Equation 1) associated with the protein's folding process, where T_m is the protein melting temperature, ΔC_p the standard heat capacity of folding, ΔH_m the standard enthalpy of folding at T_m and $\Delta G(T)$ is the folding free energy, which is a function of temperature T. Along with this information it also yields other thermodynamic properties associated with protein folding like any change in enthalpy, heat capacity, the melting temperature, and free energy.

$$\Delta G(T) = \Delta H_m \left(1 - \frac{T}{T_m}\right) - \Delta C_p \left(T_m - T + TLog\left[\frac{T}{T_m}\right]\right) \quad \text{Eq. 1}$$

The authors constructed 4 different databases and used different equations to model the individual outputs for a particular protein of choice. To submit a job on the SCooP web server, the PDB file of choice is required along with the target chain name. Once these two are submitted, the user is redirected to a page, which contains several host organisms of expression. In this study, for the 75 different PDB files, their different host organisms were selected and the queries were submitted. The outputs generated were all within one minute, which was very rapid and accurately fits the description of this tool.

4.2.4. Discarding synthetic proteins & proteins with surface Cys residues

The presence of Cysteine residues on the protein surface was also considered for further filtering the structures from the remaining list. The reason is that Cys residues are sites for drug conjugation[15], [16]. There are two residues specifically that can assist in drug conjugation i.e., Cys and Lys. The mechanisms in which they function are different. However, a reason cited in the design limitations of antibodies is the presence of multiple surface exposed Cys residues for drug conjugation. This leads to uncontrolled dosage. This was also one of the reasons for DARPins being used as the alternative protein binders of choice in the previous chapter, which discusses DARPin – α-Cobratoxin designs. The reason cited was that DARPins do not have any Cys residues, which makes them easy to engineer for drug conjugation[17], [18]. All 75 PDB files were visually inspected in UCSF Chimera for Cys residues. The ones with Cys residues were discarded. Furthermore, the ones being selected were made sure not to display steric clashes among the loops situated at both ends, and with the incorrect positioning of a loop, shifted from the framework. PDB files that did not adhere to these categories were discarded. Finally, *de novo* designed or synthetic alpha helical bundle structures were also discarded due to future copyright issues.

These were the main criteria to screen the structures based on their physico-chemical and structural characteristics. Based on this process, the PDB files and their homologs, which were shortlisted were subjected to protein multiple sequence alignment (MSA) using Clustal Omega[19], sequence design using RosettaDesign[20], [21], [22], protein solubility and thermostability improvement using CamSol[10], PoPMuSiC[23] and HoTMuSiC[24], [25] respectively, de-immunizing the protein sequences using IEDB's[26] De-immunization tool[27] and finally predicting the new protein sequences using AlphaFold[28] and RosettaFold[29].

4.2.5. Multiple Sequence Alignment using Clustal Omega

Considering that the proteins, which were selected had poor solubility and high immunogenicity, the primary motive was to redesign the sequences to provide some relative improvement to these factors. The proteins and their homologs were aligned using multiple sequence alignment (MSA) algorithms e.g., Clustal Omega[19], to check for mutational sites. These sites can help incorporate the desired mutations to steer the protein scaffold properties to the desired ones. However, very few such sites were identified, and such mutations would not be sufficient to improve the desired protein properties. It is for this reason that RosettaDesign was used to redesign the entire scaffold sequences.

4.2.6. Protein sequence design using RosettaDesign

The RosettaDesign[20], [30] application was used to re-design the entire sequence of the shortlisted alpha-helical bundle protein scaffolds. This method is used to optimize sidechain positioning or create a new sequence of residues on the original backbone of the protein. The most common example of a *de* novo designed protein structure is Top7[30]. Top7 was a 93 residue α/β highly soluble, stable protein, which was designed by Dr. Kuhlman and Dr. Dantas in the Baker's Laboratory at the University of Washington using their protein design and structure prediction

algorithms. They have further improved their algorithm using artificial intelligence and machine learning techniques known as RFDiffusion[31]. However, this was the most commonly used tool to design protein sequences when working on this project and RFDiffusion was not developed back then. The PDB structure filenames were listed in a text file 'pdblist.txt' and the number of iterations i.e., number of designs was set to fifty. The database was set to the default main Rosetta database and the latest weight function, ref2015.wts was used for scoring purposes. The linear-memory interaction graph was set as the Interaction Graph parameter. Any option for sequence similarity was disregarded because the sequences are being re-designed so that there is a relatively higher chance of the end product being more soluble and not necessarily focusing on sequence recovery. The remaining design parameters were kept to the default recommended settings. All information regarding these design parameters can be found in the RosettaDesign tutorials page under Application documentation and they are well detailed in their specific links on the RosettaCommons website. The design run was accomplished using the Auburn University HPC Easley Cluster and the following are the job specifications: 10 cores in one general node were used for this job for each of the five design runs. All jobs were completed.

4.2.7. Protein solubility improvement using CamSol

During the first set of trial runs, the top designs were selected out of the 50 from each set of design runs. The CamSol combination method of automated structure-based solubility optimization[10] was used to improve the solubility and stability of the re-designed protein sequences. The standard protocol is to use a 'PDB Cleaner' or input a PDB file after removing any hydrogen atom, or hetero-atoms, except for the heavy atoms. An additional file i.e., a .zip archive file was required containing a .hhm or a .a3m file, which is generated with HHBLITS[32]. The MPI Bioinformatics Toolkit HHBLITS webserver was used to obtain such a file[32], [33].

The entire protein sequence was submitted in the query, which generated the .hhm file. The file was archived in a .zip file and it was renamed as the filename_chainname.hhm. This additional file is only necessary if the protein is not an Ab/Nb. The default setting for excluding target residue substitution was Met and Cys. However, Phe and Trp were also included in that setting making the list of residues to be excluded as Met, Cys, Phe, and Trp. 2 more trial runs were set with a list of excluded amino acids as Cys, Met, Ile, Leu, Phe and another list being Cys, Phe, Ile, Leu, Gly, Trp. Multiple trial runs were conducted with different excluded sets of amino acids because the fundamental idea behind this is to mutate the protein sequence in the absence of highly hydrophobic residues and find the best combination to generate a stable structure with the maximum solubility value. The remaining settings were set as recommended with the maximum number of simultaneous mutations set as eight. It is worth mentioning here that multiple jobs were submitted in the same window because the jobs got queued and they started executing one after another and not simultaneously. The results get mailed to the registered email address and the contents can be downloaded from the link in that email.

4.2.8. Protein thermostability improvement

To improve the protein thermostability, two tools were used i.e., HoTMuSiC, to improve the overall melting temperature, and PoPMuSiC, to improve the thermodynamic stability. The process was fairly simple, these stabilities were based on mutations, which again was dependent on the exposure of that residue to the solvent. For example, if there is a polar residue, which has less than 30% exposure to solvent molecules then there are 19 mutations to choose from, of which the non-polar options usually have higher melting temperature improvement or have a negative free energy suggesting a stabilizing mutation. It is vice versa for a residue that has more than 60-70% exposure to solvent molecules. In this case, it will need polar mutations as they will have a stabilizing effect.

However, the motive was not to apply such mutations all throughout the protein, instead, it was to apply minimum mutations to improve the overall protein thermostability significantly. The mutations that led to a change in melting temperature or negative change in free energy were considered. In almost all the cases, the mutations were the same for both HoTMuSiC and PoPMuSiC i.e., the residue change for higher melting temperature was the same for the stabilizing mutation. The user input is the entire protein structure either from the PDB or to use a private file, an account needs to be created in the Dezyme website. This is followed by selecting the option for which software to use. It takes a few minutes to generate the results and is not as fast as SCooP.

4.2.9. Protein sequence de-immunization

The IEDB De-immunization tool was used to humanize the protein sequences. The user input is the entire protein sequence in FASTA format, and the server will break the sequence into 15mers, which will overlap in 10 residues. Each 15mer will be predicted for 26 alleles using the default recommended method. The second input is to use a limit or a threshold that corresponds to the median percentile rank for each peptide sequence (15mer sequence). A lower percentile rank will lead to a higher binding affinity of the peptide sequence. If a higher limit is set, more peptide sequences will become immunogenic in the next step of the process. However, this was set to a default value of 20 as recommended. The next step will display all the immunogenic peptides predicted based on the user choices. The peptides will be subjected to all 285 possible mutations (19 mutations each for 15 amino acids i.e., 19 x 15 = 285). The relatively less immunogenic peptides have a higher median percentile rank. However, 8.5 is the recommended value for optimum de-immunization and it is kept the same. A maximum of 5 such 15mer sequences can be further selected to undergo the de-immunization process and is confined to the framework of the helical bundles because the loops will be subjected to swapping in the design algorithm. It is

recommended to provide an email address to keep the outputs in record as .csv files. This is followed by the output generation which lists a variety of mutational recommendations for each of the 15mer sequences along with their evaluation metrics. It is worth mentioning that the time increases with a higher number of 15mer peptide sequence selections. Adding separate jobs in separate tabs for simultaneous protein sequences might result in erroneous results or no results at all.

4.2.10. Protein structure prediction using RosettaFold

The protein structures were predicted from their mutated sequences using RosettaFold[29] in the Robetta[34] server. One is required to register and create an account on the website and select the structure prediction option. The protein sequence is copied into the blank space or the FASTA file is uploaded followed by naming the target/job. For minor mutations, it is recommended to apply the comparative modeling method with the original PDB file being uploaded to the server. However, since there were plenty of mutations in the protein sequences, RosettaFold (which outperforms all the other methods) was selected without uploading any homologous structure's PDB file. The job submission results were emailed to the registered email address. The results were also observable in the visualization panel on the Robetta website.

4.2.11. Incorporation of a surface exposed Cys residue for drug conjugation

The protein sequences were mutated to enhance solubility, and thermostability and reduce immunogenicity considering that there should not be any Cys residue during the process. The presence of a single surface accessible Cys residue will be an appropriate site for drug conjugation for controlled drug dosage. Post prediction using RosettaFold, a Cys residue was incorporated after carefully observing the structure in UCSF Chimera and selecting an optimum location around the center position on the surface of the helical bundle, so that the Cys residue remains accessible to

drugs and there will be no steric clashes during the loop grafting protein design process. The choice of residue, which was mutated in this case was hydrophobic to prevent any significant reduction in solubility in case a hydrophilic surface exposed residue was mutated.

The fine-tuning of physico-chemical properties of the proteins was accomplished in 2 iterative rounds. The methodology mentioned above is common in all of the rounds. The chronology is to improve solubility, and thermostability, reduce immunogenicity, incorporate a single surface exposed Cys residue, predict the protein structure, and quantify these characteristic properties. Once the structures that fold appropriately are obtained with such optimized properties, the next set of structures is made to go through these exact steps. Therefore, instead of mentioning the entire process for different rounds, it will be mentioned only once for all of them considering the example that folds at the end with positive results for the fine-tuned properties.

4.3. Results

4.3.1. Scaffold selection and properties measurement

Almost 100,000 structures were searched and 75 structures were selected from the PDB. These 75 structures were either synthetic designs with high solubility since they were all optimized *de novo* designs or were all membrane proteins, with very poor solubility profiles. Their solubility profiles were predicted using CamSol, which generated a very poor range of solubilities as shown in Figures 4.a and 4.b. These proteins were mostly *de novo* designed, or had surface exposed Cys, or had undesired position of an end loop, or had steric clashes amongst the end loops, so many had to to discarded. However, some estimation was necessary to understand the range of mutations needed to solubilize these proteins. For this reason, the SoluProt solubility values were also considered for the initial screening. Similarly, the SCooP predicted thermostability profile was roughly around 65ºC as can be observed from Figure 5, which meant that such temperature range was sufficient to prevent the protein from degrading at room temperatures. These were the wild-type scaffold predicted structural properties as obtained from the PDB, and their values changed when they were cleaned due to the removal of miscellaneous atoms (SO4, NAG, FE, etc.). Even after this round of shortlisting, there were structures with more loops or an extra helix, which would have been detrimental to the structure. They had relatively worse values of both solubility and thermostability when the PDB was cleaned compared to their original wild type structures. Therefore, these structures were screened one final time before starting the sequence designing process. As shown in Figures 2 and 3 , only the best 5 structures i.e., PDBs 3HAS, 4KME, 1PY6, 1PXS, and 4Y9H, which passed all of the above screening parameters were selected for the final engineering process that would fine-tune their properties to steer them to the protein of our choice.

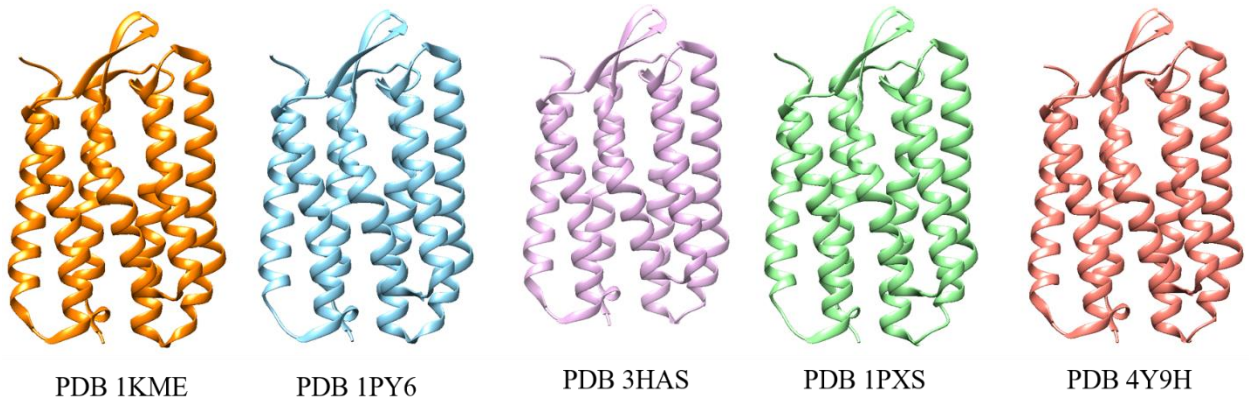

Figure 2. **Wild type helical bundle scaffold's structure**. Fully crystallized WT structures of the top 5 shortlisted alpha helical bundles from the RCSB PDB as observed in UCSF Chimera.

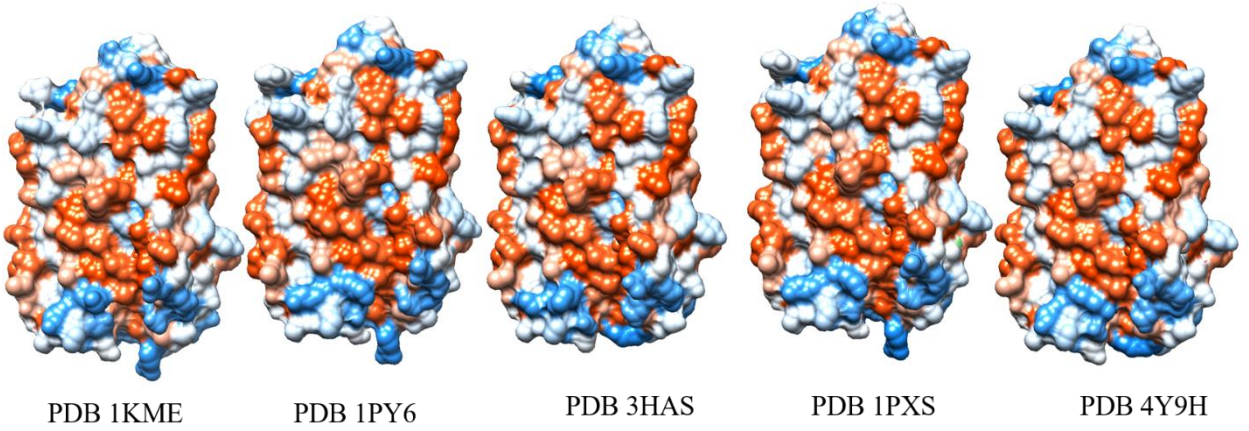

Figure 3. **Wild type structures hydrophobicity profiles**. As observed in UCSF Chimera, these are the hydrophobic profiles of the top 5 PDB structures obtained from the PDB before they were engineered to improve their physicochemical properties.

The next few sections will explain the engineering of hundreds of protein structures and their physico-chemical properties. Since it is impossible to consider individual scaffolds and explain them, only one such scaffold will be explained from improving its solubility to de-immunizing it to prevent unnecessary immune reactions. However, the same steps have been followed to engineer each of them.

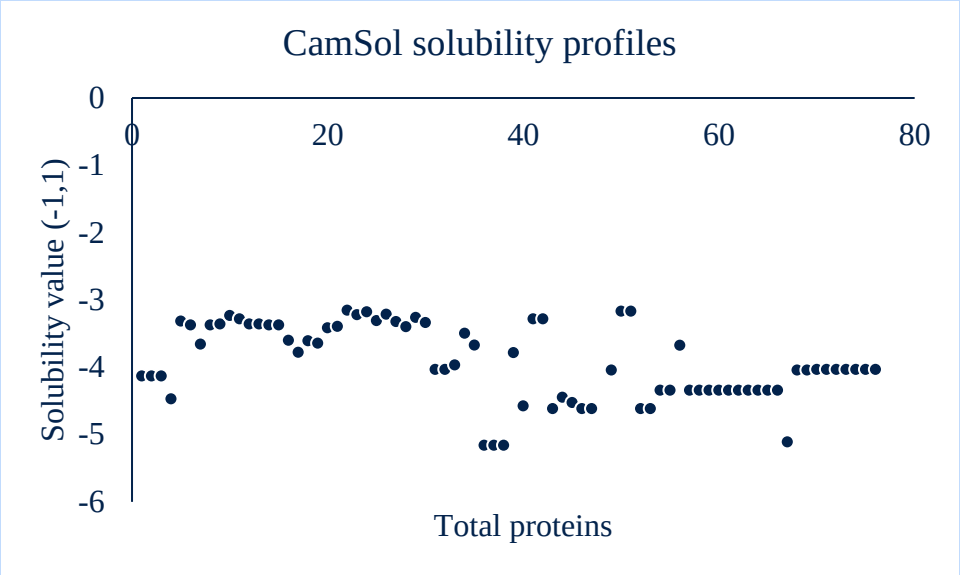

Figure 4.a. **Wild type structures solubility prediction using CamSol server.** Clearly, these represent poor solubility, which should be the case since the proteins are either membrane or transport proteins. However, plenty of *de novo* designed helices are highly soluble, that were produced in the past. Henceforth, all calculations pertaining to protein solubility has been accomplished and predicted using the CamSol server and none other.

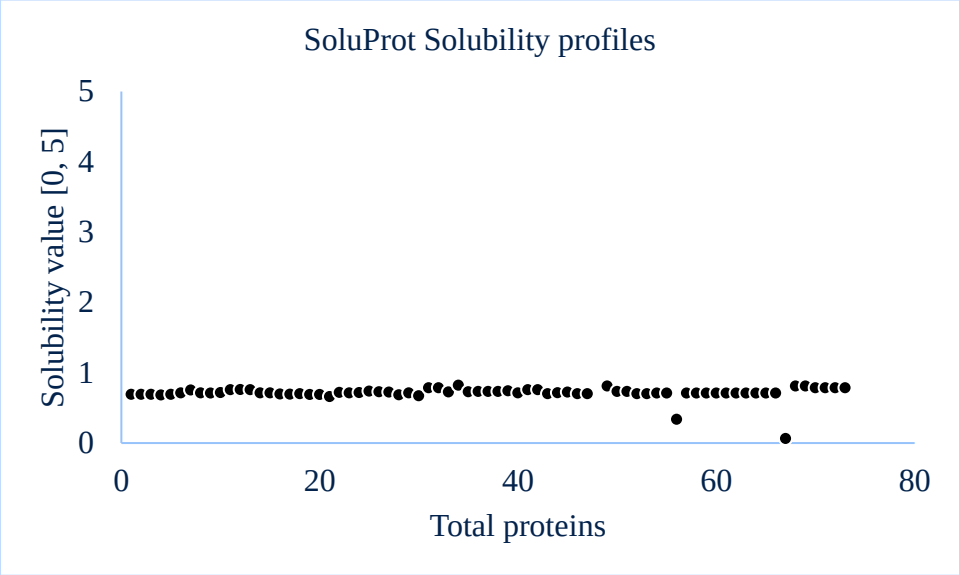

Figure 4.b. **Wild type structures solubility prediction using SoluProt server**. They also represent poor solubility, which should be the case since the proteins are either membrane or transport proteins. However, plenty of *de novo* designed helices are highly soluble, that were produced in the past. Additionally, all calculations pertaining to protein solubility has been accomplished and predicted using the CamSol server because it is a complete suite and quantification of several tools using a common data set will prevent erroneous results.

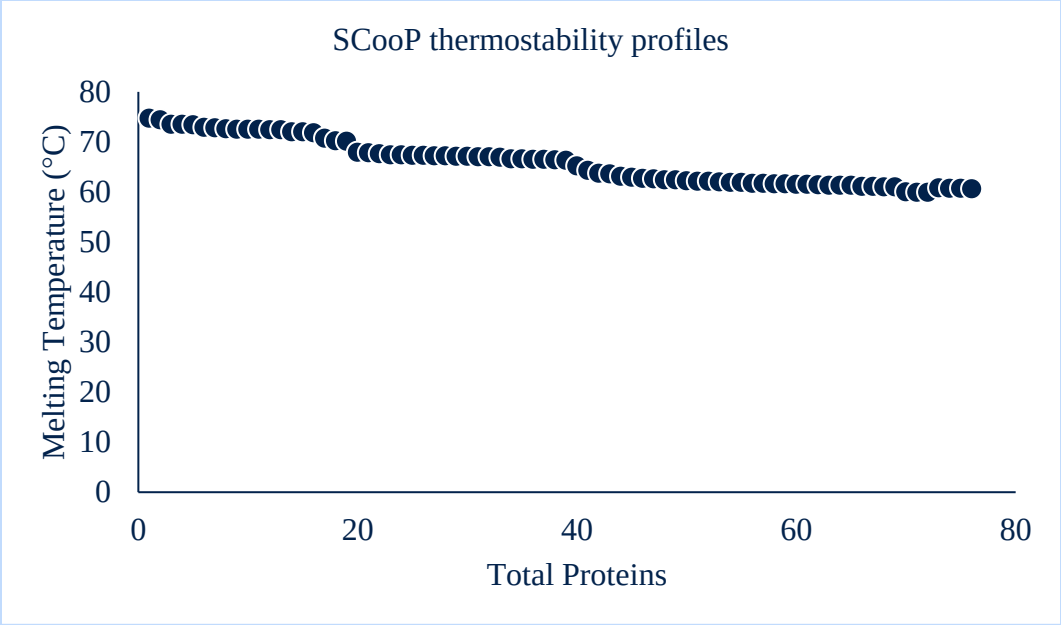

Figure 5. **SCooP thermostability prediction of wild type structures**. It can be observed that the helical bundles have an average temperature of around 65ºC, which is sufficient for a protein to not degrade at room temperatures.

4.3.2. RosettaDesign results

RosettaDesign was used to re-design the top 5 protein genetic sequences, their solubility and thermostability were improved using the tools mentioned previously and they were de-immunized as well. Each design run took approximately 50-55 hours to complete except for 1, which took approximately 93 hours. After incorporating the mutations and using RosettaFold to predict the structures, the only structure that correctly folded back to its native state was the design from PDB 3HAS (mutated). It had an improved T_m of 65.1ºC from 61.4ºC of its WT sequence, an improved CamSol solubility score of 1.58 from -3.21 and it was completely de-immunized. Thus, PDB 3HAS was used as the starting point for the next rounds instead of trying to engineer the other designs. The top 24 out of the 50 PDB 3HAS RosettaDesign designed sequences were selected for further solubility improvement based on their improvement in solubility and thermostability scores. Figures 3 and 9 show the change in these physicochemical profiles of the structures just due to the sequence redesign. They show a significant improvement in the protein solubility

profiles and the thermostability profile remained almost the same or increased slightly compared to the average WT values.

Table 1. After first round of engineering, 1 out of 5 PDBs got destabilized during solubility mutations. However, the remaining underwent the fine-tuning process and RosettaFold was used to predict their structures. As shown in the table only 1 folded correctly, PDB 3HAS. This was considered as the starting point for the remaining rounds of engineering.

PDB Files	**ScOOP (°C)**	**CamSol Solubility**	**Rosetta Fold Results**
	T_m	(-1,1)	
1kme	**62.4**	**-3.659132**	**Inaccurate folding**
1kme_3	66.5	1.449651	
1kme_13	63.7	1.382326	
1py6	**61.3**	**-3.282869**	**Inaccurate folding**
1py6_4	58.5	0.771455	
1py6_12	60.8	0.861351	
3has	**61.4**	**-3.21332**	**Correct native state folding**
3has_4	66.8	-0.876442	
3has_13	65.1	1.57854	
4y9h	**60.8**	**-3.671735**	**Inaccurate folding**
4y9h_8	62.7	1.528744	
4y9h_12	63.1	1.525698	

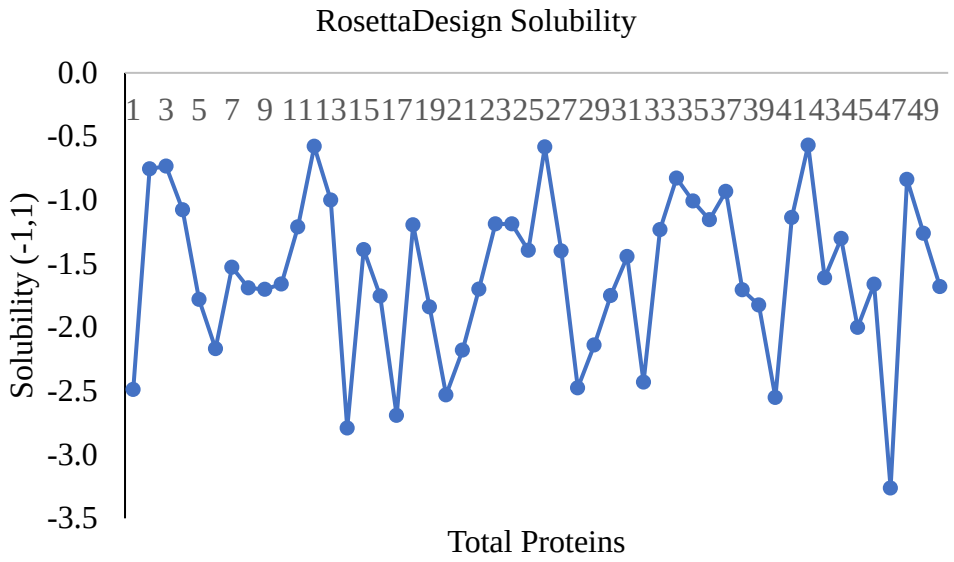

Figure 6. **Post RosettaDesign Solubility enhancement plots**. The solubility profiles of 50 PDB 3HAS designs obtained using the RosettaDesign algorithm. The difference in solubility is only due to the result of sequence design and no additional engineering has been done.

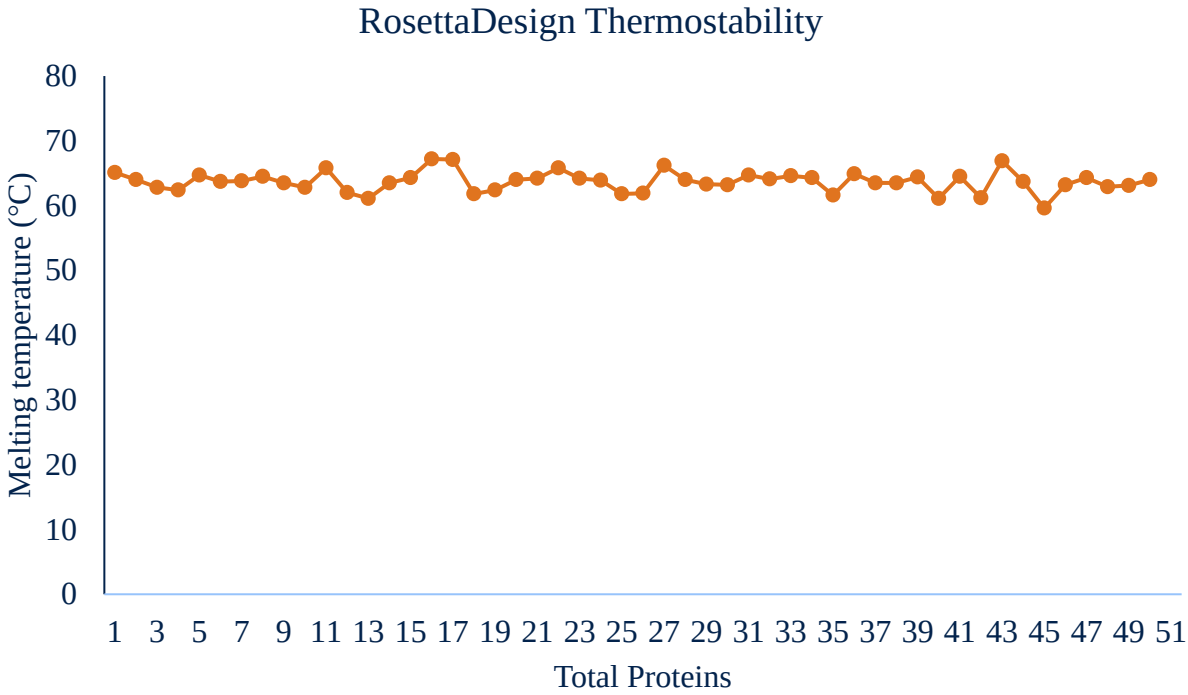

Figure 7. **Post RosettaDesign thermostability plots**. Although in Figure 6, a significant improvement in solubility was observed, there was not much of a change in the thermostability stability profile of the new designs compared to the WT protein sequences. Neither was the sequence re-designing detrimental to the thermostability of the structures, which suggests that they are innately stable.

4.3.3. Protein solubility and thermostability fine-tuning

The designs were marked 1-50. Let's take the example of design 12, which had an improved T_m of 62ºC and a solubility value of -0.578. CamSol's multiple mutation scanning report showed Cys and Met being excluded as the only two residues during mutations so that many solutions can be generated. 22 solution models were generated out of which model 15 ranked 1 in both solubility and stability. A total of 70 mutation sites were identified based on several factors, solvent exposure being one of them. Model 15 had 14 simultaneous mutations that was marked as xZABCy where x was the original residue, Z refers to the protein chain, ABC was the residue position and y was the newly mutated residue. The mutations were TA5Q, KA38R, CA49I, AA110R, TA120L, FA135Y, AA142S, VA151L, GA155R, SA158R, DA161E, TA169K, IA179L, and VA203Q. This was followed by using HoTMuSiC's T_m fine-tuning results. There were plenty of mutation recommendations (19 mutations for each protein residue); however, a single mutation recommendation at position 155, Gly to Arg showed an increment of 1.17ºC. If we look at the CamSol mutations, this was the same recommendation as GA155R, where A refers to the protein chain. Both of these tools consider the exposure of the residue to the surrounding solvent. Therefore, it is expected that this was a stabilizing mutation since two entirely different *in-silico* tools predicted the same mutations.

4.3.4. Protein de-immunization

As stated previously in the methods section, only a maximum of 5 15mer peptides undergo humanization calculations at once. The common practice was to humanize these 5 15mer peptides, and post humanization, feed this humanized sequence again to the De-immunization tool for further humanize the remaining regions. It has not happened yet that any sequence required back-to-back 3 shots of de-immunization calculations. To humanize an amino acid, it needs to have a

low deimmunization score (1, 10), to ensure that its neighboring amino acids remain unaffected or become less immunogenic. Additionally, the residue shows a high median difference, which means a higher degree of immunogenicity reduction in that amino acid. Unfortunately, there was no way to quantify if the sequence is completely humanized except for feeding it to the de-immunization tool, till it does not generate any more peptide sequences. The residue mutations for this process were L181D, I185R, I52D, A57D, F208D, 1211P, A214T, I148D, F156C, I89D, A85P, L94E, F171D, L174W, L179E, F43P, L177D, F112D, L117D, and F42D.

Apart from the mutations that were suggested in all of these processes, a few other important factors were kept in mind to conduct this research. It was made sure that no mutations crossed the effect of other mutations from a different tool e.g. if for ‘x’ residue at position ABC, a solubility mutation ‘b’ has been introduced and a de-immunization mutation ‘c’ has been recommended then the latter option is discarded and other potential mutation sites are considered since there are more de-immunization mutation options. For tools that provide a plethora of mutations, the neighboring residues, the mutating residues, and the stabilizing effect of the neighboring residues are considered before the mutation e.g., if there is an option to mutate a residue to Asp in two positions one with a neighbor Arg and another with a neighbor Asp, the one with Arg will be selected for intra-residue stabilization. Furthermore, solubility and thermostability effects are also considered before mutating residues e.g., given a chance to choose between Ser and Ile for a position with the same deimmunization score and insignificant median difference, Ser will be selected because it is hydrophilic and will reinforce protein solubility compared to Ile. Although it was known right from the beginning that using separate tools to engineer protein properties can result in improving one at the cost of another, care was taken to make sure that all factors were considered before

mutating a residue so that a single mutation could improve all of the properties and steer the protein towards its desired end goal.

4.3.5. Final structure prediction using RosettaFold

RosettaFold was used to predict the structures of all the newly designed sequences. The success rate was better than expected as out of 24 structures 16 folded (66.67%) to their native state including 12_15, with which the above engineering process was explained. Additionally, the structures were carefully observed in UCSF Chimera, and a Cys residue was planted around the middle of the structure for most of the structures in place of a Met. It was near the middle with no scope for steric clash when the loops were being used to bind a target antigen. Post sequence re-design, SCooP, and CamSol were further used to determine the T_m calculated at 59.4ºC and a solubility value of 1.55 respectively. Figure 6 below shows the image of a predicted, humanized, soluble, and thermostable 'ideal' binding protein. Table 1 shows a list of designs followed by their solubility and thermostability profiles along with the WT PDB 3HAS and the mutated PDB 3HAS, which served as the starting point for the remaining models.

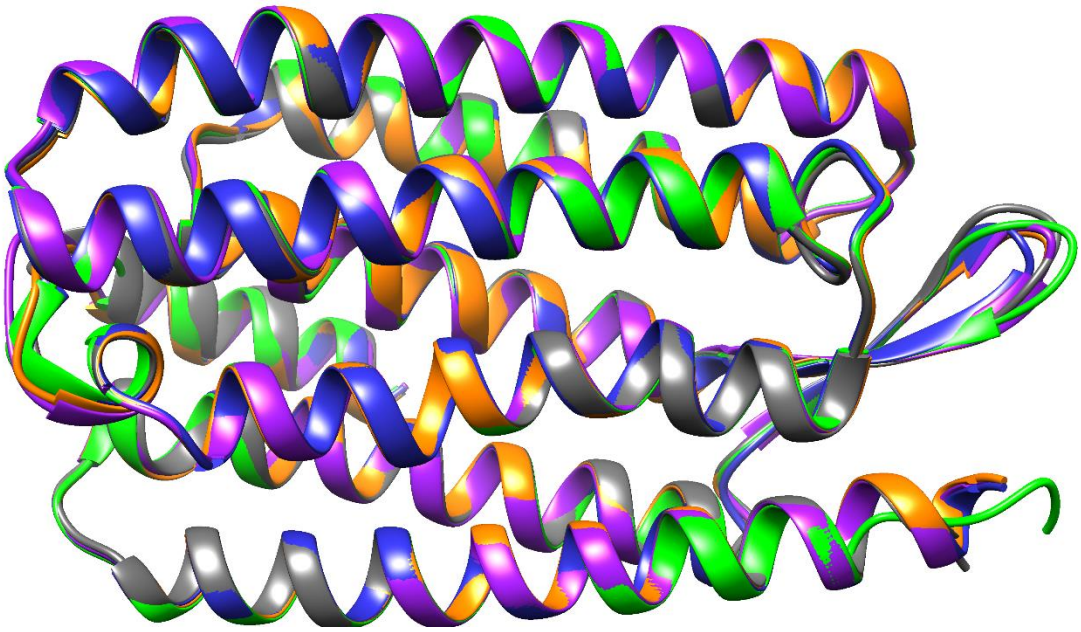

Figure 8. **Final predicted *Ideal* binding proteins structures**. Rosetta designed 'ideal' protein was predicted using RosettaFold and observed in UCSF Chimera and the matchmaking feature was used to superimpose all the 5 predicted structures to see their difference. As can be observed from the figure, there is extremely small difference to none. The RMSD average was around 0.4Å for these structures. The loops at the end connecting the helical framework will be grafted to bind target antigen.

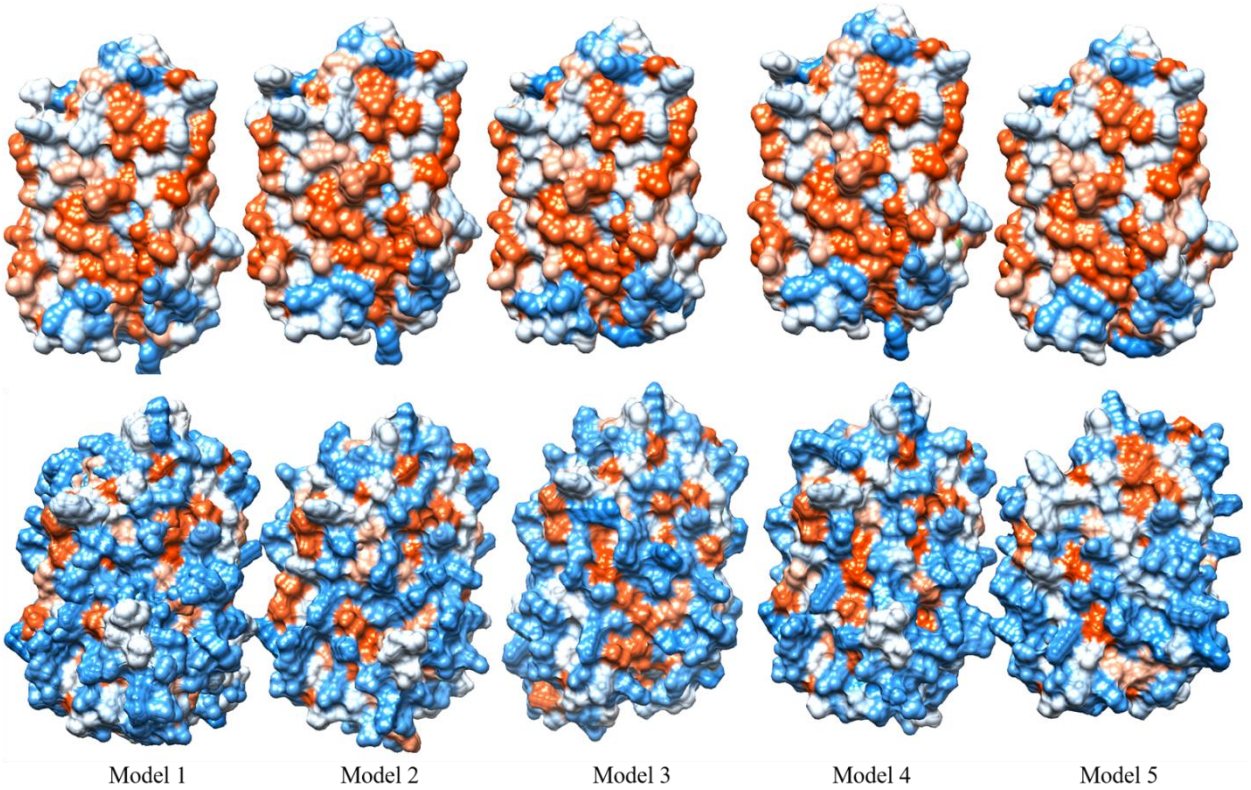

Figure 9. **Before and after hydrophobicity profiles of engineered scaffolds.** A before and after solubility engineered image of the alpha helical bundles. The top 5 structures were the ones with the WT sequences, directly obtained from the PDB. The bottom ones underwent a full round of fine-tuning and the hydrophobic profile difference is clear from visualization in UCSF Chimera.

Table 2. The final 16 'ideal' protein designs after several rounds of engineering and mutations. Unfortunately, no immunogenicity quantification tool was found to calculate their immunogenicity. And each proteins sequence had multiple hundreds of de-immunization mutations and factors, which was impossible to be tabulated in this chapter.

Structure	SCooP Thermostability (ºC)	CamSol Solubility (-1,1)
PDB 3HAS (WT structure)	61.4	-3.21332
PDB 3HAS 1st Round	65.1	1.55854
Design_1	61.3	1.553451
Design_2	63.9	1.752526
Design_3	61.6	1.754562
Design_4	67.8	1.368982

Design_5	62.5	0.928521
Design_6	66.4	1.022473
Design_7	61.8	1.403472
Design_8	64.7	1.716237
Design_9	63.8	1.243881
Design_10	61.8	1.033810
Design_11	56.3	1.228967
Design_12	65.5	1.248610
Design_13	59.4	1.554915
Design_14	59	1.138238
Design_15	60.9	1.162546
Design_16	62.3	1.799447

4.4. Discussion

In this chapter, advanced computational tools were used to convert a highly insoluble immunogenic protein to a soluble humanized version. The PDB was searched and out of almost 100,000 structures, 75 were shortlisted, whose physicochemical properties were predicted using SCooP, CamSol, and SoluProt. Based on that, they were further narrowed down to only 5 structures, and their sequences were redesigned. Their properties were fine-tuned and the sequence was predicted. Out of 5 structures only 1 folded back to its native state. That was selected as the starting point for the next rounds of trials. These steps were repeated for the remaining 23 out of 50 structures, which were shortlisted based on their properties. At the end 16/24 re-designed proteins folded back to their native state. This proves that with the appropriate tool and the basic knowledge of thermodynamics, the physicochemical properties of proteins can be fine-tuned to a great extent without destabilizing them.

The next chapter will discuss some small projects that were done during my graduate degree and their results. It will be followed by the future prospects of these projects and final chapter of my dissertation, showing the direction these projects will continue in the future.

References

[1] N. Koga *et al.*, “Principles for designing ideal protein structures,” *Nature*, vol. 491, no. 7423, pp. 222–227, 2012, doi: 10.1038/nature11600.

[2] R. Koga *et al.*, “Robust folding of a de novo designed ideal protein even with most of the core mutated to valine,” *Proc Natl Acad Sci U S A*, vol. 117, no. 49, pp. 31149–31156, 2020, doi: 10.1073/pnas.2002120117.

[3] P. S. Huang *et al.*, “High thermodynamic stability of parametrically designed helical bundles,” *Science (1979)*, vol. 346, no. 6208, pp. 481–485, 2014, doi: 10.1126/science.1257481.

[4] R. B. Hill, D. P. Raleigh, A. Lombardi, and W. F. Degrado, “De novo design of helical bundles as models for understanding protein folding and function,” *Acc Chem Res*, vol. 33, no. 11, pp. 745–754, 2000, doi: 10.1021/ar970004h.

[5] R. Koga and N. Koga, “Consistency principle for protein design.,” *Biophys Physicobiol*, vol. 16, pp. 304–309, 2019, doi: 10.2142/biophysico.16.0_304.

[6] S. K. Burley *et al.*, “RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D464--D474, 2019, doi: 10.1093/nar/gky1004.

[7] P. Sormanni, F. A. Aprile, and M. Vendruscolo, “The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility,” *J Mol Biol*, vol. 427, no. 2, pp. 478–490, 2015, doi: https://doi.org/10.1016/j.jmb.2014.09.026.

[8] P. Sormanni, L. Amery, S. Ekizoglou, M. Vendruscolo, and B. Popovic, “Rapid and accurate in silico solubility screening of a monoclonal antibody library,” *Sci Rep*, vol. 7, no. 1, p. 8200, 2017, doi: 10.1038/s41598-017-07800-w.

[9] J. Hon *et al.*, “SoluProt: prediction of soluble protein expression in Escherichia coli.,” *Bioinformatics*, vol. 37, no. 1, pp. 23–28, Apr. 2021, doi: 10.1093/bioinformatics/btaa1102.

[10] A. Rosace *et al.*, “Automated optimisation of solubility and conformational stability of antibodies and proteins,” *Nat Commun*, vol. 14, no. 1, p. 1937, 2023, doi: 10.1038/s41467-023-37668-6.

[11] W. C. Wimley and S. H. White, “At Membrane Interfaces,” *Nature Strutural Biology*, vol. 3, no. 10, pp. 842–848, 1996.

[12] F. Pucci, J. M. Kwasigroch, and M. Rooman, “SCooP: an accurate and fast predictor of protein stability curves as a function of temperature.,” *Bioinformatics*, vol. 33, no. 21, pp. 3415–3422, Nov. 2017, doi: 10.1093/bioinformatics/btx417.

[13] F. Pucci and M. Rooman, “Towards an accurate prediction of the thermal stability of homologous proteins.,” *J Biomol Struct Dyn*, vol. 34, no. 5, pp. 1132–1142, May 2016, doi: 10.1080/07391102.2015.1073631.

[14] F. Pucci and M. Rooman, “Stability Curve Prediction of Homologous Proteins Using Temperature-Dependent Statistical Potentials,” *PLoS Comput Biol*, vol. 10, no. 7, p. e1003689, Jul. 2014, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003689

[15] D. Shinmi *et al.*, “One-Step Conjugation Method for Site-Specific Antibody-Drug Conjugates through Reactive Cysteine-Engineered Antibodies,” *Bioconjug Chem*, vol. 27, no. 5, pp. 1324–1331, May 2016, doi: 10.1021/acs.bioconjchem.6b00133.

[16] A. Wagh, H. Song, M. Zeng, L. Tao, and T. K. Das, “Challenges and new frontiers in analytical characterization of antibody-drug conjugates,” *MAbs*, vol. 10, no. 2, pp. 222–243, 2018, doi: 10.1080/19420862.2017.1412025.

[17] 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.

[18] M. T. Stumpp, H. K. Binz, and P. Amstutz, “DARPins: A new generation of protein therapeutics,” *Drug Discov Today*, vol. 13, no. 15–16, pp. 695–701, 2008, doi: 10.1016/j.drudis.2008.04.013.

[19] F. Sievers *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Mol Syst Biol*, vol. 7, p. 539, Oct. 2011, doi: 10.1038/msb.2011.75.

[20] B. Kuhlman, “Designing protein structures and complexes with the molecular modeling program Rosetta,” *Journal of Biological Chemistry*, vol. 294, no. 50, pp. 19436–19443, 2019, doi: https://doi.org/10.1074/jbc.AW119.008144.

[21] 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.

[22] G. L. Butterfoss and B. Kuhlman, “Computer-based design of novel protein structures,” *Annu Rev Biophys Biomol Struct*, vol. 35, pp. 49–65, 2006, doi: 10.1146/annurev.biophys.35.040405.102046.

[23] Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman, “PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality,” *BMC Bioinformatics*, vol. 12, no. 1, p. 151, 2011, doi: 10.1186/1471-2105-12-151.

[24] F. Pucci, J. M. Kwasigroch, and M. Rooman, “Protein Thermal Stability Engineering Using HoTMuSiC.,” *Methods Mol Biol*, vol. 2112, pp. 59–73, 2020, doi: 10.1007/978-1-0716-0270-6_5.

[25] F. Pucci, R. Bourgeas, and M. Rooman, “Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC,” *Sci Rep*, vol. 6, no. 1, p. 23257, 2016, doi: 10.1038/srep23257.

[26] R. Vita *et al.*, “The Immune Epitope Database (IEDB): 2018 update.,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D339–D343, Jan. 2019, doi: 10.1093/nar/gky1006.

[27] S. K. Dhanda *et al.*, “Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity,” *Immunology*, vol. 153, no. 1, pp. 118–132, Jan. 2018, doi: 10.1111/imm.12816.

[28] 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.

[29] R. Krishna *et al.*, “Generalized biomolecular modeling and design with RoseTTAFold All-Atom.,” *Science*, vol. 384, no. 6693, p. eadl2528, Apr. 2024, doi: 10.1126/science.adl2528.

[30] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, “Design of a Novel Globular Protein Fold with Atomic-Level Accuracy,” *Science (1979)*, vol. 302, no. 5649, pp. 1364–1368, Nov. 2003, doi: 10.1126/science.1089427.

[31] J. L. Watson *et al.*, “De novo design of protein structure and function with RFdiffusion,” *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023, doi: 10.1038/s41586-023-06415-8.

[32] M. Remmert, A. Biegert, A. Hauser, and J. Söding, “HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment,” *Nat Methods*, vol. 9, no. 2, pp. 173–175, 2012, doi: 10.1038/nmeth.1818.

[33] F. Gabler *et al.*, “Protein Sequence Analysis Using the MPI Bioinformatics Toolkit,” *Curr Protoc Bioinformatics*, vol. 72, no. 1, p. e108, Dec. 2020, doi: https://doi.org/10.1002/cpbi.108.

[34] D. E. Kim, D. Chivian, and D. Baker, “Protein structure prediction and analysis using the Robetta server,” *Nucleic Acids Res*, vol. 32, no. Web Server issue, pp. W526–W531, Jul. 2004, doi: 10.1093/nar/gkh468.

5. Additional side projects during doctoral tenure

5.1. Modeling anti-CTLA4 nanobodies for immune checkpoint blockade

This has been adapted with permission from 'Nanobody-based CTLA4 inhibitors for immune checkpoint blockade therapy of canine cancer patients', Jonathan Marable, Damien Ruiz, Anil K. Jaiswal, Ritankar Bhattacharya, Robert Pantazes, Payal Agarwal, Amol S. Suryawanshi, Deepa Bedi, Amarjit Mishra, Bruce F. Smith, Maninder Sandey, Scientific Reports, Nature, 21 October 2021[1].

5.1.1. Background information

Although antibodies (Abs) extremely important therapeutic binders, they have one significant limitation when treating tumors. Due to their large size (~150 kDa), they are unable to penetrate tumors[2], [3]. It is a significant issue that such a highly effective protein binder is unable to utilize its full potential when it comes to cancer tissues/tumors. This is where smaller molecules are needed. Both Ig-like and non-Ig binders have been engineered to tackle such issues. One such particular example of an Ig-like protein binder, found in camels, sharks, and some other animals is called a VHH or a single domain antibody (sdAb) or a nanobody (Nb)[4], [5], [6]. These scaffolds are the single heavy chain variable Ab domains, which have CDRs homologous to Abs. They are almost $1/10^{th}$ the size of a conventional Ab i.e., approx. 12 kDa. Research has found that camelid (camels) Nb or VHH domains are very robust and have similar specificities as Abs. They are highly soluble and can be cultured in high concentration formulations. Since they lack the Fc region, they do not require expression in eukaryotic cells, so their bulk production becomes a lot cheaper compared to producing a conventional Ab[7]. These factors make them a high-value therapeutic binder, particularly for cancer research. In this project, experimentally tested nanobodies were modeled to bind a particular epitope of a canine CTLA4 molecule.

CTLA4[8], [9] is an important immunoregulatory protein, which downregulates T-cell activation and prevents the immune system from reacting to an antigen. This protein, when

expressed on T cells prevents B7-1 or B7-2 from binding to CD28 and blocks positive co-stimulatory action[10]. This blockade of complex formation of CD28 – B7-1 by CTLA4 initiates a cascade of events, TCR downregulation, reduction in T cell activation, and induction of T cell tolerance[11]. Therefore, a high affinity, effective, and site-specific protein binding CTLA4 and blocking its mode of action will increase the survival time and cause remission of cancer cells. Since dogs are also exposed to the same environment as us humans, they are also prone to developing cancer over their lifetime. Moreover, the human and dog CTLA4 proteins have conserved regions in their sequences. However, commercially available therapeutic blockers are unavailable for canine species. These reasons served as the motivation behind developing blockers to target canine CTLA4 protein.

This was a collaboration project where the authors had generated an Nb library and identified the ones to bind canine CTLA4. The Nbs were already characterized *in vitro* and the exact binding pocket had to be modeled. This was a classic case of protein-protein docking i.e., to be more specific, a case of global (unbiased) docking[12], [13]. Since the epitope of protein CTLA4 was unknown, the global docking method was used so that no residues can be specified to where the Nbs would bind. Below is a brief summary of this project. It was published in the Scientific Reports, in 2021[1].

5.1.2. Methods

There were three Nbs (cNb6, cNb13, and cNb17), and the canine CTLA4 genetic sequences that were provided by the collaborators. The CDRs were determined from their sequences using the multiple sequence alignment tool Clustal Omega, EMBL-EBI[14], [15]. The structures of the Nbs, the extra-cellular CTLA4 domain, and the entire canine CTLA4 protein were predicted from their genetic sequences using Robetta[16], trRosetta[17], and I-Tasser[18], [19], [20]. However, that was before the advent of machine learning or AlphaFold[21], I-Tasser was considered to be the gold standard of protein structure predictions. However, since these were all predictions, multiple tools were used to conclude if the predictions were similar or different due to the approaches. Once, the structures were predicted, ZDOCK[22], [23] was used to globally dock the predicted individual Nbs with the extra-cellular domain of the canine CTLA4 protein. 9 jobs were submitted, 1 for each nanobody predicted using a particular method with the corresponding CTLA4 protein. Each job had 10 results, and there were 90 predicted docked complexes. These complexes were visually inspected using the well-known molecular visualization tool UCSF Chimera[24].

5.1.3. Results

A total of 15 structures were obtained from the prediction tools, which were 5 structures for each tool. The structural predictions of the entire canine CTLA4 protein were consistently of poor quality for all the methods i.e., Robetta, trRosetta, and I-Tasser. However, they were all high-confidence predictions for the extra-cellular domain by itself. Since the Nbs interacted with the extra-cellular domain, the same was used for the docking calculations. Additionally, each Nb structure prediction was also generated with high confidence scores. The predictions are in Figure 1. A total of 90 docked complexes were generated out of which 85 showed the Nb CDRs

interacting with the canine CTLA4 MYPPPY epitope. This epitope is conserved in humans[9], mice[8], and dogs as shown in Figure 2. The remaining structures exhibited unrealistic binding where the CDRs were not at all interacting with the antigen. Figure 3 shows the Nb-CTLA4 docked complex and a zoomed-in view to depict the binding interactions.

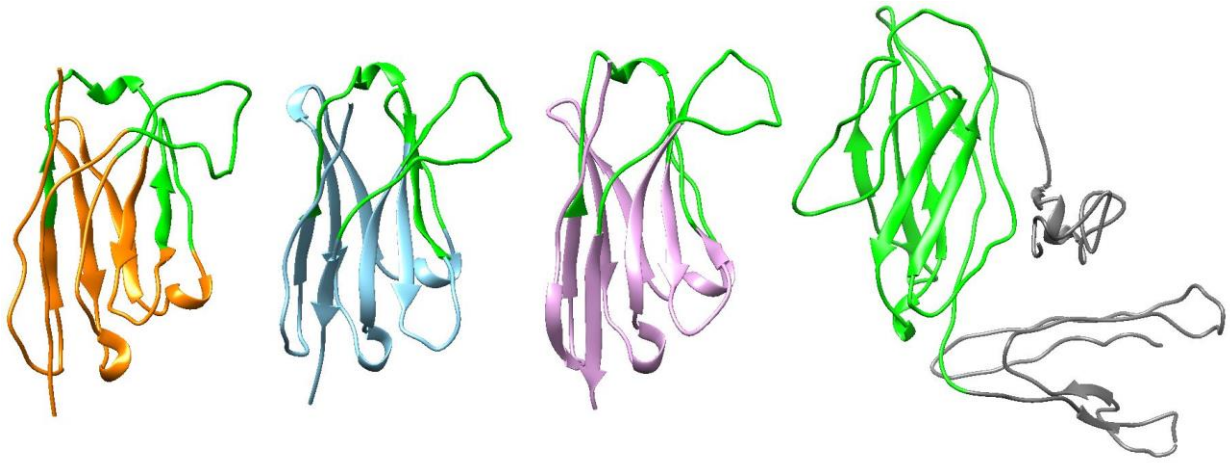

Figure 1. **I-Tasser predicted nanobodies and canine CTLA4 protein**. The green color highlights the binding region of both the Nbs and the target protein.

```
Seq_Murine      ------------------------------------IQVTQPSVVLASSHGVASFPCEY 23
Seq_Homo        ------------------------------------AMHVAQPAVVLASSRGIASFVCEY 24
Seq_Can         MAGFGFRRHGAQPDLASRTWPCTALFSLLFIPVFSKGMHVAQPAVVLASSRGVASFVCEY 60
                                                    ::*:**:******:*:*** ***

Seq_Murine      SPSHNTDEVRVTVLRQTNDQMTEVCATTFTEKNTVGFLDYPFCSGTFNESRVNLTIQGLR 83
Seq_Homo        ASPGKATEVRVTVLRQADSQVTEVCAATYMMGNELTFLDDSICTGTSSGNQVNLTIQGLR 84
Seq_Can         GSSGNAAEVRVTVLRQAGSQMTEVCAATYTVEDELAFLDDSTCTGTSSGNKVNLTIQGLR 120
                .   :: *********:..*:*****:*:   : : ***   *:** . .:*********

Seq_Murine      AVDTGLYLCKVELMYPPPYFVGMGNGTQIYVIDP------------------------- 117
Seq_Homo        AMDTGLYICKVELMYPPPYYLGIGNGTQIYVIDP------------------------- 118
Seq_Can         AMDTGLYICKVELMYPPPYYVGMGNGTQIYVIDPEPCPDSDFLLWILAAVSSGLFFYSFL 180
                *:*****:**********::*:**********

Seq_Murine      ------------------------------------------- 117
Seq_Homo        ------------------------------------------- 118
Seq_Can         ITAVSLSKMLKKRSPLTTGVYVKMPPTEPECEKQFQPYFIPIN 223
```

Figure 2. **Sequence alignment of Human, murine, canine CTLA4**. The alignment of three genetic sequences of murine, human and canine species. Except for the head and tail position of the canine domain, majority of the remaining sequence is similar to the other two genetic sequences. Additionally, the epitope MYPPPY is conserved in all three sequences of the different species and the anti-CTLA4 Nbs bound to this position.

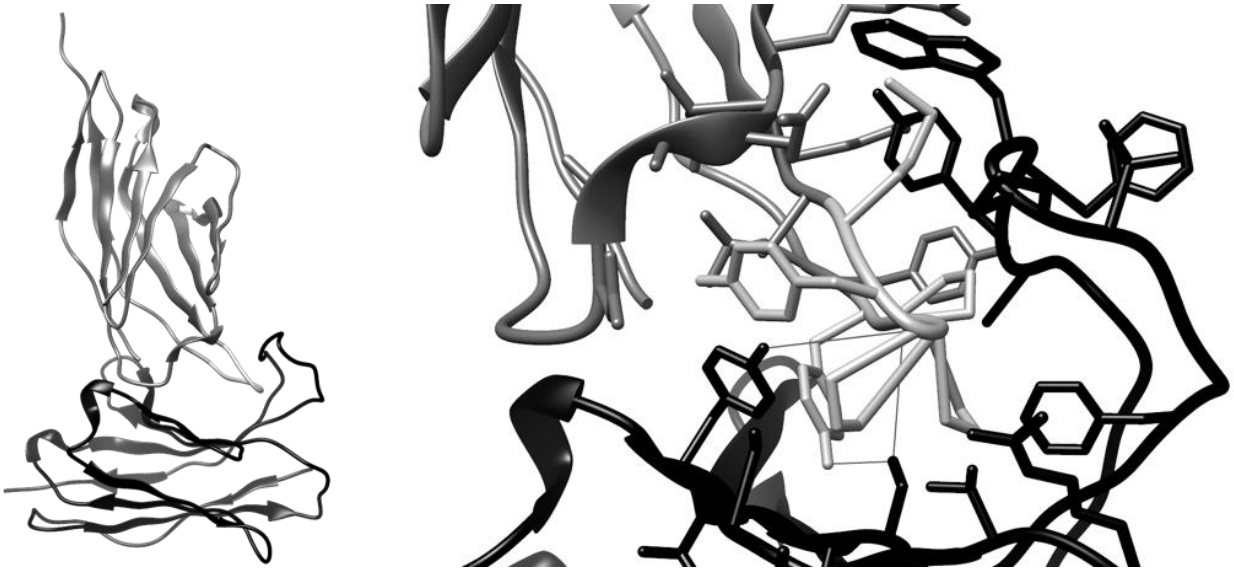

Figure 3. **Nanobodies docked to MYPPPY sequence of canine CTLA4**. The image of the left side shows the globally docked complex of anti-CTLA4 cNb6 (black) with canine CTLA4 (light gray) with the conserved epitope MYPPPY (dark gray) interacting with the cNb6 CDRs. The image on the right portrays a more detailed view of the previous image, highlighting the different binding interactions between the two molecules. The fine straight lines depict hydrogen bond formation between Ser-Pro/Ser-Tyr/Tyr-Pro residues near the center of the image.

5.1.4. Discussion

In summary, we were able to identify the anti-CTLA4 cNb CDRs, model them and their stable framework, and further predict their binding to the conserved MYPPPY epitope of the canine CTLA4 protein. This was consistent with the experimental evidence found in the laboratory.

5.2. Using in-silico designed Nanobodies to bind Interleukin-2

5.2.1. Background information

Another collaboration project also involved nanobodies (Nbs). However, this time, Nbs had to be designed to bind interleukin-2/IL-2 and then experiments were to be conducted. It was in the reverse order of the previous project. IL-2 is a protein that belongs to the cytokine family and plays a crucial role in regulating the immune system[25], [26], [27]. It has an important role to play in T cell regulation, differentiation, proliferation, assists in activating NK cells to attack and destroy cancer cells, suppression of specific immune responses which prevents autoimmune diseases, and many more[25], [28], [29]. As a result of this plethora of immune regulating activities, it is also used in cancer treatment in several ways such as target therapies, chemotherapeutic agents, peptide therapies, and many other techniques[26], [30], [31]. A Nb can potentially bind to an IL-2 to increase its half-life and stability, which will cause an increase in its efficiency to regulate the T cell activities, thus enhancing immune-system activation. Hence, using Nbs to bind IL-2 can be studied to check the effect of the complex on various immune system activities, which will open a direction towards several biomedical research.

As described in Chapter 2 of this dissertation, PETEI was a tool developed by the PROTEIN PANT(z) Lab and was used to design the Nbs to bind IL-2. Chapter 2 has a detailed explanation of the entire PETEI workflow, which will not be repeated here again.

5.2.2. Methods

The PDB 4B50 a 1.30 Å XRD structure, with no disulfide bonds and no residue gaps, was selected as the protein scaffold. The 3 CDRs were identified and used as the input for the protein database creation. This was followed by ligand selection, which was PDB 4ERJ, an IL-2 in complex with its receptor. Here, the IL-2 chain had to be extracted and used as the target antigen.

After the database creation process, 3 distinct epitopes were selected in the target IL-2 protein, and the design simulation run was executed. After the designs were generated for the 3 distinct epitopes, the designs that were screened from being discarded based on charge-charge clashes, steric hindrance, and the presence of more than 3 positive residues on the CDRs were further energy minimized using CHARMM[32], [33] and Rosetta[34] forcefields. This was further followed by visual inspection using UCSF Chimera[24] to look for good structural features and the top designs were selected from each set for experimental testing.

5.2.3. Results

PDB 4B50 CDRs were used for the database creation step. The point of initial and final attachment of each CDRs was 26-37 for H1, 57-64 for H2, and 109-114 for H3, where H1-3 represents the heavy chain CDRs. This generated a database of 669, 731, and 748 usable H1, H2, and H3 CDR loops respectively, creating a possibility of almost 366 million combinations of Nbs. There were 3 distinct epitopes selected for the design run, each with its separate set of residues based on their contact with the IL-2 receptor molecule. The epitopes were named A, B, and C. The PETEI designs were minimized using the fixed backbone CHARMM energy minimization followed by the all-atom Rosetta energy minimization, and the designs with more than 3 positive CDR residues were discarded. Finally, for Epitope A, there were 55 designs, for Epitope B, there were 121 designs, and for Epitope C, there were 51 designs. The designs were further analyzed using the Rosetta Interface Analyzer (RIA) for their binding metrics calculation. This was further subjected to visual inspection using UCSF Chimera and the best structures were screened for experiments. The structures were carefully observed to find good binding features, such as polar interactions, burial of the core hydrophobic residues, absence of too many labile residues in the CDRs, and the absence of peripheral voids around the protein core. Figure 4 shows the binding

positions of the Nb with the different IL-2 epitope binding positions and Figure 5 shows the binding metric characteristics of the designs. The green straight fine lines in Figure 4 represent hydrogen bond formation, which suggests strong interactions. They were compared with a non-redundant Ab-protein database. The reason is that therapeutic antibodies are arguably the most popular binding proteins and their experimentally determined crystallized structures outnumber the crystallized structures of other proteins by a huge margin. Moreover, nanobodies are engineered single domain Abs so they have the homology with Ab CDRs. Therefore, it made sense both structurally and statistically to consider a database of non-redundant Ab-protein to compare the *in-silico* designs. Figure 5 shows the binding energy/buried surface area of the database and the Nbs binding to three distinct epitopes. It can be observed that the majority of the designed structures are around the database average values. Additionally, the top 10 designs of each epitope were selected for experimental validation.

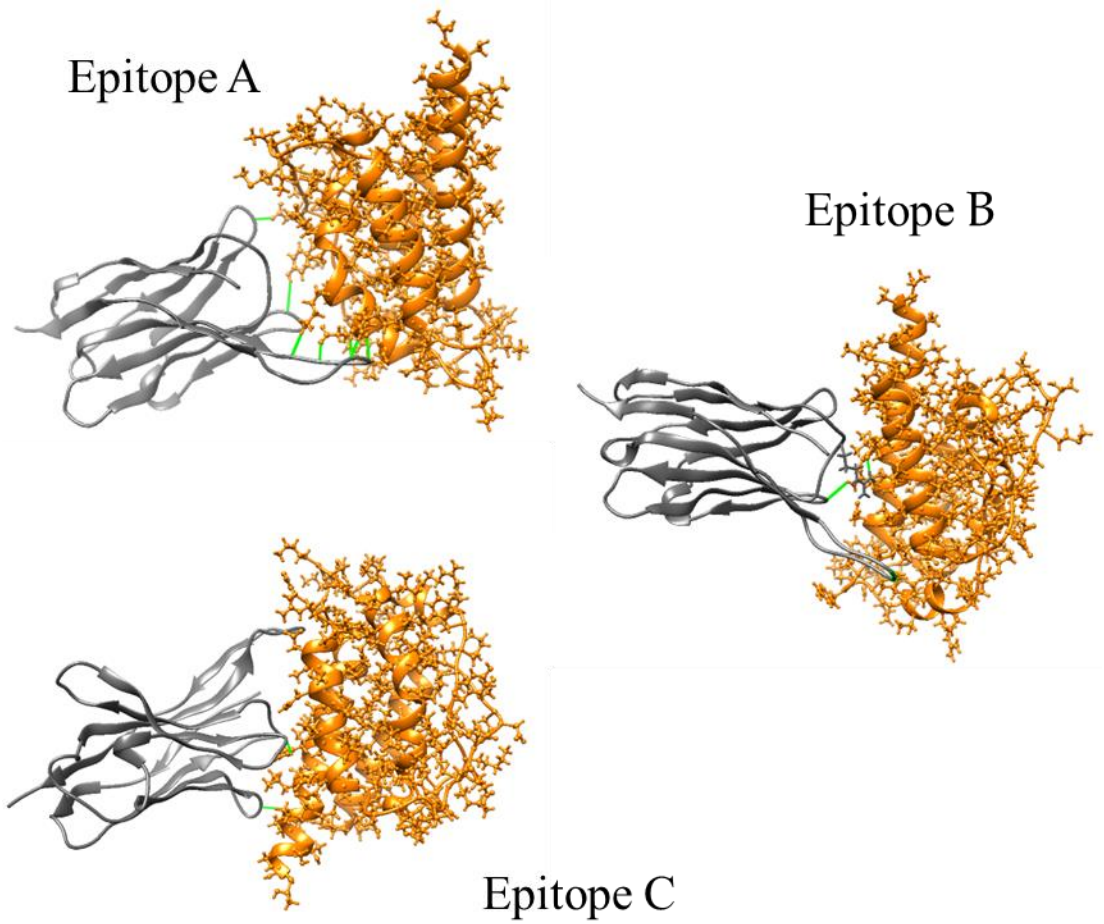

Figure **4. Nanobodies binding to 3 distinct epitopes of IL-2**. Nb PDB 4B50 (gray) in complex with three distinct epitopes of IL-2 (PDB 2ERJ, orange). This image was observed in UCSF Chimera. The fine green straight lines depict hydrogen bonds as a sign of strong optimal binding interactions between the paratope and epitope.

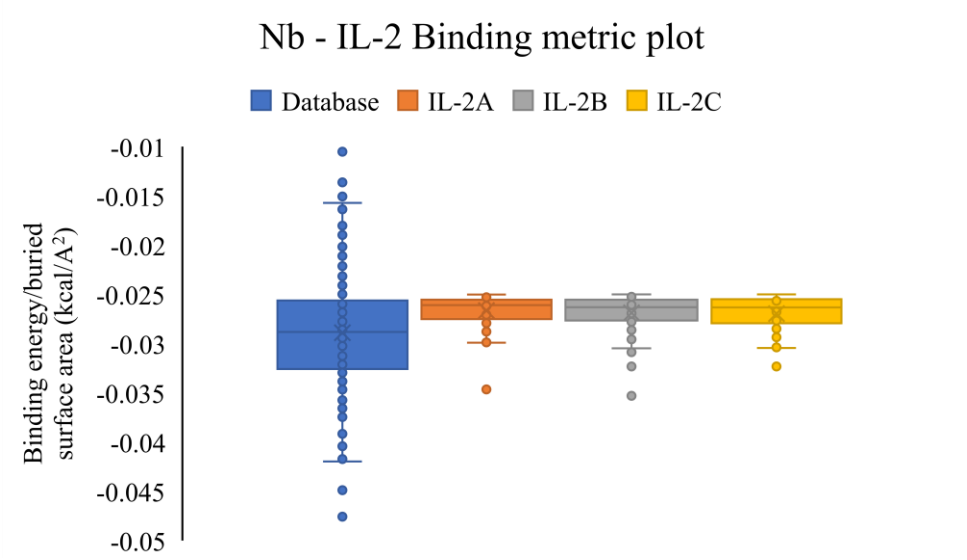

Figure 5. **Rosetta minimized binding metrics plot of Nb – IL-2 complexes**. The plot depicts the Nb binding metrics compared to the non-redundant Ab-protein database. It can be clearly observed that the designs containing all three epitopes have similar binding metric values, whose average is slightly higher than -0.025 kCal/mol. $Å^2$ compared to the database average which is around -0.028 kCal/mol. $Å^2$.

5.2.4. Discussion

PETEI was successfully able to design Nbs to target and bind three distinct epitopes of IL-2. It made a database of almost 366 million possible combinations of Nbs, a one-time process to bind the target antigen. This same database can be used in the future in case research with Nbs is being conducted. The designed proteins were minimized using CHARMM and Rosetta and their binding metrics were calculated using RIA. Then they were further visually analyzed using UCSF Chimera to find designs with good structural features to screen them out. Finally, 10 best designs were selected for each epitope i.e., a total of 30 designs for wet laboratory experiments.

References

[1] J. Marable *et al.*, “Nanobody-based CTLA4 inhibitors for immune checkpoint blockade therapy of canine cancer patients,” *Sci Rep*, vol. 11, no. 1, p. 20763, 2021, doi: 10.1038/s41598-021-00325-3.

[2] D. R. Burton, “Antibody: the flexible adaptor molecule,” *Trends in Biochemical Sciences*. 1990. doi: 10.1016/0968-0004(90)90178-E.

[3] et al. Janeway CA Jr, Travers P, Walport M, “The structure of a typical antibody molecule.,” *Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001*.

[4] P. Bannas, J. Hambach, and F. Koch-Nolte, “Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics,” *Front Immunol*, vol. 8, no. NOV, pp. 1–13, 2017, doi: 10.3389/fimmu.2017.01603.

[5] S. Muyldermans, “Nanobodies: Natural Single-Domain Antibodies,” *Annu Rev Biochem*, vol. 82, no. 1, pp. 775–797, 2013, doi: 10.1146/annurev-biochem-063011-092449.

[6] C. Vincke, R. Loris, D. Saerens, S. Martinez-Rodriguez, S. Muyldermans, and K. Conrath, “General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold,” *Journal of Biological Chemistry*, vol. 284, no. 5, pp. 3273–3284, 2009, doi: 10.1074/jbc.M806889200.

[7] A. Ghannam, S. Kumari, S. Muyldermans, and A. Q. Abbady, “Camelid nanobodies with high affinity for broad bean mottle virus: a possible promising tool to immunomodulate plant resistance against viruses,” *Plant Mol Biol*, vol. 87, no. 4, pp. 355–369, 2015, doi: 10.1007/s11103-015-0282-5.

[8] D. A. Ostrov, W. Shi, J.-C. D. Schwartz, S. C. Almo, and S. G. Nathenson, “Structure of Murine CTLA-4 and Its Role in Modulating T Cell Responsiveness,” *Science (1979)*, vol. 290, no. 5492, pp. 816 LP – 819, Oct. 2000, doi: 10.1126/science.290.5492.816.

[9] W. J. Metzler *et al.*, “Solution structure of human CTLA-4 and delineation of a CD80/CD86 binding site conserved in CD28,” *Nat Struct Biol*, vol. 4, no. 7, pp. 527–531, 1997, doi: 10.1038/nsb0797-527.

[10] L. S. K. Walker and D. M. Sansom, “The emerging role of CTLA4 as a cell-extrinsic regulator of T cell responses.,” *Nat Rev Immunol*, vol. 11, no. 12, pp. 852–863, Nov. 2011, doi: 10.1038/nri3108.

[11] S. C. Wei *et al.*, “Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade.,” *Cell*, vol. 170, no. 6, pp. 1120-1133.e17, Sep. 2017, doi: 10.1016/j.cell.2017.07.024.

[12] N. S. Pagadala, K. Syed, and J. Tuszynski, “Software for molecular docking: a review,” *Biophys Rev*, vol. 9, no. 2, pp. 91–102, 2017, doi: 10.1007/s12551-016-0247-1.

[13] I. A. Vakser, “Protein-protein docking: From interaction to interactome,” *Biophys J*, vol. 107, no. 8, pp. 1785–1793, 2014, doi: 10.1016/j.bpj.2014.08.033.

[14] F. Sievers *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Mol Syst Biol*, vol. 7, p. 539, Oct. 2011, doi: 10.1038/msb.2011.75.

[15] F. Madeira *et al.*, “The EMBL-EBI search and sequence analysis tools APIs in 2019,” *Nucleic Acids Res*, vol. 47, no. W1, pp. W636–W641, 2019, doi: 10.1093/nar/gkz268.

[16] D. E. Kim, D. Chivian, and D. Baker, “Protein structure prediction and analysis using the Robetta server,” *Nucleic Acids Res*, vol. 32, no. Web Server issue, pp. W526–W531, Jul. 2004, doi: 10.1093/nar/gkh468.

[17] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, “Improved protein structure prediction using predicted interresidue orientations,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 3, pp. 1496 LP – 1503, Jan. 2020, doi: 10.1073/pnas.1914677117.

[18] A. Roy, A. Kucukural, and Y. Zhang, “I-TASSER: A unified platform for automated protein structure and function prediction,” *Nat Protoc*, vol. 5, no. 4, pp. 725–738, 2010, doi: 10.1038/nprot.2010.5.

[19] Y. Zhang, “I-TASSER server for protein 3D structure prediction,” *BMC Bioinformatics*, vol. 9, pp. 1–8, 2008, doi: 10.1186/1471-2105-9-40.

[20] J. Yang and Y. Zhang, “Protein Structure and Function Prediction Using I-TASSER,” *Curr Protoc Bioinformatics*, vol. 52, pp. 5.8.1-5.8.15, Dec. 2015, doi: 10.1002/0471250953.bi0508s52.

[21] 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.

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

[23] 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.

[24] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[25] W. Liao, J. Lin, and W. J. Leonard, “IL-2 Family Cytokines,” *Curr Opin Immunol*, vol. 23, no. 5, pp. 418–423, 2011, doi: 10.1016/j.coi.2011.08.003.IL-2.

[26] T. Jiang, C. Zhou, and S. Ren, “Role of IL-2 in cancer immunotherapy,” *Oncoimmunology*, vol. 5, no. 6, pp. 1–10, 2016, doi: 10.1080/2162402X.2016.1163462.

[27] D. J. Stauber, E. W. Debler, P. A. Horton, K. A. Smith, and I. A. Wilson, “Crystal structure of the IL-2 signaling complex: Paradigm for a heterotrimeric cytokine receptor,” *Proc Natl Acad Sci U S A*, vol. 103, no. 8, pp. 2788–2793, 2006, doi: 10.1073/pnas.0511161103.

[28] S. L. Gaffen and K. D. Liu, “Overview of interleukin-2 function, production and clinical applications,” *Cytokine*, vol. 28, no. 3, pp. 109–123, 2004, doi: 10.1016/j.cyto.2004.06.010.

[29] O. Boyman and J. Sprent, “The role of interleukin-2 during homeostasis and activation of the immune system,” *Nat Rev Immunol*, vol. 12, no. 3, pp. 180–190, 2012, doi: 10.1038/nri3156.

[30] A. Valle-Mendiola, A. Gutiérrez-Hoya, M. D. C. Lagunas-Cruz, B. Weiss-Steider, and I. Soto-Cruz, “Pleiotropic Effects of IL-2 on Cancer: Its Role in Cervical Cancer,” *Mediators Inflamm*, vol. 2016, 2016, doi: 10.1155/2016/2849523.

[31] G. Ju, K. M. Heaton, and E. A. Grimm, “Human Interleukin 2 Analogues That Preferentially Bind the Intermediate-Affinity Interleukin 2 Receptor Lead to Reduced Secondary Cytokine Secretion: Implications for the Use of These Interleukin 2 Analogues in Cancer Immunotherapy,” *Cancer Res*, vol. 53, no. 11, pp. 2597–2602, 1993.

[32] A. Manuscript, “CHARMM.pdf,” vol. 31, no. 4, pp. 671–690, 2011, doi: 10.1002/jcc.21367.CHARMM.

[33] J. Huang *et al.*, “CHARMM36m: an improved force field for folded and intrinsically disordered proteins.,” *Nat Methods*, vol. 14, no. 1, pp. 71–73, Jan. 2017, doi: 10.1038/nmeth.4067.

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

6. Future Works

6.1. Experimental testing for the DARPin designs

In this project, a novel algorithm called MutDock[1], developed by the PROTEIN PANT(z) Group was used to design DARPin[2], [3] and α-Cobratoxin[4], [5] complexes. As discussed in Chapter 3, 39 out of 2145 CHARMM and Rosetta forcefield energy minimized MutDock designs were analyzed and shortlisted for further research using the binding metrics threshold of experimentally determined Ab-proteins. Rosetta affinity maturation generated 3900 designs, out of which the top 73 were shortlisted based on similar protocols and visual inspection in UCSF Chimera[6]. However, this time the threshold of binding metrics increased, which suggests the improved binding affinity of the designs. 73 designs were further analyzed using VMD[7] and NAMD[8] MD simulations out of which 51 designs were finally shortlisted for laboratory experiments.

Before the binding assay experiments are conducted, the sequence should be ordered from a vendor. DARPins should be expressed in bacterial cells like a compatible strain of *E. coli*. Since DARPins do not have Fc sites, they do not require mammalian cells. Then the system is purged of unwanted proteins, which can tamper with the experimental results. There protein purification is an important step in the experimental validation[9], [10], [11]. The steps for these processes are protein-system specific, and a plethora of protein purification handbooks are available to carefully review the details of the expression system and use the necessary reagents to remove protein impurities. Once the DARPins are expressed and purified, they are ready for binding assay experiments such as competitive Enzyme Linked Immuno-Sorbent Assay, commonly known as

ELISA. These techniques can be used to measure the dissociation constant or K_d of the binder and the lower its value, the better the binding, which means the better the *in-silico* design quality. K_d has a direct relationship with dG of binding and the interaction can also be calculated for the same and compared to the forcefield generated energy. The steps of experiments mentioned previously are very generalized because the experiments will be conducted by a research faculty collaborator at the University of Florida, Florida. Therefore, we will be sharing the sequences of the 51 designed complexes and they will express, purify, characterize the DARPins, and conduct binding assay experiments.

This project was undertaken to design a therapeutic binder to treat venomous snakebites. After a plethora of experiments conducted for this research, the results established that the 51 designs shortlisted were the most likely to work in wet laboratory experiments. If they turn out to be an experimental success then that will suggest that rational protein design using computational tools, good structural features, and an appropriate pipeline of complex scientific protocols to shortlist the most promising designs can help develop therapeutic binders in a faster and cost-effective way. The results also prove the benefits of exploiting structural features to select protein epitopes. Such decisions, if made correctly can result in good-quality designs that will have a higher chance of binding experimentally.

After the advent of machine learning in the field of protein engineering, scientists have taken great interest in it; however, the foundation of this new technology was established by rational protein design and therefore, it has the same importance today as before the advent of machine learning. These concepts altogether bring humanity a step closer to designing on-demand therapeutics, that will in the coming future pave the way for a faster, cost-effective, and reliable technology to revolutionize the field of medical science. These shortlisted DARPin designs are the

proof that after being compared with the binding metrics of experimentally validated structures from the PDB, they have consistently exhibited comparable to better results. Furthermore, the MDs results helped reinforce this judgment since they were analyzed w.r.t time. During this analysis, structures with decent binding metrics proved to have detrimental structural features. Therefore, such decisions are extremely significant for shortlisting the most promising structures for experiments. Such protein design algorithms when coupled with scientific rationale, experimental data, and sophisticated data analysis and down-selection process, can give rise to the development of a powerful technique to establish a fast and cost-effective way to come up with on-demand protein binders.

6.2. In-silico protein design using the 'Ideal' binding proteins

As mentioned in Chapter 4, PDB 3HAS was re-designed from scratch using RosettaDesign[12] to improve its physicochemical properties. This was followed by the CamSol combination method[13], which significantly improved the solubility with structurally stabilizing mutations. Furthermore, HoTMuSiC[14] and PoPMuSiC[15] were used to improve protein thermostability. The sequences were also de-immunized using the IEDB[16] De-immunization[17] tool to prevent unnecessary *in vivo* immune reactions. Lastly, one solvent exposed non-polar surface accessible residue was mutated to Cys to serve as a single drug conjugation[18] site for controlled dosage. This was a very brief account of the experiments conducted to arrive at the re-designed 16 scaffolds. Unlike the previous chapters, which utilized a rank-ordering system to shortlist designs for experiments, or a linear progression of protocols to help select the best designs, this project used a more complex iterative technique to arrive at 16 completely re-designed protein sequences to mimic therapeutic Abs. This list of scientific protocols serves as the commonality in all the chapters and helps connect the chapters. In between Chapters 2 and 4, this process got progressively more complex whereas in Chapter 2, the design phase and evaluation phase were completely separate, whereas in Chapter 3, the design and evaluation phase although separate were repeated in a series of subsequent methods. Finally, in Chapter 4, the two phases were no longer related, and the characteristic properties of individual designs were assessed and compared with the others to shortlist them. Therefore, several parameters had to be considered before designing and shortlisting them in this chapter.

Since there are 16 different humanized scaffold sequences, the PETEI algorithm will make protein designs. The database generation step will generate binding loops to be swapped with wild type loops of the helical bundles in residue positions 34-36, 62-80, 101-104, 128-130, 161-164, and 192-200. This will be followed by selecting PDB 7U9G[19] as the antigen, a prefusion state of rabies glycoprotein in complex with an Ab. Epitope selection will be based on the presence of pre-stabilized residues and 4 different epitopes will be selected. Epitope_1 residues are 150, 151, 155, 166, 167, 170, 172, epitope_2 residues are 38, 197, 202, 203, epitope_3 residues are 44, 46, 179, 180, 218, 229, epitope_4 residues are 238, 239, 240, 241, 242, and 245. UniProt[20] shows that the glycoprotein sequence (final sequence 402) is far away from the membrane protein (starting sequence 460) of rabies. Therefore, the epitopes on either end of the molecule will not provide steric hindrance during the design runs. Since the end loops are short, even MutDock can be explored to generate the designs. However, the paratope residues will remain the same in this case.

Once the designs are generated, they will be energy minimized using CHARMM and Rosetta forcefields and their binding metrics will be calculated using the Rosetta Interface Analyzer. Based on the energy minimized non-redundant database of 231 Ab-protein complexes, the BE/BSA threshold is -0.028 ± 0.006 kcal/mol.Å^2. This will also serve as a reference to shortlist the analyzed structures from the thousands generated. The shortlisted designs will be subjected to Rosetta affinity maturation to improve binding affinity. This will again follow the energy minimization and data analysis protocol, followed by visual inspection in UCSF Chimera for good complex formation features such as H-bonds, salt bridges, and complete burial of the interfacial residues. Based on the quality and number of designs generated, the top 1-5% will be tested experimentally. The proteins will be expressed in *E. coli* bacterial cells and purified using the standardized

protocols. Different experiments will be conducted to analyze the protein solubility, stability and binding, once it has been expressed and purified.

Additionally, if these designs are successfully expressed, purified, and characterized then their next step is binding to the target antigen. Therefore, it will be safe to say that an antibody mimetic has been designed from scratch using computational techniques. Just as explained in the previous chapter, this will also pave the way for designing on-demand binders using stabilizing point mutations as compared to working through an algorithm. Either way, it benefits humanity by providing the technology for a faster, cheaper, and more reliable protein design method.

References

[1] V. M. Chauhan and R. J. Pantazes, “MutDock: A computational docking approach for fixed-backbone protein scaffold design,” *Front Mol Biosci*, vol. 9, no. August, pp. 1–15, 2022, doi: 10.3389/fmolb.2022.933400.

[2] 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.

[3] M. T. Stumpp, H. K. Binz, and P. Amstutz, “DARPins: A new generation of protein therapeutics,” *Drug Discov Today*, vol. 13, no. 15–16, pp. 695–701, 2008, doi: 10.1016/j.drudis.2008.04.013.

[4] H. F. Williams *et al.*, “The urgent need to develop novel strategies for the diagnosis and treatment of snakebites,” *Toxins (Basel)*, vol. 11, no. 6, pp. 1–29, 2019, doi: 10.3390/toxins11060363.

[5] T. Lynagh *et al.*, “Peptide Inhibitors of the α-Cobratoxin–Nicotinic Acetylcholine Receptor Interaction,” *J Med Chem*, vol. 63, no. 22, pp. 13709–13718, Nov. 2020, doi: 10.1021/acs.jmedchem.0c01202.

[6] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[7] W. Humphrey, A. Dalke, and K. Schulten, “VMD: visual molecular dynamics.,” *J Mol Graph*, vol. 14, no. 1, pp. 27-28,33-38, Feb. 1996, doi: 10.1016/0263-7855(96)00018-5.

[8] J. C. Phillips *et al.*, “Scalable molecular dynamics with NAMD,” *J Comput Chem*, vol. 26, no. 16, pp. 1781–1802, Dec. 2005, doi: 10.1002/jcc.20289.

[9] N. K. Tripathi, “ Production and Purification of Recombinant Proteins from Escherichia coli ,” *ChemBioEng Reviews*, vol. 3, no. 3, pp. 116–133, 2016, doi: 10.1002/cben.201600002.

[10] D. Ruano-Gallego, S. Fraile, C. Gutierrez, and L. Á. Fernández, “Screening and purification of nanobodies from E. coli culture supernatants using the hemolysin secretion system,” *Microb Cell Fact*, vol. 18, no. 1, pp. 1–13, 2019, doi: 10.1186/s12934-019-1094-0.

[11] Y. Li, “A brief introduction of IgG-like bispecific antibody purification: Methods for removing product-related impurities,” *Protein Expr Purif*, vol. 155, no. November 2018, pp. 112–119, 2019, doi: 10.1016/j.pep.2018.11.011.

[12] B. Kuhlman, “Designing protein structures and complexes with the molecular modeling program Rosetta,” *Journal of Biological Chemistry*, vol. 294, no. 50, pp. 19436–19443, 2019, doi: https://doi.org/10.1074/jbc.AW119.008144.

[13] A. Rosace *et al.*, “Automated optimisation of solubility and conformational stability of antibodies and proteins,” *Nat Commun*, vol. 14, no. 1, p. 1937, 2023, doi: 10.1038/s41467-023-37668-6.

[14] F. Pucci, J. M. Kwasigroch, and M. Rooman, “Protein Thermal Stability Engineering Using HoTMuSiC.,” *Methods Mol Biol*, vol. 2112, pp. 59–73, 2020, doi: 10.1007/978-1-0716-0270-6_5.

[15] Y. Dehouck, J. M. Kwasigroch, D. Gilis, and M. Rooman, “PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality,” *BMC Bioinformatics*, vol. 12, no. 1, p. 151, 2011, doi: 10.1186/1471-2105-12-151.

[16] R. Vita *et al.*, “The Immune Epitope Database (IEDB): 2018 update.,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D339–D343, Jan. 2019, doi: 10.1093/nar/gky1006.
[17] S. K. Dhanda *et al.*, “Development of a strategy and computational application to select candidate protein analogues with reduced HLA binding and immunogenicity,” *Immunology*, vol. 153, no. 1, pp. 118–132, Jan. 2018, doi: 10.1111/imm.12816.
[18] D. Shinmi *et al.*, “One-Step Conjugation Method for Site-Specific Antibody-Drug Conjugates through Reactive Cysteine-Engineered Antibodies,” *Bioconjug Chem*, vol. 27, no. 5, pp. 1324–1331, May 2016, doi: 10.1021/acs.bioconjchem.6b00133.
[19] H. M. Callaway *et al.*, “Structure of the rabies virus glycoprotein trimer bound to a prefusion-specific neutralizing antibody,” *Sci Adv*, vol. 8, no. 24, p. eabp9151, Jul. 2024, doi: 10.1126/sciadv.abp9151.
[20] A. Bateman, “UniProt: A worldwide hub of protein knowledge,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D506--D515, 2019, doi: 10.1093/nar/gky1049.

Appendix 1

Chapter 3: Additional Information

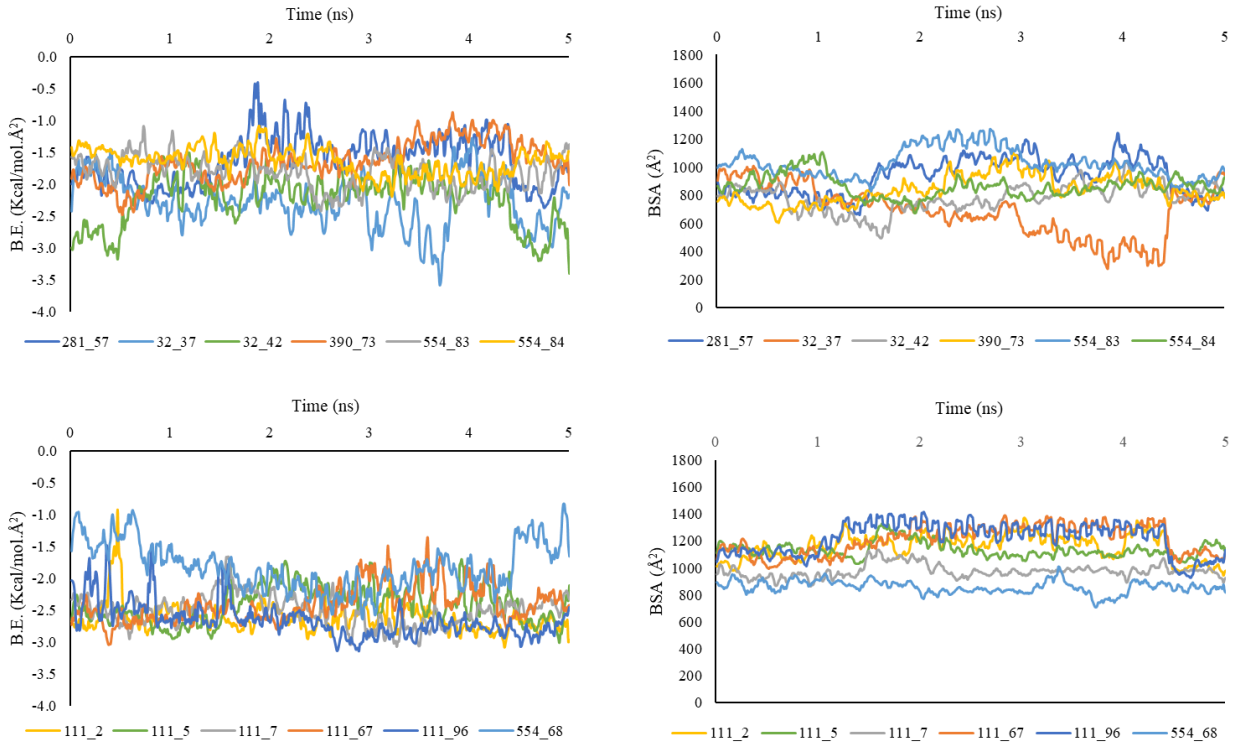

Figure 1. The figure shows the MDs characteristics of 4 distinct plots, 2 with B.E/BSA vs. time on the left and 2 with BSA vs. time on the right. A detailed explanation has been provided in Chapter 2 to understand these plots.

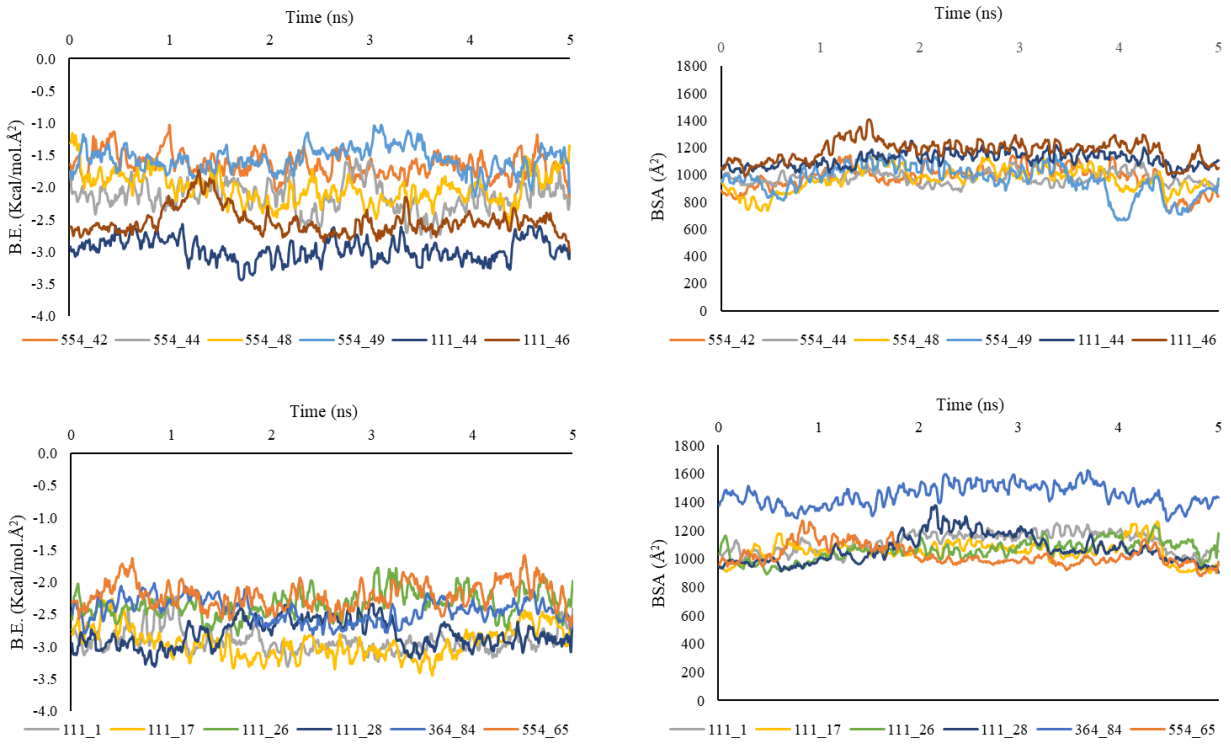

Figure 2. A detailed explanation has been provided in Chapter 2 to understand these plots.

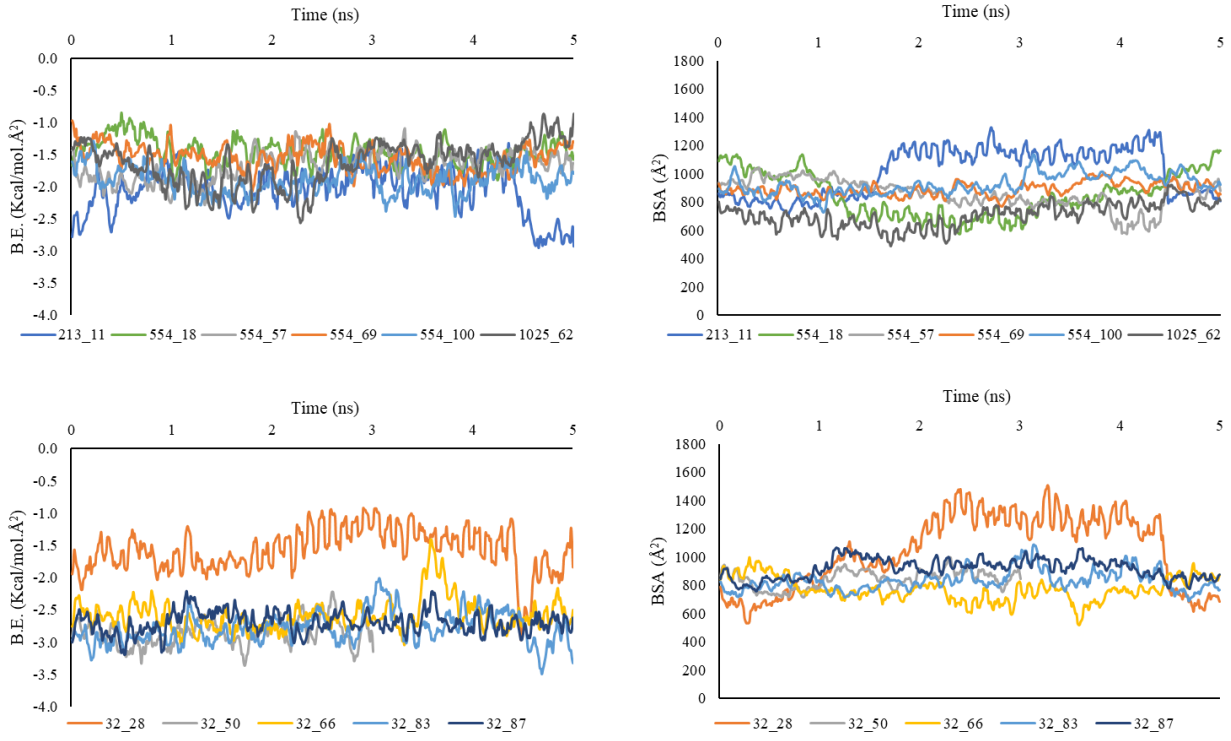

Figure 3. A detailed explanation has been provided in Chapter 2 to understand these plots.

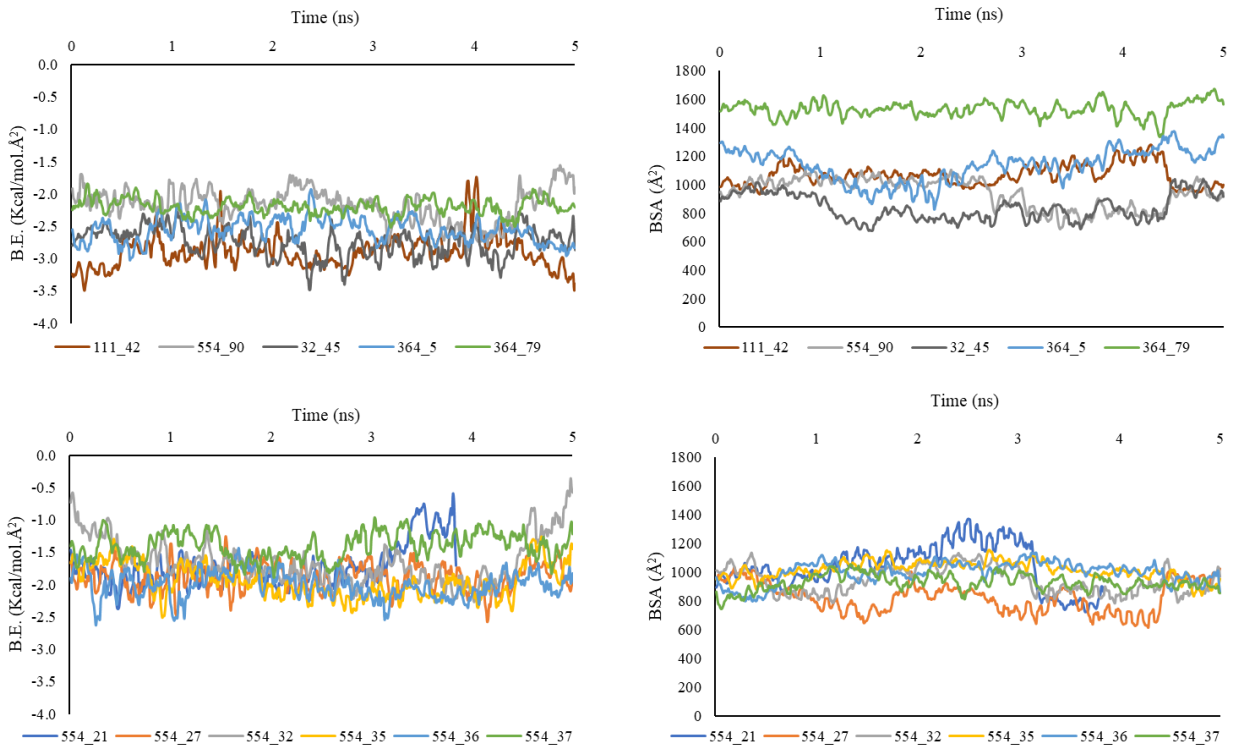

Figure 4. At first glance, 554_32 can be seen to have a wide range of fluctuations and it also ends at a state of poor B.E/BSA conformation. The sudden fluctuations in its BSA plot on the right also suggests that there is a decrease in its stability as a complex. So, it is also ranked towards the end.

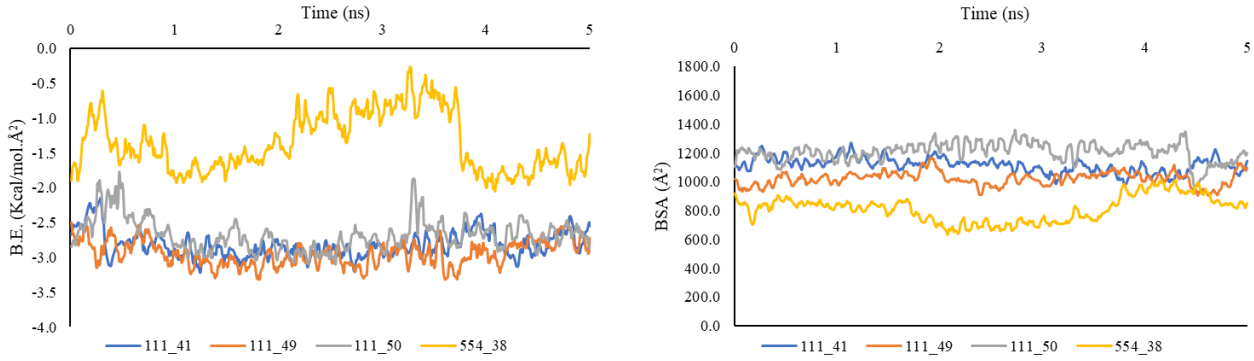

Figure 5. From the plot, it can be easily determined that 554_38 has a wide range of fluctuations and the characteristics of a bad curve. And judging by its BSA plot, which does not show such fluctuations, the B.E. itself was the major factor for such behavior or in other words, the complex formation was not strong enough throughout the MDs run. So, after matching it was found that this complex also belonged to the bottom SD rankings.

Table 1. Table showing the SD for B.E/BSA of the complexes

Ranks	PDB ID	SD (B.E/BSA)	Ranks	PDB ID	SD (B.E/BSA)
1	364_79	0.190240831	37	554_36	0.340897761
2	111_29	0.227593382	38	554_100	0.346205435
3	111_32	0.233058835	39	554_83	0.346717713
4	111_34	0.239255269	40	32_83	0.346896402
5	111_57	0.242334666	41	32_45	0.346910878
6	364_84	0.242908386	42	111_50	0.348224162
7	111_94	0.246483281	43	554_48	0.351733495
8	111_97	0.247339352	44	554_57	0.353441863
9	111_85	0.251634841	45	554_35	0.354458228
10	111_78	0.259219776	46	1025_36	0.356898953
11	111_1	0.26881807	47	111_59	0.3641755
12	364_5	0.270545062	48	554_44	0.3682385
13	111_63	0.273618471	49	32_66	0.368384255
14	111_41	0.27442873	50	111_2	0.377129791
15	111_93	0.277699384	51	111_42	0.378153253
16	111_46	0.279336396	52	390_73	0.380466238
17	111_89	0.281926236	53	554_90	0.384081338
18	111_49	0.287891285	54	554_27	0.385102616
19	111_28	0.288207475	55	111_5	0.414379134
20	111_91	0.288922713	56	1025_62	0.425837608
21	111_44	0.290653001	57	554_52	0.429927657
22	554_49	0.291001072	58	554_67	0.430050362
23	554_94	0.298542724	59	111_67	0.435439885
24	111_52	0.299498785	60	111_96	0.440606668
25	32_87	0.299968165	61	554_32	0.440922973
26	554_84	0.306338601	62	554_21	0.451804086
27	554_69	0.310979697	63	554_61	0.460973799
28	32_50	0.31334538	64	32_42	0.498278555
29	554_42	0.314722051	65	554_68	0.502928413
30	111_7	0.317502373	66	554_38	0.517871914
31	111_17	0.317640932	67	111_62	0.519313383
32	111_26	0.321138746	68	32_37	0.519898997
33	554_37	0.325726893	69	213_11	0.52233762
34	554_65	0.327156567	70	281_57	0.526868345
35	554_18	0.338926401	71	32_28	0.538977389
36	111_58	0.339534142	72	245_45	0.544821427
			73	213_81	0.767519969

Table 2. Table showing the SD for BSA of the complexes

Ranks	PDB ID	SD	Ranks	PDB ID	SD
1	111_32	52.35408	37	32_83	94.3379
2	111_29	55.84232	38	364_84	100.4289
3	554_44	57.26305	39	111_52	101.252
4	111_63	58.7681	40	32_66	101.625
5	554_69	60.3717	41	111_50	101.7162
6	554_68	60.61795	42	554_48	102.2862
7	111_89	61.36574	43	554_38	102.4204
8	111_7	61.50434	44	111_46	102.4836
9	111_34	61.71085	45	32_45	102.7904
10	111_49	61.91084	46	111_59	104.3247
11	111_97	62.92951	47	554_42	106.4861
12	111_94	67.33185	48	32_42	108.4581
13	111_58	68.30222	49	554_32	110.7232
14	111_5	70.14354	50	111_28	112.7425
15	554_35	70.2429	51	554_57	114.6274
16	111_41	71.08983	52	554_90	114.8709
17	554_37	71.48291	53	1025_62	115.443
18	111_57	72.45493	54	554_83	116.4633
19	111_85	73.90608	55	554_61	117.0182
20	364_79	75.37392	56	111_2	117.763
21	32_50	80.70588	57	390_73	118.2041
22	111_93	80.86411	58	554_52	121.1276
23	111_44	81.19863	59	554_27	121.4296
24	111_1	82.46945	60	111_91	129.3394
25	111_26	82.76506	61	364_5	130.8063
26	554_94	82.93288	62	111_67	131.0254
27	554_67	83.39686	63	554_49	135.7099
28	554_100	85.11691	64	281_57	145.9444
29	554_36	85.2485	65	111_96	149.9693
30	111_17	87.41493	66	1025_36	154.069
31	111_78	87.73521	67	213_81	154.5076
32	32_87	88.71184	68	554_18	167.1743
33	111_42	88.86793	69	111_62	168.0611
34	554_65	89.13067	70	554_21	182.9889
35	554_84	89.38995	71	213_11	189.8715
36	245_45	89.86362	72	32_37	195.7161
			73	32_28	287.1189

Table 3. Table showing the Sc of the complexes

Rank	PDB ID	SD	Rank	PDB ID	SD
1	111_91	0.040049	37	111_62	0.060416
2	111_85	0.043083	38	554_49	0.061057
3	111_42	0.043731	39	554_36	0.061551
4	111_28	0.044843	40	111_93	0.062288
5	111_17	0.044923	41	111_7	0.063512
6	111_78	0.045095	42	554_32	0.063846
7	111_34	0.046317	43	554_94	0.063922
8	111_96	0.046513	44	554_42	0.064534
9	111_67	0.047003	45	111_58	0.065017
10	32_50	0.047055	46	245_45	0.065155
11	111_1	0.047221	47	213_81	0.065199
12	111_50	0.047231	48	554_83	0.065305
13	32_87	0.047434	49	554_18	0.065925
14	111_2	0.047565	50	364_5	0.067604
15	111_46	0.04783	51	554_35	0.068594
16	111_29	0.048838	52	554_100	0.069504
17	32_83	0.048843	53	554_90	0.070849
18	111_32	0.049187	54	32_37	0.070941
19	111_89	0.049598	55	554_57	0.07599
20	111_44	0.049888	56	554_37	0.076269
21	32_45	0.050243	57	554_48	0.076469
22	111_97	0.05041	58	554_38	0.076632
23	111_94	0.050622	59	554_67	0.076667
24	111_52	0.052057	60	32_42	0.077724
25	111_57	0.05303	61	554_68	0.079401
26	364_79	0.053409	62	213_11	0.080552
27	364_84	0.053562	63	390_73	0.080723
28	111_59	0.053714	64	554_69	0.080781
29	111_5	0.053888	65	554_84	0.083611
30	554_65	0.054859	66	281_57	0.084994
31	111_49	0.055619	67	554_27	0.088801
32	111_41	0.056409	68	32_28	0.088867
33	32_66	0.057264	69	554_61	0.09026
34	554_44	0.057897	70	554_21	0.091654
35	111_63	0.058929	71	554_52	0.09989
36	111_26	0.059475	72	1025_36	0.102537
			73	1025_62	0.11083

Table 4. Table showing the RMSD rankings for each complex

Rank	PDB_ID	SD	Rank	PDB_ID	SD
1	364_79	0.339	37	554_38	0.628
2	111_63	0.343	38	111_34	0.634
3	364_5	0.37	39	111_1	0.635
4	111_26	0.375	40	32_42	0.636
5	111_67	0.375	41	111_44	0.643
6	364_84	0.378	42	554_84	0.655
7	111_62	0.401	43	554_42	0.659
8	111_57	0.421	44	554_100	0.667
9	111_89	0.426	45	111_7	0.675
10	111_5	0.427	46	554_90	0.677
11	32_87	0.43	47	111_2	0.701
12	111_91	0.431	48	554_69	0.701
13	111_49	0.453	49	32_83	0.709
14	32_50	0.454	50	554_18	0.735
15	111_52	0.456	51	554_37	0.751
16	111_46	0.462	52	111_59	0.765
17	111_85	0.491	53	32_45	0.774
18	111_41	0.492	54	281_57	0.786
19	111_97	0.492	55	111_32	0.835
20	111_28	0.501	56	111_93	0.85
21	554_21	0.503	57	111_29	0.863
22	111_50	0.515	58	111_17	0.871
23	111_94	0.522	59	554_67	0.879
24	111_42	0.53	60	1025_36	0.881
25	554_48	0.551	61	554_32	0.9
26	554_94	0.555	62	32_28	0.998
27	213_11	0.558	63	245_45	1.02
28	554_36	0.559	64	554_68	1.042
29	554_65	0.572	65	554_83	1.053
30	111_96	0.591	66	32_66	1.065
31	554_49	0.591	67	213_81	1.155
32	1025_62	0.602	68	554_52	1.16
33	111_78	0.604	69	554_35	1.174
34	554_44	0.609	70	554_57	1.38
35	111_58	0.615	71	390_73	1.382
36	32_37	0.628	72	554_61	1.474
			73	554_27	4.224

Supplementary Files

Charts, images and tables information

All images were originally made by the author using UCSF Chimera v1.16. All charts and tables were made using Microsoft Excel by the author.

MutDock Simulation File for Epitope II

Run name: cobra_unknown2

Receptor path: ./structures/6fpa_minimized2.pdb

Ligand path: ./structures/cobratoxin_min.pdb

Probe: 1.4

Interaction limit: 3

Distance match limit: 1.8

Angle match limit: 70.0

Max rot limit: 5

Phobic dis limit: 2

Solution folder: results/DARP_COBRA_UNKNOWN2

Receptor residues: B34, B35, B36, B37, B38, B42, B43, B46, B67, B68, B69, B70, B71, B76, B79, B100, B101, B102

Receptor residues: B104, B109, B112, B113, B133, B135, B136

Ligand residues: D1, D2, D3, D4, D9, D10, D11, D12, D13, D14, D15, D16, D17, D18, D19

Ligand residues: D57, D58, D59, D60, D61, D62, D63, D64

Mutation residues: B35, B36, B38, B46, B69, B71, B79, B80, B101, B102, B104, B112, B113

Rosetta affinity maturation DESIGN file for 1 complex

-database /home/shared/rjp0029_lab/Rosetta/main/database

-s /home/rzb0071/Input_Structures/Best/known_6.pdb

-resfile /home/rzb0071/affmat.resfile

-packing:resfile

-nstruct 100

-minimize_sidechains

-score:weights ref2015.wts

-linmem_ig 10

-multi_cool_annealer 10

-ex1

-ex2

-ex3

-ex4

-no_his_his_pairE

-out:path:pdb /home/rzb0071/Results/DarpinAffMat/known/

Rosetta affinity maturation RESFILE

NATAA # this default command applies to all residues that are not given non-default commands

start

6 B ALLAAxc # allow all amino acids except cys

9 - 10 B ALLAAxc

33 - 39 B ALLAAxc

42 - 43 B ALLAAxc

46 B ALLAAxc

66 - 72 B ALLAAxc

75 - 76 B ALLAAxc

79 - 80 B ALLAAxc

99 - 105 B ALLAAxc

108 - 109 B ALLAAxc

111 - 113 B ALLAAxc

132 - 138 B ALLAAxc

141 - 142 B ALLAAxc

145 - 146 B ALLAAxc

Antigen sequence:

IRCFITPDITSKDCPNGHVCYTKTWCDAFCSIRGKRVDLGCAATCPTVKTGVDIQCCSTDNCN PFPTRKRP.

Final 51 DARPin sequences:

364_79

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSRGATPLDLAASSGHLEIVEVLLKTGADVNA SDKYGLTPLDYAAAKGHLEIVEVLLKAGADVNAKDMTGLTPLVFAAMAGHLEIVEVLLKHGAD VNATDKLGLTPFDWAILTGNEDIAEVLQKAAKLN

111_29

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

111_3

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGATPLDLAAASGHLEIVEVLLKTGADVNA SSKMGDTPLDLAAGAGHLEIVEVLLKAGADVNAKDKDGLTPLAWAAMYGHLEIVEVLLKHGA DVNASDKKGKTPFDYAIENGNEDIAEVLQKAAKLN

111_34:

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGATPLHLAAASGHLEIVEVLLKTGADVNA SSKMGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARDGHLEIVEVLLKHGAD VNASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_57

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA SSSQGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDAEGLTPLAYAARHGHLEIVEVLLKHGADV NASDKRGLTPFDYAILNGNEDIAEVLQKAAKLN

364_84

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSAGRTPLALAASSGHLEIVEVLLKTGADVNA SDKYGLTPLDYAAAKGHLEIVEVLLKAGADVNAKDMSGLTPLVFAAMAGHLEIVEVLLKHGAD VNATDKKGLTPFDWAILTGNEDIAEVLQKAAKLN

111_94

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA SSSQGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDAEGLTPLAYAARHGHLEIVEVLLKHGADV NASDKRGLTPFDYAILNGNEDIAEVLQKAAKLN

111_97

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGKTPLVYAAKHGHLEIVEVLLKHGAD
VNASDKDGKTPFDHAIENGNEDIAEVLQKAAKLN

111_85

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKMGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGKTPLVYAAKHGHLEIVEVLLKHGAD
VNASDKDGKTPFDHAIENGNEDIAEVLQKAAKLN

111_78

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNALDADGLSPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_1

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

364_5

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSAGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKYGLTPLDYAAAKGHLEIVEVLLKAGADVNAKDDTGLTPLVFAAMAGHLEIVEVLLKHGAD
VNATDKKGMTPFDYAIANGNEDIAEVLQKAAKLN

111_63

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARDGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_41

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_93

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARDGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_46

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKMGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGAD
VNASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

111_89

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_49

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARDGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

111_28

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_91

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

111_44

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

554_49

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSEGATPLVLAARSGHLEIVEVLLKTGADVNA
SDKKGRTPLDDAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGA
DVNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_94

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_52

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

32_87

DLGEKLLKAARAGQDDEVRILMANGADVNAEDYDGATPLVLAARSGHLEIVEVLLKTGADVNA
SDAMGRTPLVDAAEAGHLEIVEVLLKAGADVNATDAYGMTPLDYAAKLGHLEIVEVLLKHGAD
VNASDKQGKTPFDYAIENGNEDIAEVLQKAAKLN

554_84

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSEGATPLVLAARSGHLEIVEVLLKTGADVNA
SDKKGRTPLDDAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGA
DVNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_69

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

32_50

DLGEKLLKAARAGQDDEVRILMANGADVNAEDYDGRTPLHLAAMSGHLEIVEVLLKTGADVN
ASDAMGLTPLDFAAEAGHLEIVEVLLKAGADVNATDAYGRTPLYFAAANGHLEIVEVLLKHGAD
VNASDKQGKTPFDAAIDNGNEDIAEVLQKAAKLN

554_42

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSEGRTPLALAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_7

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGATPLHLAAASGHLEIVEVLLKTGADVNA
SSKMGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGAD
VNASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_17

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

111_26

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGKTPLVYAAKHGHLEIVEVLLKHGAD
VNASDKDGKTPFDHAIENGNEDIAEVLQKAAKLN

554_37

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_65

DLGEKLLKAARAGQDDEVRILMANGADVNAESSTGDTPLAYAAQSGHLEIVEVLLKTGADVNA
SNKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_18

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_58

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

554_36

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSEGATPLALAARSGHLEIVEVLLKTGADVNA
SDKKGRTPLVDAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_100

DLGEKLLKAARAGQDDEVRILMANGADVNAESSTGDTPLAYAAQSGHLEIVEVLLKTGADVNA
SNKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_83

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_50

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

554_48

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLALAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_57

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSEGRTPLALAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_35

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSKGRTPLDLAASSGHLEIVEVLLKTGADVNA
SDKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

1025_36

DLGEKLLKAARAGQDDEVRILMANGADVNAEDYLGKTPLAYAAESGHLEIVEVLLKTGADVNA
SDKFGQTPLVLAAAAGHLEIVEVLLKAGADVNATDADGLTPLDYAARHGHLEIVEVLLKHGADV
NATDKKGLTPFDLAILNGNEDIAEVLQKAAKLN

111_59

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

554_44

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGDTPLALAAKSGHLEIVEVLLKTGADVNA
SNKKGRTPLDDAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGA
DVNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_2

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN

111_42

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKKGLTPFDYAILNGNEDIAEVLQKAAKLN

554_90

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGDTPLALAAKSGHLEIVEVLLKTGADVNA
SNKKGQTPLVAAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

554_27

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSNGDTPLALAAQSGHLEIVEVLLKTGADVNA
SNKKGLTPLDYAAEAGHLEIVEVLLKAGADVNAQDADGMTPLDYAAMRGHLEIVEVLLKHGAD
VNATDKHGYTPFDYAIMNGNEDIAEVLQKAAKLN

111_67

DLGEKLLKAARAGQDDEVRILMANGADVNAEDSLGATPLDLAAASGHLEIVEVLLKTGADVNA
SSKEGLTPLDLAAGAGHLEIVEVLLKAGADVNAKDADGLTPLAYAARHGHLEIVEVLLKHGADV
NASDKHGKTPFDYAIENGNEDIAEVLQKAAKLN