Studying the Effects of Mutations on the Structures and Binding of Therapeutic Proteins Towards Improving the Engineering of Protein Functions

by

Sumaiya Islam

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
May 7, 2022

Keywords: protein mutations, protein-protein interactions, binding interfaces, statistical analyses, computational protein design, force fields

Approved by

Robert J. Pantazes, Chair, Assistant Professor of Chemical Engineering
F. Stephen Dobson, Professor of Biological Sciences
Qinghua (Peter) He, Associate Professor of Chemical Engineering
Andrew J. Adamczyk, Assistant Professor of Chemical Engineering

Abstract

The high effectiveness of proteins in biological systems is a result of their mutational histories. A single amino acid substitution, the most frequent of changes in proteins, can alter their physiochemical properties and functions. Mutations impact protein folding and interactions, and thus their functions in a biological system. These mutations may be beneficial and identifying them can provide insights towards improving the engineering of proteins. Mutational analyses are extensively used to study protein structures and functions for different purposes. For therapeutic purposes, such analyses can lead to more efficient engineering of proteins: from identifying beneficial interactions for improving protein binding to identifying how point mutations in pathogens can impact immune responses in humans. Alternatively, the coronavirus pandemic has shown examples of how single point mutations in pathogens, like viruses, can lead to different variants with drastic consequences to human health; some variants may spread more easily in humans or show signs of resistance to existing treatment options. This has led to a belief that improved engineering of proteins towards therapeutic developments requires the study of protein interactions responsible for protein binding and functions. Both engineering native proteins and designing new proteins require computational techniques to overcome the perplexities of traditionally used experimental techniques. This dissertation is directed towards bridging the gap between computational protein structure and function by building a statistical understanding of various aspects of protein interactions and functions, thus contributing towards protein engineering techniques for therapeutic purposes. It is based on the hypothesis that knowledge about point mutations can then be directed towards developing an understanding of the structure to function relationship in proteins. Each chapter focuses on the effect of mutations on protein functions from a unique perspective. The chapters progress from characterizing the

binding interfaces of therapeutic proteins to quantifying the effects of point mutations on protein binding, making an in-depth analysis of the effects of antigenic mutations on therapeutic protein interactions, and identifying the impact of viral mutations on immune responses in humans. The findings in each chapter can contribute to the study of engineering proteins to meet specific therapeutic needs.

Acknowledgments

First and foremost, praises and thanks to the Almighty, the Most Gracious and the Most Merciful.

I would like to express my sincerest gratitude to my academic advisor, Robert Pantazes, for his wisdom and guidance throughout my doctoral journey. He has shown incredible patience and helped me grow into a researcher at my own pace. This dissertation would not have been possible without his constant support. I would also like to thank my other dissertation committee members for their insights and recommendations throughout my graduate school experience. In addition, I would like to thank my friends and colleagues in the department for the support, recommendations, inspirations and, most of all, friendship for the last several years

I am truly indebted to my parents, Shafiqul Islam and Moslema Khatun, without whom I would not have been here today. They have always supported my education and inspired me to pursue my engineering career. I would like to thank my brother, Moshfique, for his love and support for many years. I would also like to thank all my friends and family for all their encouragement.

I dedicate this dissertation to all women who have remained steadfast at pursuing their dreams and not allowed their lives to be governed by “social norms”.

Table of Contents

Abstract 2

Acknowledgments 4

List of Tables 7

List of Figures 11

List of Abbreviations 13

Chapter 1 Introduction 14

References 26

Chapter 2 Development and Analyses of a Database of Antibody – Antigen Complexes 43

Introduction 44

Methods 44

Results and Discussions 46

References 53

Chapter 3 Development of Similarity Matrices based on Protein Binding Interactions 57

Introduction 58

Methods 59

Results and Discussions 63

References 89

Chapter 4 Analysis of How Mutations Disrupt Hotspot Binding Interactions 93

Introduction 94

Methods....95

Results and Discussions....97

References....115

Chapter 5 The Effects of SARS-CoV-2 Spike Protein Mutations on the Immune System....121

Introduction....122

Methods....124

Results and Discussions....128

References....155

Chapter 6 Conclusion....161

References....167

List of Tables

Table 1: Matrix showing the means of percentage of changes in binding energy due to point mutations in protein antigens using the CHARMM force field. 70

Table 2: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the CHARMM force field. 71

Table 3: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the CHARMM force field. 72

Table 4: Matrix showing the percentage of changes in binding energy due to point mutations in protein antigens using the Amber force field. 73

Table 5: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the Amber force field. 74

Table 6: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the Amber force field. 75

Table 7: Matrix showing the percentage of changes in binding energy due to point mutations in protein antigens using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. 76

Table 8: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the Rosetta force field. 77

Table 9: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the Rosetta force field. 78

Table 10: Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the CHARMM force field. 79

Table 11: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the CHARMM force field..... 80
Table 12: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the CHARMM force field. 81
Table 13:Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the Amber force field..... 82
Table 14: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the Amber force field..... 83
Table 15: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the Amber force field..... 84
Table 16: Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the Rosetta force field..... 85
Table 17: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the Rosetta force field..... 86
Table 18: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the Rosetta force field..... 87
Table 19: Total count of each type of residue in the epitopes of the antigens and the paratopes of the antibodies, for each force fields. 88
Table 20: Heat Map showing all interactions of antibody-antigen interfaces in the selected complexes with the CHARMM force field..... 106
Table 21: Heat Map showing all interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field..... 107

Table 22: Heat Map showing the best five interactions of antibody-antigen interfaces in the selected complexes with a CHARMM force field. 108
Table 23: Heat Map showing the best five interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field. 109
Table 24: Heat Map showing the worst three interactions of antibody-antigen interfaces in the selected complexes with the CHARMM force field. 110
Table 25: Heat Map showing the worst three interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field. 111
Table 26: Mutational analysis results compared to the corresponding values in the antibody-antigen database for the CHARMM force field. 112
Table 27: Mutational analysis results compared to the corresponding values in the antibody-antigen database for the Amber force field. 113
Table 28 Mutational analysis results compared to the corresponding values in the antibody-antigen database for the Rosetta force field. 114
Table 29: Allele Distribution for each Ethnicity 138
Table 30: Distribution of the Ethnicities for each Allele. 139
Table 31: Top (50%) alleles of each Ethnicity. 141
Table 32: Alpha Variant Mutations Effects on Various Ethnic Groups. 142
Table 33: Beta Variant Mutations Effects on Various Ethnic Groups 144
Table 34: Gamma Variant Mutations Effects on Various Ethnic Groups. 145
Table 35: Delta Variant Mutations Effects on Various Ethnic Groups. 146
Table 36: Omicron Variant Mutations Effects on Various Ethnic Groups. 147
Table 37: Mutations Predicted to Extremely Adverse Effects on T-Cell Epitopes. 150

Table 38: Predicted Mutation Effects for Position 505 and for Position 455 on the Spike protein. 151

List of Figures

Figure 1: Flowchart to show the arrangement of the projects and their relationship to the development of therapeutic proteins. 25

Figure 2:Cumulative percentage binding energies of the antigen residues. 49

Figure 3:Cumulative percentage binding energies of the antibody residues. 49

Figure 4: Distribution of amino acids in the paratopes of the antibodies. 50

Figure 5: Distribution of amino acids in the epitopes of the antigens. 50

Figure 6: Log2 change in amino acid usage in the significant residues versus the antibody as a whole. 51

Figure 7: Log2 change in amino acid usage in the significant residues versus the antigen as a whole 51

Figure 8: Percentage of binding threshold in the best 12 consecutive antigen amino acid sequence. 52

Figure 9: Algorithm for the calculations required to make the matrix 69

Figure 10: Disrupted Hotspot Interactions in Antibody M396 with the SARS-CoV-2 RBD.. .. 101

Figure 11: Disrupted Hotspot Interactions in Antibody S230 with the SARS-CoV-2 RBD..... 102

Figure 12: Disrupted Hotspot Interactions in Antibody 80R with the SARS-CoV-2 RBD. 103

Figure 13: Distribution of the antibody contact residues in all the structures. 104

Figure 14: Distribution of the antigen contact residues in all the structures. 104

Figure 15: Distribution of the shape complementarity in all the structures. 105

Figure 16: Distribution of the Buried Surface Area in all the structures. 105

Figure 17: Structure details of the SARS-CoV-2 Spike protein. 137

Figure 18: Distribution of the Mutations across the Spike protein in the different variants 152

Figure 19: Distribution of the Surface Exposed Residues. 153

Figure 20: Residues with small solvent exposed surface area can contribute to binding. 154

List of Abbreviations

ACE-2	Angiotensin-Converting Enzyme 2
CHARMM	Chemistry at Harvard Molecular Mechanics
COVID-19	Coronavirus Disease 2019
HLA	Human Leukocyte Antigen
IMGT	International Immunogenetics
MHC	Major Histocompatibility Complex
PPIs	Protein-Protein Interactions
RBD	Receptor Binding Domain
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
VDWs	van der Waals
VOC	Variant of Concern

Chapter 1 Introduction

"Every living being is also a fossil. Within it, all the way down to the microscopic structure of its proteins, it bears the traces if not the stigmata of its ancestry."

Jacques Monod

Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology (1972) [1]

Proteins make up a diverse set of the most important and precisely structured macromolecules in living systems. In the ribosome of the cell, polypeptides are created by combining individual amino acids, which then fold and assemble into functioning proteins. Their functions range from enzymatic catalysis, transportation of ions and molecules, counteracting foreign substances, and the regulation of cellular and physiological activities [2]. Examples of the many functions of proteins in biological systems include receptors and transcription factors binding to extra and intracellular molecules that are integral to cellular signaling, antibodies binding to foreign objects to protect a host organism from disease or invasion, enzymes carrying out life-sustaining reactions to keep an organism functioning, and a vast selection of different proteins working to maintain the structural integrity of cells and the movement of molecules between compartments.

The existence of such a diverse pool of proteins is a testament to the evolutionary forces that have shaped the complex biological systems [3], and unraveling these forces has been a challenge for evolutionary biologists. The high effectiveness of proteins is a result of their mutational histories driven by natural selection. Zuckerkandl and Pauling introduced the idea of the molecular clock that protein sequences change at an almost uniform rate over a long period, laying the foundations of molecular evolution [4]. It is largely believed that genotypic and phenotypic changes are driven by the selection of "favorable" mutations, though Kimura's

"neutral theory" says that, at a molecular level, evolution is driven by random drift of neutral and nearly neutral changes [5]. It is understood that there are other factors to be considered, like population size and structure, ecological opportunities and changes, and much more, for natural selection of variations produced by genotypic changes to drive adaptive evolution. The more important the protein or parts of the protein, the less likely they are to change [6], leading to the belief that when changes are made to important proteins, they are not conservative. Amino acid substitutions may lead to relatively little change, whereas at certain times a single amino acid substitution may lead to a radical change in function. Yet, the functional effect of any given substitution may depend on the presence or absence of other substitutions. This has led to applying artificial selection to create new proteins, directed evolution [7], [8], which has become one of the most widespread tools used for engineering proteins for improved or novel functions.

Biological engineers attempt to modify or design proteins for an assortment of applications. The engineered design of proteins has a wide variety of practical impacts, ranging from industrial applications such as catalytic activities for chemicals and biosensors to biomedical applications such as therapeutic medicine, the creation of gene switches, and signal transduction pathways. Engineered proteins are also used in agriculture and the food industry, environmental monitoring and bioremediation, detergents, biopolymer production and nanotechnology, and biofuel production [9]. Both engineering native proteins and designing new proteins require computational techniques to overcome the perplexities of traditionally used experimental techniques.

Protein functions greatly depend on their three-dimensional structure, but accurately predicting the native structure of a protein from its amino acid sequence was a longstanding challenge in the field of computational molecular biology. However, with improvements in the

methods and expansion of computational resources, structure prediction has advanced to the level of reliably designing novel proteins "from scratch" (i.e., *de novo* design) [10]. More recently, after years of scientific research to answer the question of the "protein folding problem" [11], AlphaFold has shown success in highly accurate protein structure prediction using a deep learning algorithm [12]. The accurate modeling of the structure, thermodynamics, and protein interactions with other molecules are fundamental to successful *de novo* protein design [13]. There have been attempts to computationally design proteins with improved stability [14], [15], improved binding [15] – [17], and novel functions [17] – [19]; from the ratio of successful designs that have shown desired properties to the unsuccessful designs, it is evident that there is a need for improvements in protein design techniques. Accurate protein structure prediction and design necessitate understanding the properties and behavior of proteins from a theoretical perspective.

Given the significant progress in computational techniques for *de novo* protein design, the leap from "static" protein structures to dynamic, functional proteins is not yet well understood [22]. The inability to accurately model electrostatic and hydrogen bonding interactions, to appropriately balance electrostatic and solvent effects, and to understand the necessity for quantum mechanical effects (e.g., the polarizability of electron clouds) to calibrate potentials are all ongoing challenges in designing protein binding interfaces [23] – [25]. Another important unsolved question is correctly balancing the energetics when replacing one type of residue with another and understanding the relationship of local interactions and backbone motions [22], [26].

Although protein structures are usually robust to sequence changes in the natural selection process [27], minor sequence changes from a single point mutation can sometimes

cause disturbances to the protein structure. The static and dynamic properties of a protein are the results of its sequence, and mutations in the sequence can affect protein folding and stability [28], [29]. Most mutations in proteins either act as destabilizers or have relatively no effect on stability, yet a fraction of mutations can increase the stability over the wild type [27].

Protein-protein interactions (PPIs) are crucial for biological systems to function effectively [25] – [28]. Mutations can alter the kinetics and thermodynamics of PPIs [34], [35], thus affecting the functions of proteins. Identifying mutations that are beneficial in known proteins and understanding the mode of their impact can provide insight into the engineering of proteins for desired applications. To understand PPIs, identification of the interaction sites is required before using analyses to study the impact of point mutations on protein-protein affinities. Such understanding can be related to the function [32] – [34] of proteins and used in therapeutic developments such as targeting interaction mutants [39] and in catalytic reactions such as developing biological catalysts.

PPIs are complex and factored by large sets of variations. Deciphering the complicated details of PPIs requires the use of both physical chemistry analysis of and observed interaction potentials from already known structures of protein complexes [40], [41]. Protein interfaces have been studied extensively to develop a clear perception of the forces and the recognition processes at a molecular level guiding these interactions. Comprehending the specific types of residues that have more involvement in the level of affinity and specificity of PPIs is paramount to understanding these interactions and designing proteins. Understanding the interaction-function relationships for proteins in general and attempting to improve design techniques entails making improvements in the models and force fields used.

The introduction of mutations into the native protein structures can be used to modulate their physical activities and understand how they affect the different aspects of protein interactions and thermodynamics. This would also help provide a detailed understanding of the conformational and dynamic formations of the secondary structures (i.e., α-helices, β-sheets, and loops) of proteins and their folding mechanisms, which is crucial to solving any problem related to function. Identification of the favorable and unfavorable mutations in protein binding can also help improve experimental studies like mutagenesis and be used in evaluating predicted interfaces [42]. Attempted modifications of the specificity and affinity of PPIs can present interesting results that can impact both industrial and therapeutic fields [39] – [41].

Various experimental and computational techniques have been used to build a broader and more thorough understanding of protein interactions. Both experimental and computational Alanine-scanning mutagenesis of interfacial residues [42] – [47], analysis of crystallographic structures of protein interactions [48] – [50], studies focusing on the physical and chemical properties of protein interfaces [55], [56], evolutionary trace method based on conservation of important residues [57], and structural and thermodynamic studies of protein interfaces [54] – [59] are examples of some of the techniques used. Computational tools [64] have been developed to investigate the role of electrostatics in PPIs and analyze the effects of disturbances in protein structure and protein-protein complexes.

Statistical analysis has played a crucial role in the study of PPIs. A large non-redundant set of known structures [40] has been used to derive protein-protein interface residue composition and residue-residue contact preferences. A statistical analysis of structurally non-redundant protein-protein interfaces and symmetry-related oligomeric interfaces was conducted

to understand the role of the hydrophobic effect in PPIs [65]. Atom-based statistical pair potentials [66] were developed to evaluate the strength of PPIs.

Mutations in proteins can cause them to assume different conformational states than their native state, and thus improve, or even evolve their function [67]. Therefore, understanding the intricate details of PPIs entails focusing on mutations in native protein sequences and analyzing the impact of these mutations on protein binding and folding. Besides gaining structural insights, the thermodynamics of protein interactions is crucial to understanding these details of PPIs, understanding the effects of mutations on them, and using them to meet specific needs.

The binding affinity change upon mutation is a key indicator of the effects of the mutation on the interactions, functions, and stability of the protein. Binding affinity can be determined experimentally, which can be expensive and laborious, or predicted computationally, which requires accurate descriptive parameters. There are available databases [68], [69] of the binding free energy changes upon mutation, measured experimentally. Different methods have been evaluated for the prediction of the effects on mutation using experimentally derived data. Web tools and algorithms [70], [71] have also been created to predict the effects of point mutations on protein stability. Different methods have concentrated on the effects of single point mutations on protein stability and interactions [73] – [77]. A detailed review [77] of the different experimental and computational methods used for predicting the effects of mutations on protein-protein binding interactions is available.

The first two projects in this dissertation are based on studying protein-protein interfaces and understanding the effects of mutations on protein binding. In the first project, the antibody-antigen interface features are identified, e.g., the number of residues important to binding in terms of energetics and the types of residues for binding compared to those found in the core

structure. Antibody-antigen interfaces are different from other binding interfaces [62], and studying mutations on antibody-antigen interfaces requires knowledge of their specific features. Identifying the features of the interface provided information for building the second project. In the second project, the effects of mutations on antibody-antigen interface binding are studied. The most important residues to binding in terms of energetics identified in the first project are used for the mutational analysis in this project. The mutational analysis was conducted for both the antibody and antigen residues. This also revealed the relative importance of amino acids in binding in the antibodies and protein antigens.

Starting in late 2019, the novel Coronavirus Disease 2019 (COVID-19), caused by the virus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), rapidly spread to become a global pandemic. As the pandemic progressed, all known facts about viral antigens and protein engineering techniques became important resources against combating the virus. SARS-CoV-2 proved to be a fast-mutating virus. After the initial surge of infection, variants of the virus were documented with increased severity in transmissibility and infectivity in humans. Until early 2021, four Variants of Concern (VOCs) were identified by WHO, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1 or B.1.1.28), and Delta (B.1.617.2). Of these, the Beta and Gamma variants have not spread to the same extent as the Alpha and Delta variants. In mid-2021, there was a new VOC, Omicron (B.1.1.529) with 30 mutations in the Spike protein alone compared to the 15 Spike mutations of the Delta variant.

The virus enters the host cell by binding to a cell receptor, Angiotensin-Converting Enzyme 2 (ACE-2) on the cell membrane. Mutations in the Spike protein of SARS-CoV-2 alter the protein function in two ways: affecting the stability of the Spike protein and altering the binding properties of the receptor-binding domain (RBD) to ACE-2 receptors[78]. The binding

properties of the Spike protein to the ACE-2 receptor change the infectivity and transmissibility of the virus variants. More importantly, mutations can aid the spread of viruses in the human body by immune evasion. The evasion of neutralization by antibodies can be possible due to the mutated residues in the Spike protein altering the targeted epitopes. Studies have found that the variants of SARS-CoV-2 with RBD mutations, including K417N, E484K, and N501Y, were highly resistant to neutralization by antibodies in vaccinated individuals by increased RBD binding to the ACE-2 receptor[79], [80].

The pandemic taught an important lesson that led to studies included in the successive chapters: understanding how mutations in antigenic proteins affect binding could potentially be our way of "getting ahead of the game". When the first publication of the structure of the SARS-CoV-2 Spike protein[81] reported that three anti-SARS antibodies, M396[82], S230[83], and 80R[84], showed no appreciable binding to SARS-CoV-2, despite the high structural homology of the Spike proteins of the viruses, an analysis of why this happened was made. The findings from that analysis led to the third project in this dissertation. A detailed study of how antigenic mutations disrupt hotspot interactions was necessary. An understanding of the different phenomena that disrupt antibody binding when antigen interfaces mutate is important for the development of therapeutics against possible future pandemics.

The development of vaccines to combat the novel virus was imminent, with the first mRNA vaccine developed in 2020. Currently the vaccines available use four different technologies: whole virus vaccine (Sinopharm and Sinovac[85]), RNA or mRNA vaccine (Pfizer-BioNTech[86], Moderna[87]), non-replicating viral vector (Johnson & Johnson[88], Oxford-AstraZeneca[89], Sputnik V[90]), and protein subunit (Novavax[91]). The most effective vaccines against SARS-CoV-2[92] use mRNA to encode for the Spike protein building on the

adaptive immunity in humans[93]. Many effective neutralizing anti-SARS-CoV-2 antibodies that target the RBD to prevent binding to the ACE-2 receptors can be potentially used as antibody therapeutics[94]. Neutralizing antibodies have been studied as a potential treatment for SARS[95] due to their ability to block the biological effects of viral particles[96]. Different antibodies have been reported to have neutralizing activity against the SARS-CoV-2 infection including monoclonal antibodies[94] – [98], synthetic nanobodies[103] – [107], and a variety of antibody cocktails[108] – [111].

Eventually, with newer variants infecting humans at different rates and causing re-infection in vaccinated people led to the fourth project in this dissertation. A detailed analysis to understand how the Spike protein mutations affect the immune systems of people with different ethnic backgrounds was designed. Predictions on the effect of Spike protein mutations on the B-cell and T-cell epitopes were used to identify which mutations are notorious for evading the immune systems in humans. The findings from this study will help identify future mutations to "look out for" and, also, identify the ethnicities that would likely be most impacted by the next variant and would require more targeted therapeutics than others.

Therapeutic proteins, or protein-based therapeutics, have proved to be successful treatments against diseases such as diabetes, cancer, infectious diseases, hemophilia, and anemia. It is expected that the global therapeutic proteins market will become $112.17 billion in 2022, according to the Therapeutic Proteins Global Market Report 2022. Based on their molecular types, therapeutic proteins can be grouped into different: antibody-based drugs, Fc fusion proteins, engineered protein scaffolds, enzymes, growth factors, hormones, anticoagulants, blood factors, bone morphogenetic proteins, interferons, interleukins, and thrombolytics [111].

The protein-engineering platform is the "enabling discipline" [111] for protein therapeutics, helping to develop better functionality and targeting, to reduce immunogenicity, and to improve the delivery of therapeutic protein drugs. Since the early 1980s, the development and continued improvement of protein engineering tools have revolutionized the use of therapeutic protein drugs, by modifying existing proteins or creating novel proteins for specific purposes. Though therapeutic monoclonal antibodies (mAbs) are the reigning proteins in the market [112], other therapeutic proteins are proving to be more efficient treatments against several diseases. Working on improving the engineering of proteins for therapeutic applications requires the study of the naturally occurring components in the human immune system to gather knowledge on the structure-to-function relationship of proteins.

The immune system is divided into two parts based on its functionality: innate and adaptive. Both the systems are closely related and work together to protect against pathogens. The adaptive immune system is of major interest for therapeutic applications, given its ability to recognize and remember specific pathogens to generate immunity. The adaptive immune system has two major components, antibodies and the lymphocytes (B and T type lymphocytes). Some potential uses of computational techniques in protein therapeutic design are affinity maturation improving antibody-antigen binding, specificity engineering optimizing protein-protein interactions, and identification of immunologic peptides studying MHC II–antigen interactions [113], among many others. Figure 1 shows the flow of the dissertation and the relationship between the different projects.

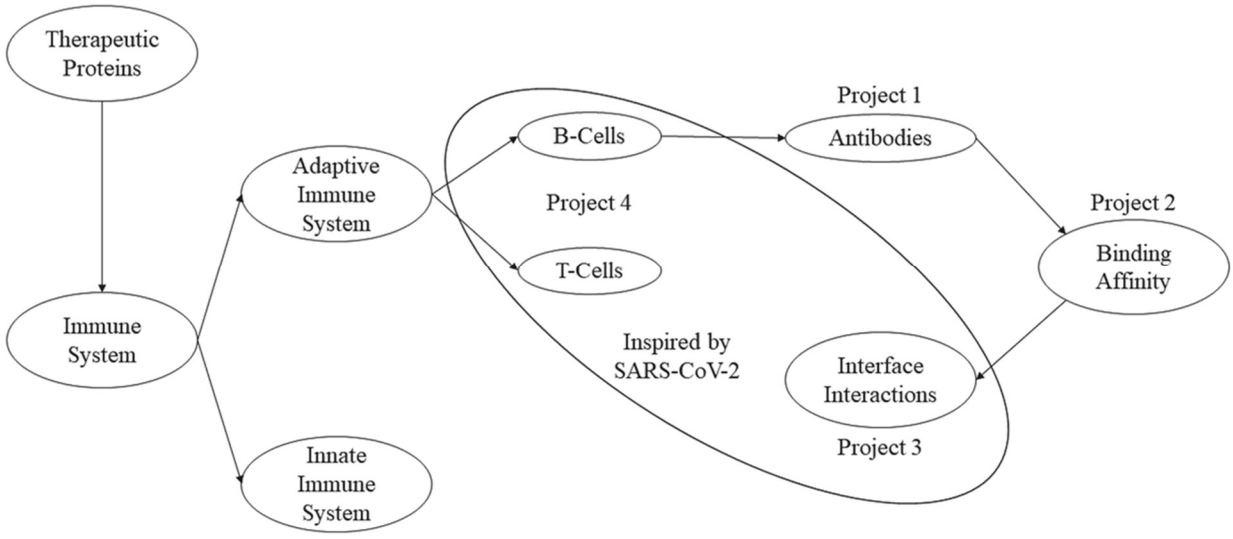


Figure 1: Flowchart to show the arrangement of the projects and their relationship to the development of therapeutic proteins.

References

[1] Jacques. Monod, *Chance and necessity; an essay on the natural philosophy of modern biology. LK - https://auburnuniversity.on.worldcat.org/oclc/1028863*. London SE - 187 pages illustrations 23 cm: Collins, 1972.

[2] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, "Protein function in the post-genomic era," *Nature*, vol. 405, no. 6788, pp. 823–826, 2000, doi: 10.1038/35015694.

[3] J. D. Bloom and F. H. Arnold, "In the light of directed evolution: Pathways of adaptive protein evolution," *Proceedings of the National Academy of Sciences*, vol. 106, no. Supplement 1, p. 9995, Jun. 2009, doi: 10.1073/pnas.0901522106.

[4] E. ZUCKERKANDL and L. PAULING, "Evolutionary Divergence and Convergence in Proteins," in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. Academic Press, 1965, pp. 97–166. doi: https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.

[5] M. KIMURA, "Evolutionary Rate at the Molecular Level," *Nature*, vol. 217, no. 5129, pp. 624–626, 1968, doi: 10.1038/217624a0.

[6] J. L. King and T. H. Jukes, "Non-Darwinian Evolution," *Science (1979)*, vol. 164, no. 3881, pp. 788–798, May 1969, doi: 10.1126/science.164.3881.788.

[7] Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, and H. Zhao, "Directed Evolution: Methodologies and Applications," *Chemical Reviews*, vol. 121, no. 20, pp. 12384–12444, Oct. 2021, doi: 10.1021/acs.chemrev.1c00260.

[8] M. S. Packer and D. R. Liu, "Methods for the directed evolution of proteins," *Nature Reviews Genetics*, vol. 16, no. 7, pp. 379–394, 2015, doi: 10.1038/nrg3927.

[9] B. Turanli-Yildiz and Z. P. C. E.-P. Kaumaya, “Protein Engineering Methods and Applications,” in *Protein Engineering*, C. Alkim, Ed. Rijeka: IntechOpen, 2012, p. Ch. 2. doi: 10.5772/27306.

[10] 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.

[11] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, “The protein folding problem.,” *Annu Rev Biophys*, vol. 37, pp. 289–316, 2008, doi: 10.1146/annurev.biophys.37.092707.153558.

[12] 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.

[13] R. J. Pantazes, M. J. Grisewood, and C. D. Maranas, “Recent advances in computational protein design,” *Current Opinion in Structural Biology*, vol. 21, no. 4, pp. 467–472, 2011, doi: 10.1016/j.sbi.2011.04.005.

[14] B. I. Dahiyat and S. L. Mayo, “De Novo Protein Design: Fully Automated Sequence Selection,” *Science (1979)*, vol. 278, no. 5335, pp. 82 LP – 87, Oct. 1997, doi: 10.1126/science.278.5335.82.

[15] 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 LP – 1368, Nov. 2003, doi: 10.1126/science.1089427.

[16] L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga, “Computational design of receptor and sensor proteins with novel functions,” *Nature*, vol. 423, no. 6936, pp. 185–190, 2003, doi: 10.1038/nature01556.

[17] T. Kortemme and D. Baker, “Computational design of protein–protein interactions,” *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 91–97, 2004, doi: https://doi.org/10.1016/j.cbpa.2003.12.008.

[18] J. Ashworth *et al.*, “Computational redesign of endonuclease DNA binding and cleavage specificity,” *Nature*, vol. 441, no. 7093, pp. 656–659, 2006, doi: 10.1038/nature04818.

[19] L. Regan, “The Design of Metal-Binding Sites in Proteins,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 22, no. 1, pp. 257–281, Jun. 1993, doi: 10.1146/annurev.bb.22.060193.001353.

[20] L. Jiang *et al.*, “De Novo Computational Design of Retro-Aldol Enzymes,” *Science (1979)*, vol. 319, no. 5868, pp. 1387 LP – 1391, Mar. 2008, doi: 10.1126/science.1152692.

[21] D. Röthlisberger *et al.*, “Kemp elimination catalysts by computational enzyme design,” *Nature*, vol. 453, p. 190, Mar. 2008.

[22] L. Regan, D. Caballero, M. R. Hinrichsen, A. Virrueta, D. M. Williams, and C. S. O’Hern, “Protein design: Past, present, and future,” *Biopolymers*, vol. 104, no. 4, pp. 334–350, 2015, doi: 10.1002/bip.22639.

[23] S. J. Fleishman *et al.*, “Community-wide assessment of protein-interface modeling suggests improvements to design methodology,” *Journal of Molecular Biology*, vol. 414, no. 2, pp. 289–302, 2011, doi: 10.1016/j.jmb.2011.09.031.

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

[25] L. Regan, D. Caballero, M. R. Hinrichsen, A. Virrueta, D. M. Williams, and C. S. O’Hern, “Protein design: Past, present, and future,” *Biopolymers*, vol. 104, no. 4, pp. 334–350, 2015, doi: 10.1002/bip.22639.

[26] S. J. Fleishman *et al.*, “Computational design of proteins targeting the conserved stem region of influenza hemagglutinin,” *Science*, vol. 332, no. 6031, pp. 816–821, May 2011, doi: 10.1126/science.1202617.

[27] D. M. Taverna and R. A. Goldstein, “Why are proteins so robust to site mutations?,” *Journal of Molecular Biology*, vol. 315, no. 3, pp. 479–484, 2002, doi: https://doi.org/10.1006/jmbi.2001.5226.

[28] M. Lorch, J. M. Mason, R. B. Sessions, and A. R. Clarke, “Effects of Mutations on the Thermodynamics of a Protein Folding Reaction: Implications for the Mechanism of Formation of the Intermediate and Transition States,” *Biochemistry*, vol. 39, no. 12, pp. 3480–3485, Mar. 2000, doi: 10.1021/bi9923510.

[29] S. Muff and A. Caflisch, “Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein,” *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 4, pp. 1185–1195, Mar. 2008, doi: https://doi.org/10.1002/prot.21565.

[30] I. M. A. Nooren and J. M. Thornton, “Diversity of protein–protein interactions,” *The EMBO Journal*, vol. 22, no. 14, pp. 3486–3492, Jul. 2003, doi: 10.1093/emboj/cdg359.

[31] S. E. Acuner Ozbabacan, H. B. Engin, A. Gursoy, and O. Keskin, “Transient protein–protein interactions,” *Protein Engineering, Design and Selection*, vol. 24, no. 9, pp. 635–648, Jun. 2011, doi: 10.1093/protein/gzr025.

[32] T. L. Blundell *et al.*, “Protein-protein interactions in receptor activation and intracellular signalling,” *Biological Chemistry*, vol. 381, no. 9–10, pp. 955–959, 2000, doi: 10.1515/BC.2000.117.

[33] A. Valencia and F. Pazos, “Computational methods for the prediction of protein interactions,” *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 368–373, 2002, doi: https://doi.org/10.1016/S0959-440X(02)00333-0.

[34] C. J. Layton and H. W. Hellinga, “Quantitation of protein–protein interactions by thermal stability shift analysis,” *Protein Science*, vol. 20, no. 8, pp. 1439–1450, Aug. 2011, doi: https://doi.org/10.1002/pro.674.

[35] D. J. Huggins, M. Marsh, and M. C. Payne, “Thermodynamic Properties of Water Molecules at a Protein–Protein Interaction Surface,” *Journal of Chemical Theory and Computation*, vol. 7, no. 11, pp. 3514–3522, Nov. 2011, doi: 10.1021/ct200465z.

[36] S. E. Acuner Ozbabacan, H. B. Engin, A. Gursoy, and O. Keskin, “Transient protein–protein interactions,” *Protein Engineering, Design and Selection*, vol. 24, no. 9, pp. 635–648, Jun. 2011, doi: 10.1093/protein/gzr025.

[37] T. L. Blundell *et al.*, “Protein-protein interactions in receptor activation and intracellular signalling,” *Biological Chemistry*, vol. 381, no. 9–10, pp. 955–959, 2000, doi: 10.1515/BC.2000.117.

[38] J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, and C. Orengo, “Transient Protein-Protein Interactions: Structural, Functional, and Network Properties,” *Structure*, vol. 18, no. 10, pp. 1233–1243, 2010, doi: https://doi.org/10.1016/j.str.2010.08.007.

[39] H. C. Jubb, A. P. Pandurangan, M. A. Turner, B. Ochoa-Montaño, T. L. Blundell, and D. B. Ascher, “Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health,” *Progress in Biophysics and Molecular Biology*, vol. 128, pp. 3–13, 2017, doi: https://doi.org/10.1016/j.pbiomolbio.2016.10.002.

[40] F. Glaser, D. M. Steinberg, I. a Vakser, N. Ben-tal, and I. E-mail, “Residue Frequencies and Pairing Preferences at Protein – Protein Interfaces protein – protein interfaces of known high-resolu- residue – residue contact preferences . The residue statistical strength of the data set . Differences be- tween amino acid dist,” *Proteins: Structure, Function and Genetics*, vol. 102, no. November 2000, pp. 89–102, 2001.

[41] L. Lo Conte, C. Chothia, and J. Janin, “The atomic structure of protein– protein recognition sites,” *J Mol Biol*, vol. 285, pp. 2177–2198, 1999, doi: 10.1006/jmbi.1998.2439.

[42] J. Meller, “Computational Methods for Prediction of Protein-Protein Interaction Sites,” A. P. E.-W. C. E.-H. Hong, Ed. Rijeka: IntechOpen, 2012, p. Ch. 1. doi: 10.5772/36716.

[43] O. Coskuner-Weber and N. V. Uversky, “Insights into the Molecular Mechanisms of Alzheimer’s and Parkinson’s Diseases with Molecular Simulations: Understanding the Roles of Artificial and Pathological Missense Mutations in Intrinsically Disordered Proteins Related to Pathology,” *International Journal of Molecular Sciences* , vol. 19, no. 2. 2018. doi: 10.3390/ijms19020336.

[44] L. K. Iyer and P. K. Qasba, “Molecular dynamics simulation of α-lactalbumin and calcium binding c-type lysozyme,” *Protein Engineering, Design and Selection*, vol. 12, no. 2, pp. 129–139, Feb. 1999, doi: 10.1093/protein/12.2.129.

[45] T. Kortemme and D. Baker, “Computational design of protein–protein interactions,” *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 91–97, 2004, doi: https://doi.org/10.1016/j.cbpa.2003.12.008.

[46] A. A. Bogan and K. S. Thorn, “Anatomy of hot spots in protein interfaces11Edited by J. Wells,” *Journal of Molecular Biology*, vol. 280, no. 1, pp. 1–9, 1998, doi: https://doi.org/10.1006/jmbi.1998.1843.

[47] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, “Computational alanine scanning mutagenesis—An improved methodological approach,” *Journal of Computational Chemistry*, vol. 28, no. 3, pp. 644–654, Feb. 2007, doi: 10.1002/jcc.20566.

[48] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, “Unraveling the Importance of Protein−Protein Interaction: Application of a Computational Alanine-Scanning Mutagenesis to the Study of the IgG1 Streptococcal Protein G (C2 Fragment) Complex,” *The Journal of Physical Chemistry B*, vol. 110, no. 22, pp. 10962–10969, Jun. 2006, doi: 10.1021/jp054760d.

[49] C. Kristensen *et al.*, “Alanine scanning mutagenesis of insulin,” *Journal of Biological Chemistry*, vol. 272, no. 20, pp. 12978–12983, 1997, doi: 10.1074/jbc.272.20.12978.

[50] K. A. Daggett, M. Layer, and T. A. Cropp, “A General Method for Scanning Unnatural Amino Acid Mutagenesis,” *ACS Chemical Biology*, vol. 4, no. 2, pp. 109–113, Feb. 2009, doi: 10.1021/cb800271f.

[51] T. Kortemme, D. E. Kim, and D. Baker, “Computational Alanine Scanning of Protein-Protein Interfaces,” *Science’s STKE*, vol. 2004, no. 219, p. pl2 LP-pl2, Feb. 2004, doi: 10.1126/stke.2192004pl2.

[52] R. B. Russell *et al.*, “A structural perspective on protein–protein interactions,” *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 313–324, 2004, doi: https://doi.org/10.1016/j.sbi.2004.04.006.

[53] B. Kobe *et al.*, “Crystallography and protein–protein interactions: biological interfaces and crystal contacts,” *Biochemical Society Transactions*, vol. 36, no. 6, pp. 1438 LP – 1441, Dec. 2008, doi: 10.1042/BST0361438.

[54] H. Erlandsen, E. E. Abola, and R. C. Stevens, “Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites,” *Current Opinion in Structural Biology*, vol. 10, no. 6, pp. 719–730, 2000, doi: https://doi.org/10.1016/S0959-440X(00)00154-8.

[55] A. Fernández and H. A. Scheraga, “Insufficiently dehydrated hydrogen bonds as determinants of protein interactions,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 113 LP – 118, Jan. 2003, doi: 10.1073/pnas.0136888100.

[56] S. Schmitt, D. Kuhn, and G. Klebe, “A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology,” *Journal of Molecular Biology*, vol. 323, no. 2, pp. 387–406, 2002, doi: https://doi.org/10.1016/S0022-2836(02)00811-2.

[57] O. Lichtarge, H. R. Bourne, and F. E. Cohen, “An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families,” *Journal of Molecular Biology*, vol. 257, no. 2, pp. 342–358, 1996, doi: https://doi.org/10.1006/jmbi.1996.0167.

[58] G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz, and R. C. Stevens, “Structural Insights into the Evolution of an Antibody Combining Site,” *Science (1979)*, vol. 276, no. 5319, pp. 1665 LP – 1669, Jun. 1997, doi: 10.1126/science.276.5319.1665.

[59] M. H. Kubala, O. Kovtun, K. Alexandrov, and B. M. Collins, “Structural and thermodynamic analysis of the GFP:GFP-nanobody complex,” *Protein Science*, vol. 19, no. 12, pp. 2389–2401, Dec. 2010, doi: 10.1002/pro.519.

[60] S.-H. Chong, J. Hong, S. Lim, S. Cho, J. Lee, and S. Ham, “Structural and Thermodynamic Characteristics of Amyloidogenic Intermediates of β-2-Microglobulin,” *Scientific Reports*, vol. 5, p. 13631, Sep. 2015.

[61] R. A. Mariuzza, D. Immunologie, I. Pasteur, and P. Cedex, “THE STRUCTURAL BASIS OF ANTIGEN-ANTIBODY RECOGNITION Poljak,” 1987.

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

[63] W. E. Stites, “Protein-protein interactions: Interface structure, binding thermodynamics, and mutational analysis,” *Chemical Reviews*, vol. 97, no. 5, pp. 1233–1250, 1997, doi: 10.1021/cr960387h.

[64] R. E. S. Harrison, R. R. Mohan, R. D. Gorham, C. A. Kieslich, and D. Morikis, “AESOP: A Python Library for Investigating Electrostatics in Protein Interactions,” *Biophysical Journal*, vol. 112, no. 9, pp. 1761–1766, 2017, doi: https://doi.org/10.1016/j.bpj.2017.04.005.

[65] C.-J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, "Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect," *Protein Science*, vol. 6, no. 1, pp. 53–64, Jan. 1997, doi: 10.1002/pro.5560060106.

[66] W. T. M. Mooij and M. L. Verdonk, "General and targeted statistical potentials for protein–ligand interactions," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 2, pp. 272–287, Nov. 2005, doi: 10.1002/prot.20588.

[67] N. Tokuriki and D. S. Tawfik, "Protein Dynamism and Evolvability," *Science (1979)*, vol. 324, no. 5924, pp. 203 LP – 207, Apr. 2009, doi: 10.1126/science.1169375.

[68] I. H. Moal and J. Fernández-Recio, "SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models," *Bioinformatics*, vol. 28, no. 20, pp. 2600–2607, Aug. 2012, doi: 10.1093/bioinformatics/bts489.

[69] R. Moretti *et al.*, "Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions," *Proteins: Structure, Function, and Bioinformatics*, vol. 81, no. 11, pp. 1980–1987, Nov. 2013, doi: 10.1002/prot.24356.

[70] V. Parthiban, M. M. Gromiha, and D. Schomburg, "CUPSAT: prediction of protein stability upon point mutations," *Nucleic Acids Research*, vol. 34, no. suppl_2, pp. W239–W242, Jul. 2006, doi: 10.1093/nar/gkl190.

[71] N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik, "The Stability Effects of Protein Mutations Appear to be Universally Distributed," *Journal of Molecular Biology*, vol. 369, no. 5, pp. 1318–1332, 2007, doi: https://doi.org/10.1016/j.jmb.2007.03.069.

[72] N. Berliner, J. Teyra, R. Çolak, S. Garcia Lopez, and P. M. Kim, "Combining Structural Modeling with Ensemble Machine Learning to Accurately Predict Protein Fold Stability and Binding Affinity Effects upon Mutation," *PLOS ONE*, vol. 9, no. 9, p. e107353, Sep. 2014.

[73] D. E. V Pires, D. B. Ascher, and T. L. Blundell, "mCSM: predicting the effects of mutations in proteins using graph-based signatures," *Bioinformatics*, vol. 30, no. 3, pp. 335–342, Nov. 2013, doi: 10.1093/bioinformatics/btt691.

[74] Y. Dehouck, J. M. Kwasigroch, M. Rooman, and D. Gilis, "BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations," *Nucleic Acids Research*, vol. 41, no. W1, pp. W333–W339, May 2013, doi: 10.1093/nar/gkt450.

[75] I. C. M. Simões, I. P. D. Costa, J. T. S. Coimbra, M. J. Ramos, and P. A. Fernandes, "New Parameters for Higher Accuracy in the Computation of Binding Free Energy Differences upon Alanine Scanning Mutagenesis on Protein–Protein Interfaces," *Journal of Chemical Information and Modeling*, vol. 57, no. 1, pp. 60–72, Jan. 2017, doi: 10.1021/acs.jcim.6b00378.

[76] M. Li, M. Petukh, E. Alexov, and A. R. Panchenko, "Predicting the Impact of Missense Mutations on Protein–Protein Binding Affinity," *Journal of Chemical Theory and Computation*, vol. 10, no. 4, pp. 1770–1780, Apr. 2014, doi: 10.1021/ct401022c.

[77] C. Geng, L. C. Xue, J. Roel-Touris, and A. M. J. J. Bonvin, "Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?," *WIREs Computational Molecular Science*, vol. 9, no. 5, p. e1410, Sep. 2019, doi: 10.1002/wcms.1410.

[78] T. N. Starr *et al.*, “Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding,” *Cell*, vol. 182, no. 5, pp. 1295-1310.e20, 2020, doi: https://doi.org/10.1016/j.cell.2020.08.012.

[79] W. F. Garcia-Beltran *et al.*, “Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity,” *Cell*, vol. 184, no. 9, pp. 2372-2383.e9, Apr. 2021, doi: 10.1016/j.cell.2021.03.013.

[80] G. S. M.-C. *et al.*, “Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity,” *Science (1979)*, vol. 373, no. 6555, p. eabi6226, Dec. 2021, doi: 10.1126/science.abi6226.

[81] D. Wrapp *et al.*, “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation,” *Science (1979)*, p. eabb2507, Feb. 2020, doi: 10.1126/science.abb2507.

[82] P. Prabakaran *et al.*, “Structure of Severe Acute Respiratory Syndrome Coronavirus Receptor-binding Domain Complexed with Neutralizing Antibody,” *Journal of Biological Chemistry* , vol. 281, no. 23, pp. 15829–15836, Jun. 2006, doi: 10.1074/jbc.M600697200 .

[83] Z. Zhu *et al.*, “Potent cross-reactive neutralization of SARS coronavirus isolates by human monoclonal antibodies,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 29, pp. 12123 LP – 12128, Jul. 2007, doi: 10.1073/pnas.0701000104.

[84] W. C. Hwang *et al.*, “Structural Basis of Neutralization by a Human Anti-severe Acute Respiratory Syndrome Spike Protein Antibody, 80R,” *Journal of Biological Chemistry* , vol. 281, no. 45, pp. 34610–34616, Nov. 2006, doi: 10.1074/jbc.M603275200 .

[85] S. Xia *et al.*, "Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBIBP-CorV: a randomised, double-blind, placebo-controlled, phase 1/2 trial," *The Lancet Infectious Diseases*, vol. 21, no. 1, pp. 39–51, 2021, doi: https://doi.org/10.1016/S1473-3099(20)30831-8.

[86] F. P. Polack *et al.*, "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine," *New England Journal of Medicine*, vol. 383, no. 27, pp. 2603–2615, Dec. 2020, doi: 10.1056/NEJMoa2034577.

[87] L. R. Baden *et al.*, "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine," *New England Journal of Medicine*, vol. 384, no. 5, pp. 403–416, Dec. 2020, doi: 10.1056/NEJMoa2035389.

[88] J. Sadoff *et al.*, "Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19," *New England Journal of Medicine*, vol. 384, no. 23, pp. 2187–2201, Apr. 2021, doi: 10.1056/NEJMoa2101544.

[89] M. D. Knoll and C. Wonodi, "Oxford–AstraZeneca COVID-19 vaccine efficacy," *The Lancet*, vol. 397, no. 10269, pp. 72–74, 2021, doi: https://doi.org/10.1016/S0140-6736(20)32623-4.

[90] I. Jones and P. Roy, "Sputnik V COVID-19 vaccine candidate appears safe and effective," *The Lancet*, vol. 397, no. 10275, pp. 642–643, 2021, doi: https://doi.org/10.1016/S0140-6736(21)00191-4.

[91] P. T. Heath *et al.*, "Safety and Efficacy of NVX-CoV2373 Covid-19 Vaccine," *New England Journal of Medicine*, vol. 385, no. 13, pp. 1172–1183, Jun. 2021, doi: 10.1056/NEJMoa2107659.

[92] J. Lopez Bernal *et al.*, "Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant," *New England Journal of Medicine*, vol. 385, no. 7, pp. 585–594, Jul. 2021, doi: 10.1056/NEJMoa2108891.

[93] E. Dolgin, "THE TANGLED HISTORY OF MRNA VACCINES Hundreds of scientists had worked on mRNA vaccines for decades before the coronavirus pandemic brought a breakthrough," 2021.

[94] Y. Cao *et al.*, "Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells," *Cell*, vol. 182, no. 1, pp. 73-84.e16, Jul. 2020, doi: 10.1016/j.cell.2020.05.025.

[95] N. Yuchun *et al.*, "Neutralizing Antibodies in Patients with Severe Acute Respiratory Syndrome-Associated Coronavirus Infection," *The Journal of Infectious Diseases*, vol. 190, no. 6, pp. 1119–1126, Sep. 2004, doi: 10.1086/423286.

[96] J. E. Crowe, R. O. Suara, S. Brock, N. Kallewaard, F. House, and J.-H. Weitkamp, "Genetic and structural determinants of virus neutralizing antibodies," *Immunologic Research*, vol. 23, no. 2, pp. 135–145, 2001, doi: 10.1385/IR:23:2-3:135.

[97] M. A. Tortorici *et al.*, "Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms," *Science*, vol. 370, no. 6519, pp. 950–957, Nov. 2020, doi: 10.1126/science.abe3354.

[98] J. Kreye *et al.*, "A Therapeutic Non-self-reactive SARS-CoV-2 Antibody Protects from Lung Pathology in a COVID-19 Hamster Model," *Cell*, vol. 183, no. 4, pp. 1058-1069.e19, Nov. 2020, doi: 10.1016/j.cell.2020.09.049.

[99] L. Piccoli *et al.*, "Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology," *Cell*, vol. 183, no. 4, pp. 1024-1042.e21, Nov. 2020, doi: 10.1016/j.cell.2020.09.037.

[100] W. Dejnirattisai *et al.*, "The antigenic anatomy of SARS-CoV-2 receptor binding domain," *Cell*, vol. 184, no. 8, pp. 2183-2200.e22, Apr. 2021, doi: 10.1016/j.cell.2021.02.032.

[101] J. Wan *et al.*, "Human-IgG-Neutralizing Monoclonal Antibodies Block the SARS-CoV-2 Infection," *Cell Rep*, vol. 32, no. 3, p. 107918, Jul. 2020, doi: 10.1016/j.celrep.2020.107918.

[102] M. Schoof *et al.*, "An ultrapotent synthetic nanobody neutralizes SARS-CoV-2 by stabilizing inactive Spike," *Science*, vol. 370, no. 6523, pp. 1473–1479, Dec. 2020, doi: 10.1126/science.abe3255.

[103] J. Huo *et al.*, "Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2," *Nature Structural & Molecular Biology*, vol. 27, no. 9, pp. 846–854, 2020, doi: 10.1038/s41594-020-0469-6.

[104] T. F. Custódio *et al.*, "Selection, biophysical and structural analysis of synthetic nanobodies that effectively neutralize SARS-CoV-2," *Nat Commun*, vol. 11, no. 1, p. 5588, Nov. 2020, doi: 10.1038/s41467-020-19204-y.

[105] L. Hanke *et al.*, "An alpaca nanobody neutralizes SARS-CoV-2 by blocking receptor interaction," *Nat Commun*, vol. 11, no. 1, p. 4420, Sep. 2020, doi: 10.1038/s41467-020-18174-5.

[106] P.-A. Koenig *et al.*, "Structure-guided multivalent nanobodies block SARS-CoV-2 infection and suppress mutational escape," *Science*, vol. 371, no. 6530, p. eabe6230, Feb. 2021, doi: 10.1126/science.abe6230.

[107] N. Wang *et al.*, "Structure-based development of human antibody cocktails against SARS-CoV-2," *Cell Res*, vol. 31, no. 1, pp. 101–103, Jan. 2021, doi: 10.1038/s41422-020-00446-w.

[108] H. Yao *et al.*, "Rational development of a human antibody cocktail that deploys multiple functions to confer Pan-SARS-CoVs protection," *Cell Res*, vol. 31, no. 1, pp. 25–36, Jan. 2021, doi: 10.1038/s41422-020-00444-y.

[109] J. Hansen *et al.*, "Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail," *Science*, vol. 369, no. 6506, pp. 1010–1014, Aug. 2020, doi: 10.1126/science.abd0827.

[110] T. N. Starr, A. J. Greaney, A. S. Dingens, and J. D. Bloom, "Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016," *Cell Rep Med*, vol. 2, no. 4, p. 100255, Apr. 2021, doi: 10.1016/j.xcrm.2021.100255.

[111] P. J. Carter, "Introduction to current and future protein therapeutics: A protein engineering perspective," *Experimental Cell Research*, vol. 317, no. 9, pp. 1261–1269, 2011, doi: https://doi.org/10.1016/j.yexcr.2011.02.013.

[112] G. Walsh, "Biopharmaceutical benchmarks 2018," *Nature Biotechnology*, vol. 36, no. 12, pp. 1136–1145, 2018, doi: 10.1038/nbt.4305.

[113] I. Hwang and S. Park, “Computational design of protein therapeutics,” *Drug Discovery Today: Technologies*, vol. 5, no. 2, pp. e43–e48, 2008, doi: https://doi.org/10.1016/j.ddtec.2008.11.004.

Chapter 2 Development and Analyses of a Database of Antibody – Antigen Complexes

Introduction

To develop an understanding of the structure to function relationship of proteins, gathering knowledge about both the physical and chemical properties of the binding site is needed [1]. The properties to be studied are the number of residues required to describe interactions efficiently, the location of the interaction sites, the types of residues present at the interaction sites, and the linearity of the epitopes. This study is focused on antibody-protein interactions, given that antibodies bind with high affinity and specificity, and they are a class of binding proteins that have been extensively studied previously. This gives us more data to validate our results and ensure the usability of the workflow for other protein classes. Antigenic epitopes can be either linear (or sequential), made from a single length of continuous residues of the polypeptide chain, or conformational (or structural), made from several discrete residues that are widely spread in the primary sequence coming together on the surface during protein folding. Structural epitopes are known to be more common than linear epitopes [2], [3]. Here, the CHARMM22 [4], [5], Amber [6], and Rosetta [7], [8] molecular mechanics force fields were used to quantify the percent contribution to binding of every residue and identify the amino acid composition of the residues in the exposed parts of the antigen surface (the epitope) based on their binding contributions in a database of 384 antibody-protein complexes.

Methods

Selection of Complexes for the Database

For the creation of a non-redundant database, initially, 2498 antibody-antigen experimentally determined structures were collected from the International Immunogenetics (IMGT) Information System 3D Structure Database [9]. Of these, 1344 structures were selected based on having at least five mutations from one another in the complementary determining

regions of the antibodies, making the list a non-redundant one based on antibody structures. Out of the 1344 non-redundant structures, 492 structures were identified to have protein antigens. To remove the effects of small antigens biasing results in the calculations, only protein antigens of 50 or more amino acids were considered. Eventually, a non-redundant list consisting of 384 structures of antibody-protein complexes was made.

Analysis of the Interfaces in the Database

In each force field used, missing atoms were added to the structures and the overall energies of each of these complexes were minimized. The energy of interactions between every antibody residue and every antigen residue was calculated using pairwise additive energy functions, and the total interaction energy for each antigenic residue was summed. The total binding energy varied significantly between complexes and depended on the total number of amino acids residues. To better facilitate the analysis of the different complexes, the energies of each residue in a complex were converted to the percentage of the total energy of the complex.

Using the percentage of total binding energy contributed by each antigenic residue, they were ranked in descending order and the top seven residues were chosen as the most significant. It was analyzed and estimated that seven residues contributed around 70% of the total binding energy in all the force fields. The top seven residues were treated as the epitope of the antigen and further analyses were conducted accordingly. The amino acid distribution was analyzed for the epitope residues versus the distribution for all antigenic residues. In addition, the linearity of the epitopes was analyzed by fitting the epitope residues into a twelve-residue frame and calculating the percentage of binding energy that could be contributed by each frame.

Results and Discussions

Experimental studies have found that in most complexes only a few antigen residues (on average five) are sufficient for specific and high-affinity binding [10], [11]. The analyses reflect the same for the non-redundant database. The seven most attractive antigenic residues contribute an average of 66.2%, 67.9%, and 73.5% of the total CHARMM, Amber, and Rosetta binding energies, respectively. The seven most attractive antibody residues contribute an average of 72.5%, 68.7%, and 85.5% of the total CHARMM, Amber, and Rosetta binding energies, respectively. That means only seven residues were enough to describe around 70% of the total binding energy for both antigens and antibodies, making the remainder of the peptide sequence in both cases almost insignificant to the binding. Figures 1 and 2 show the logarithmic progression in the cumulative percentage contribution to binding energy for the antigenic and antibody residues respectively. The trend may be described as an exponential decrease in the contribution to binding energy by consecutive residues. The sequential analyses made for the database used the top seven residues in each antigen and in each antibody. Confirming that seven residues were enough to represent the epitope, the amino acid distribution for the top seven residues versus the antigen as a whole was explored.

Figures 3 and 4 show the probabilities of the amino acids as they are found in the protein antigens and antibodies, respectively. The probabilities for Cysteine and Methionine are lower than the rest, as expected; sulfur-containing amino acids are specific to their functions in proteins. Methionine is one of the most hydrophobic residues and its primary function is the initiation of translation of proteins rather than protein structure [12]. Cysteine is critical in protein structure with its ability to form disulfide bonds with other Cysteine residues and is found mostly in the hydrophobic core of the protein structure. A similar distribution analysis was

conducted for the epitopes, considering the seven top binding residues for the protein complex. The trends show charged and acidic residues have much higher probabilities to be present on the interface than most other residues. Analysis of PPIs provides strong evidence of a distinctive composition of residues [13], [14]. Figures 5 and 6 depicts the Log_2 change in the relative usage preferences of amino acids in the significant residues versus the protein as a whole. The trends are mostly consistent for all the force fields used: charged and acidic residues are overrepresented in the significant residues while hydrophobic residues are underrepresented, and most of the polar residues are significant based on their usage in antigens. Arginine is significant to protein binding, and this has been depicted well in the figure, as Arginine is one of the top three most used residues according to all three force fields. Previous analyses [13], [15] of protein interfaces have found a higher prevalence of Arginine, owing to their capability of forming multiple hydrogen bonds. Studies [16] – [18] have found hydrophobic residues are preferred in the hydrophobic core while hydrophilic residues are preferred in the exterior of the proteins, corresponding to Figure 6. Protein folding is mainly driven by the "hydrophobic collapse" brings the hydrophobic side chains compacted into the core structure to make the process as energetically favorable as possible: maximizing van der Waals (VDWs) forces, avoiding unfavorable atomic overlaps, and minimizing the overall size occupied by the protein structure [19]. The balance between the interactions of the hydrophobic residues in the packed core, also known as the hydrophobic effect, and exterior interactions of the polar residues that provide structural specificity and solvation maintains the folded structure of the protein in design [20]. Single aromatic residues were preferred in some protein interactions[16], [21], [22], which means residues capable of multiple favorable interactions were preferred. Tryptophan is capable of aromatic p-interactions, hydrogen bonding, and hydrophobic effect contributions. Similarly,

tyrosine can contribute with a hydrophobic surface, aromatic p-interactions, and hydrogen bonding.

Understanding the degree of linearity of the epitopes is an important part of understanding how sequence and structure contribute to interactions. Instead of using binary definitions of linear and structural, it is preferable to quantify the degree of linearity of epitopes accepting that all structural epitopes are linear to varying extents. Figure 7 shows the fraction of the threshold binding energy reached by the best consecutive 12 amino acid sequence in the antigens. 72 % of epitopes were at least 50 % linear, 23 % were at least 80 % linear and 5 % were at least 100 % linear in the Amber results. The corresponding CHARMM values were 67 %, 27 % and 6 %, respectively, while the corresponding Rosetta values were 67 %, 26 % and 7 % respectively. The linearity study of antigen epitopes has the potential to extend into future projects. One such project could be the creation of mimitopes to replace the usage of antigen proteins for therapeutic purposes. Understanding the linearity of epitopes is specifically needed for biomarker detection through peptide display systems.

This study was a step towards identifying the features of the binding interfaces of antibody-antigen complexes. Identifying the features is a necessary step in deciding how traditional approaches towards modifications of these interfaces can be more effective. The findings from this study are used in the consecutive chapters of this dissertation. This chapter has been published as Chauhan et al. [23].

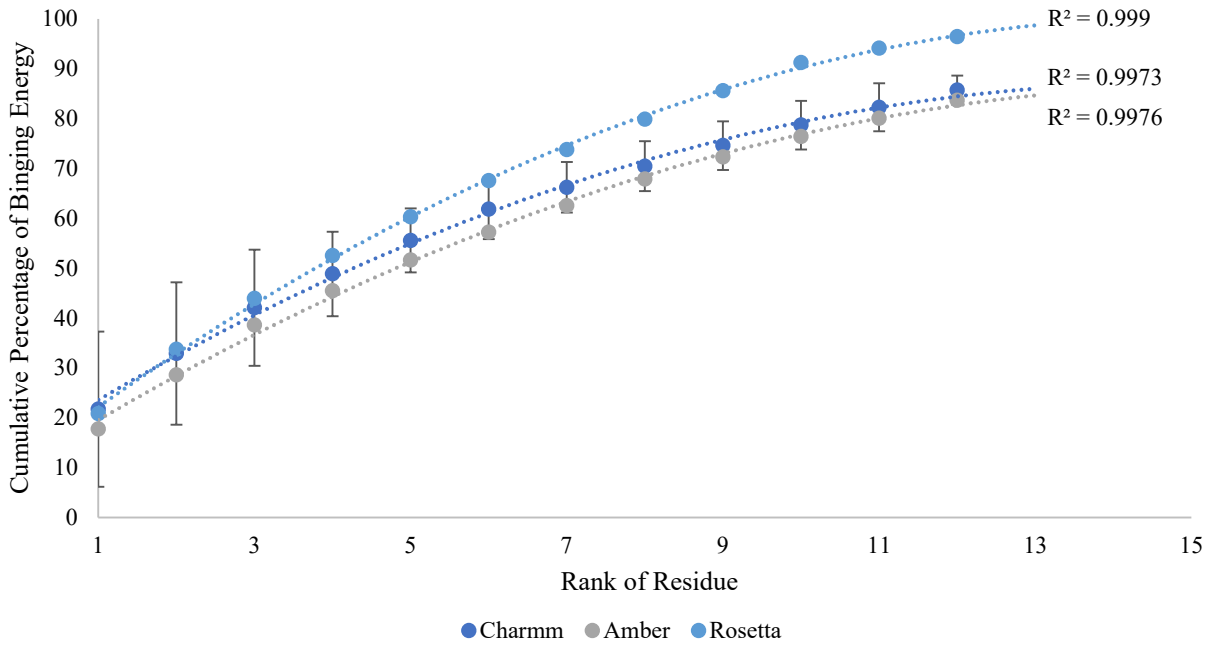


Figure 2:Cumulative percentage binding energies of the antigen residues.

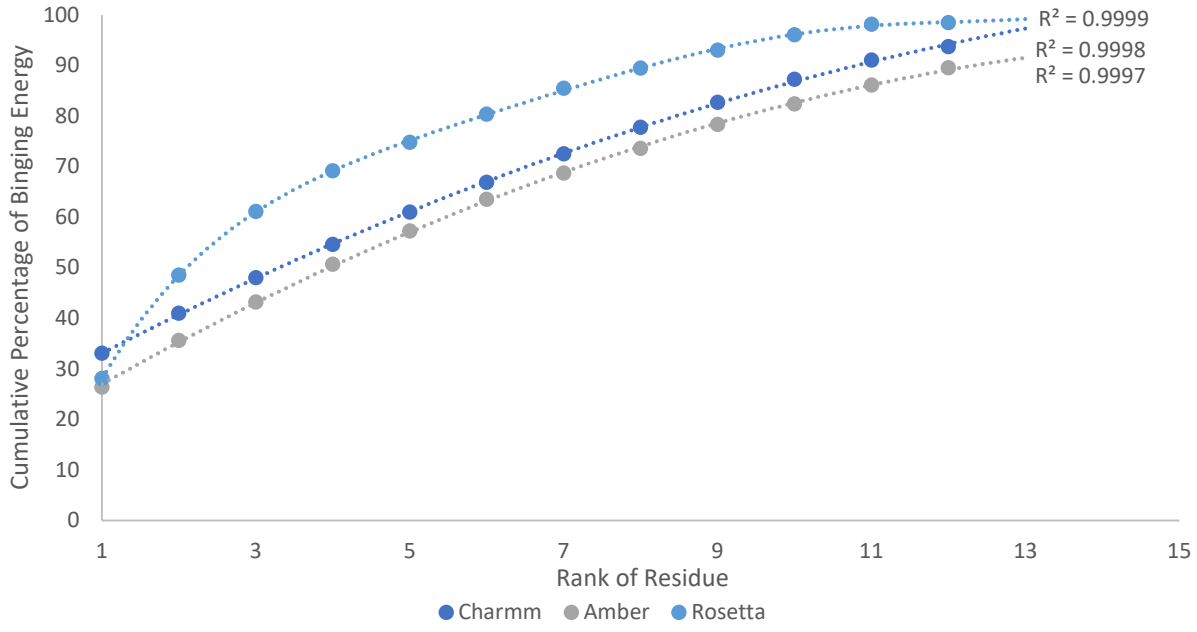


Figure 3:Cumulative percentage binding energies of the antibody residues.

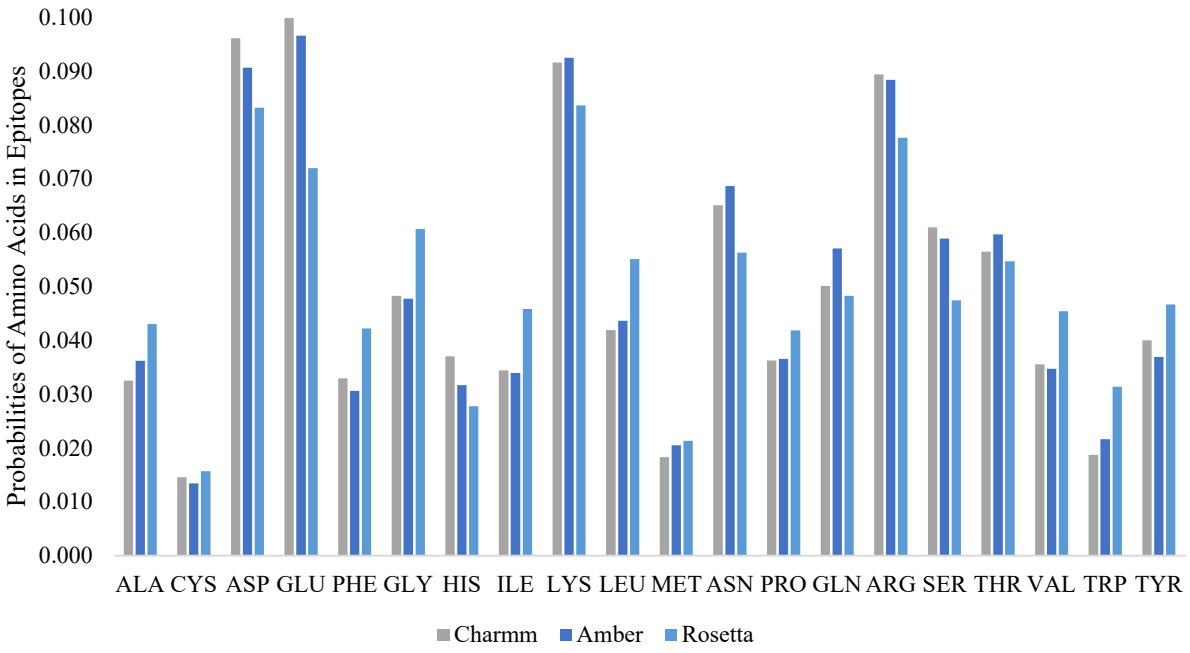


Figure 5: Distribution of amino acids in the epitopes of the antigens. As seen in the plot, Aspartic Acid, Glutamic Acid, Lysine, and Arginine are most preferred on the epitopes.

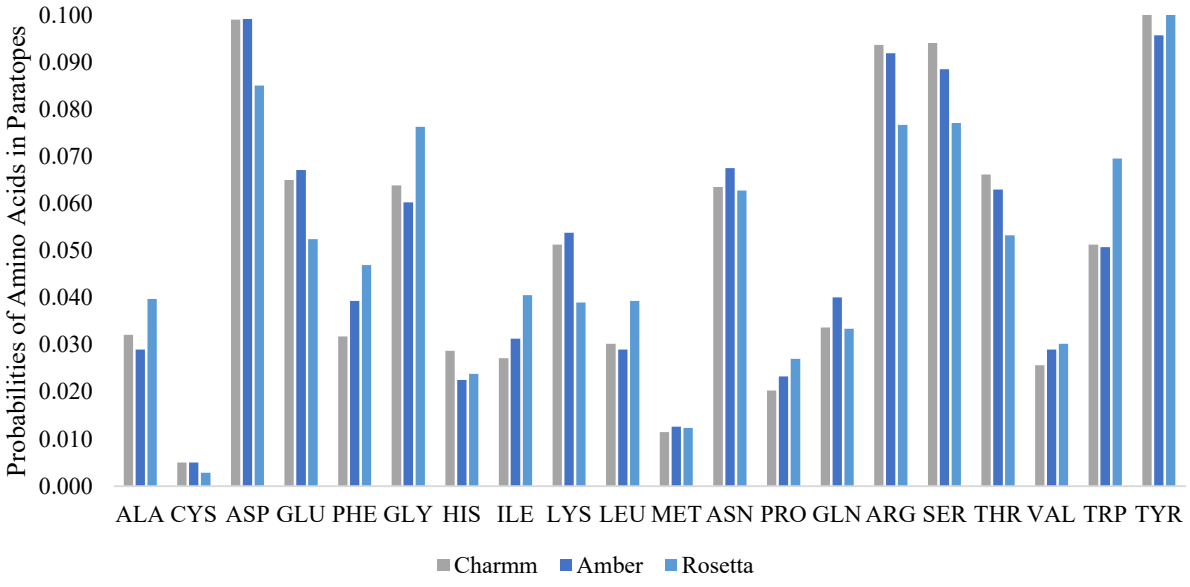


Figure 4: Distribution of amino acids in the paratopes of the antibodies. As seen in the plot, Aspartic Acid, Arginine, Serine, and Tyrosine are most preferred on the paratopes.

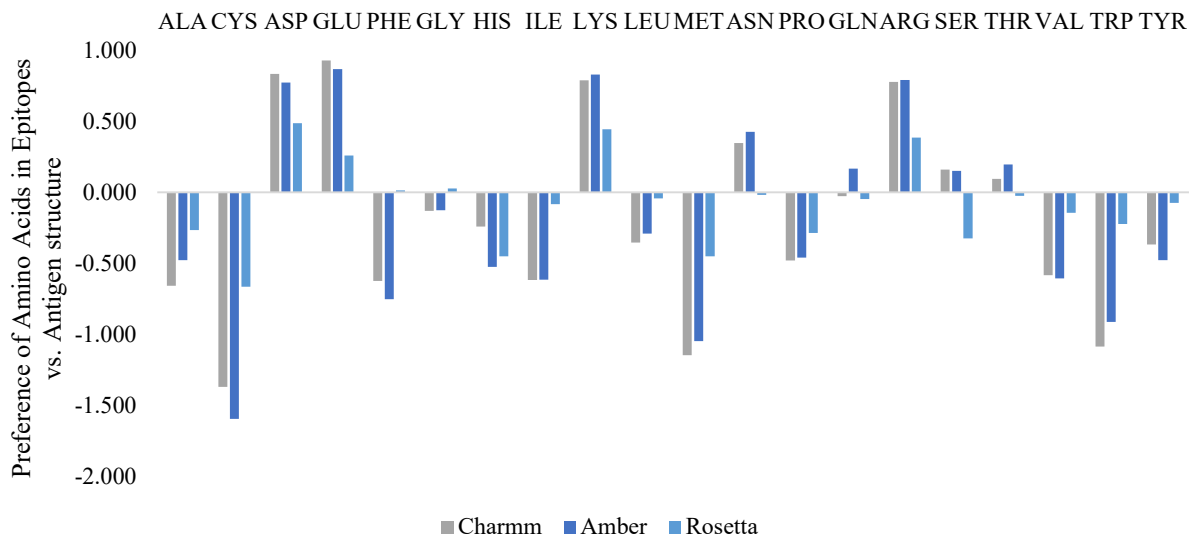


Figure 7: Log2 change in amino acid usage in the significant residues versus the antigen as a whole. This plot shows the residues that are preferred for intermolecular binding in contrast to the ones that are preferred for intramolecular binding.

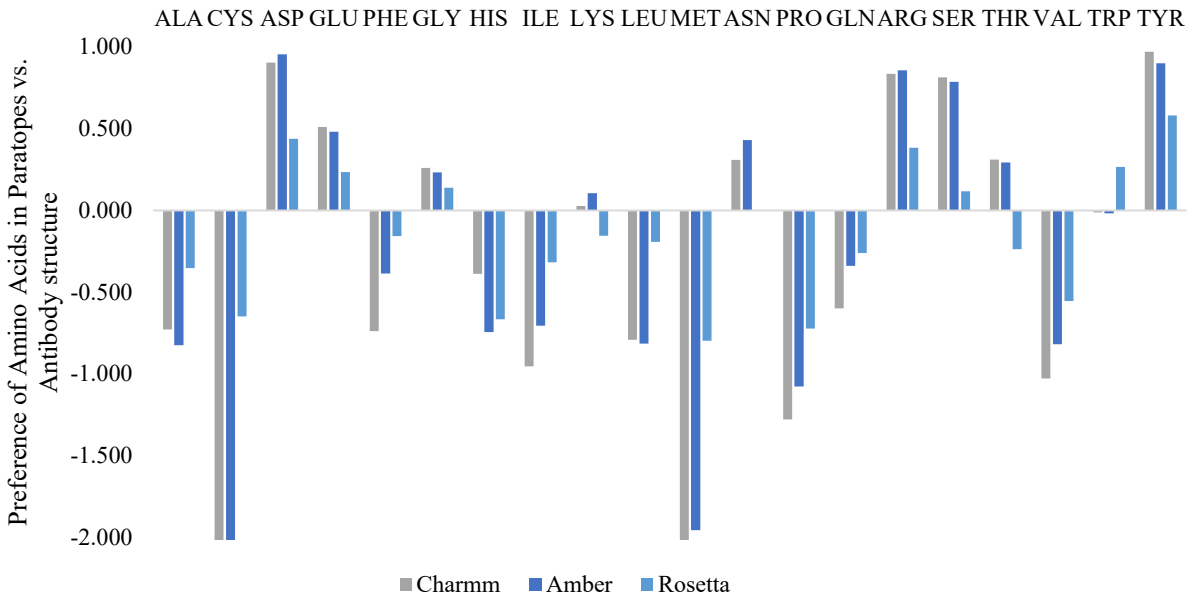


Figure 6: Log2 change in amino acid usage in the significant residues versus the antibody as a whole. This plot shows the residues that are preferred for intermolecular binding in contrast to the ones that are preferred for intramolecular binding.

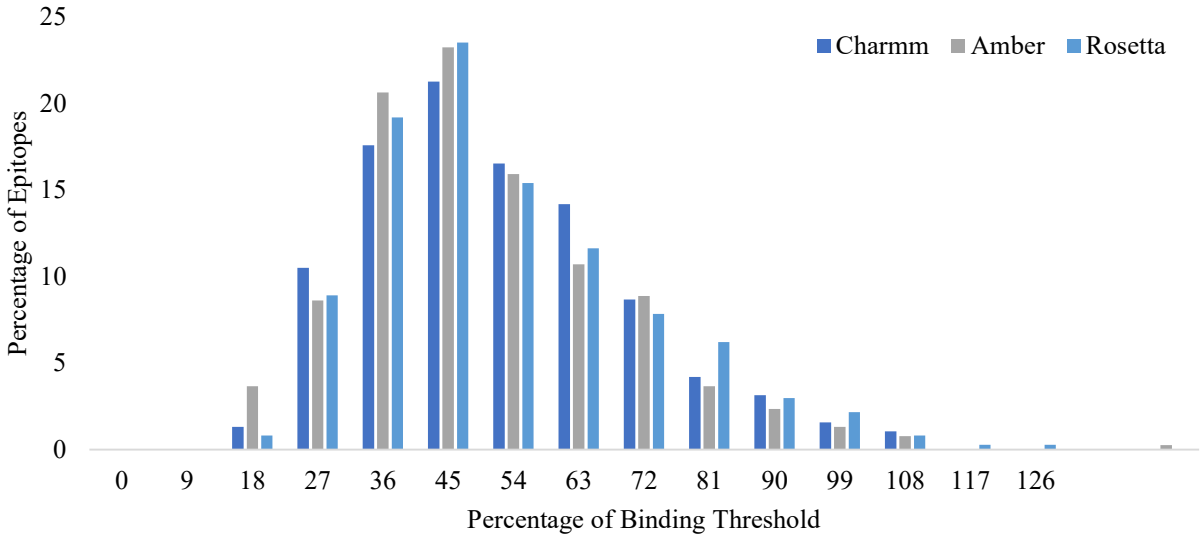


Figure 8: Percentage of binding threshold in the best 12 consecutive antigen amino acid

References

[1] T. A. Halgren, "Identifying and Characterizing Binding Sites and Assessing Druggability," *Journal of Chemical Information and Modeling*, vol. 49, no. 2, pp. 377–389, Feb. 2009, doi: 10.1021/ci800324m.

[2] W. G. Laver, G. M. Air, R. G. Webster, and S. J. Smith-Gill, "Epitopes on protein antigens: Misconceptions and realities," *Cell*, vol. 61, no. 4, pp. 553–556, 1990, doi: 10.1016/0092-8674(90)90464-P.

[3] W.-L. Chang, Y.-T. Lo, and T.-W. Pai, *A Conformational Epitope Prediction System Based on Sequence and Structural Characteristics*, vol. 9799. 2016. doi: 10.1007/978-3-319-42007-3_41.

[4] B. R. Brooks *et al.*, "CHARMM: the biomolecular simulation program," *Journal of computational chemistry*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[5] X. Zhu, P. E. M. Lopes, and A. D. Mackerell, "Recent developments and applications of the CHARMM force fields," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 1, pp. 167–185, 2012, doi: 10.1002/wcms.74.

[6] D. A. Case *et al.*, "AMBER 2017." University of California, San Francisco, 2017.

[7] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," *Methods in Enzymology*, vol. 383, no. 2003, pp. 66–93, 2004, doi: 10.1016/S0076-6879(04)83004-0.

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

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

[10] K. F. Sykes, J. B. Legutki, and P. Stafford, “Immunosignaturing: A critical review,” *Trends in Biotechnology*, vol. 31, no. 1, pp. 45–51, 2013, doi: 10.1016/j.tibtech.2012.10.012.

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

[12] J. T. Brosnan and M. E. Brosnan, “The Sulfur-Containing Amino Acids: An Overview,” *The Journal of Nutrition*, vol. 136, no. 6, pp. 1636S-1640S, Jun. 2006, doi: 10.1093/jn/136.6.1636S.

[13] A. A. Bogan and K. S. Thorn, “Anatomy of hot spots in protein interfaces,” *Journal of Molecular Biology*, vol. 280, no. 1, pp. 1–9, 1998, doi: 10.1006/jmbi.1998.1843.

[14] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, “Hot spots—A review of the protein–protein interface determinant amino-acid residues,” *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 803–812, Sep. 2007, doi: 10.1002/prot.21396.

[15] S. Jones and J. M. Thornton, “Protein-protein interactions: A review of protein dimer structures,” *Progress in Biophysics and Molecular Biology*, vol. 63, no. 1, pp. 31–65, 1995, doi: https://doi.org/10.1016/0079-6107(94)00008-W.

[16] C.-J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, “Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect,” *Protein Science*, vol. 6, no. 1, pp. 53–64, Jan. 1997, doi: 10.1002/pro.5560060106.

[17] L. Young, R. L. Jernigan, and D. G. Covell, “A role for surface hydrophobicity in protein-protein recognition,” *Protein Science*, vol. 3, no. 5, pp. 717–729, May 1994, doi: 10.1002/pro.5560030501.

[18] A. P. Korn and R. M. Burnett, “Distribution and complementarity of hydropathy in mutisunit proteins,” *Proteins: Structure, Function, and Bioinformatics*, vol. 9, no. 1, pp. 37–55, Jan. 1991, doi: 10.1002/prot.340090106.

[19] 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.

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

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

[22] E. A. Padlan, “On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands,” *Proteins: Structure, Function, and Bioinformatics*, vol. 7, no. 2, pp. 112–124, Jan. 1990, doi: 10.1002/prot.340070203.

[23] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” *Computer Aided Chemical Engineering*, vol. 44, pp. 2113–2118, Jan. 2018, doi: 10.1016/B978-0-444-64241-7.50347-5.

Chapter 3 Development of Similarity Matrices based on Protein Binding Interactions

Introduction

In search of a tool to quantify the effect of a point mutation of protein binding, it was found that existing matrices such as Point Accepted Mutation (PAM) and Blocks Substitution Matrix (BLOSUM) were not appropriate for scoring protein sequences based on their binding affinities. These matrices were created based on accumulated statistical data of how tolerated evolutionary amino acid mutations are in multiple sequence alignments in proteins. Thus, a scoring matrix for protein binding was a necessity for protein engineering in general. A single mutation of the interface residues can influence binding affinity by changing the chemical properties and the conformation of the interface [1]. Comprehending the specific types of residues having more involvement in the level of affinity and specificity of PPIs is paramount to understanding these interactions and designing proteins. Studying the effects of point mutations on protein binding can give us a clear perception of the forces and the recognition processes at a molecular level guiding these interactions [2] – [9]. Antibodies interfaces show preference for different amino acids like Arginine, Aspartic Acid, Tyrosine, and Serine [10] – [13]. Antigens are not known to show the same preference for amino acids. This may be attributed to the evolutionary history of antibodies, i.e., they show preferences to increase their binding affinity to antigens. Antigens do not have high specificity requirements to bind to antibodies and are continuously evolving, thus they do not show any preference to any specific amino acids.

Computational techniques used for affinity maturation require force fields that can be used to predict the accurate mutated structure of the protein. Different molecular mechanics force fields use different approaches to calculate the energy for a protein complex and using different force fields will help to identify the limitations of each force field. Additive force fields like CHARMM and Amber are initially based on the first principle and use calculated and

observed potentials to describe the structure of the proteins [17]. Force fields like Rosetta use observed potentials to describe individual residue environments and residue interactions [18].

In this project, statistical analyses are used to quantitatively study the effects of point mutations on protein binding interfaces and present the data as a grid to be used as a similarity matrix based on protein binding. Analyses from the non-redundant database in the previous chapter lead to the fact that around seven residues are enough to describe binding energy efficiently. To generate such a matrix, CHARMM [14], Amber [15], and Rosetta [16], [17] molecular mechanics force fields were used to calculate predicted changes in binding energy from mutations to the most important antibody and antigen amino acids. The similarity matrices for the different force fields were constructed from numerical data describing the effects on binding due to mutations of the significant residues in protein interfaces, thus, providing numerical scores to these effects of amino acid mutations on binding interactions. Binding energy is calculated as $\Delta G = G_{BA,\,min} - G_{DMs} - G_{TMs}$ [18], where a binding assembly (BA) is the entire complex of all the participating molecules, design molecules (DMs) are the mutated molecules, and target molecules (TMs) are the molecules binding to the DMs. The change in binding free energy ($\Delta\Delta G_{bind}$) is quantified using $\Delta\Delta G_{bind} = \Delta G_{MT_{bind}} - \Delta G_{WT_{bind}}$ [19], where MT is the mutated-type complex and WT is the wild-type complex.

Methods

The following steps were taken for the selected antigen residues for each force field and then for the selected antibody residues for each force field.

Mutational Analysis of the Selected Residues

The algorithm in Figure 8 was developed and executed to create mutations and analyze the data.

Step 1: The wild-type complex is minimized, and the interaction energy of the wild-type complex was calculated (ΔG_{WT}).

Step 2: Using the database analysis results, the seven important residues, based on their contribution to binding energies, were selected.

Since each of the three force fields predicts the total energy of the system using a different method, the residues included in the seven significant residues (paratopes for antibodies and epitopes for antigens) are different for the three force fields. For each of the selected residues,

Step 3A: 19 different mutations were made using a complete rotamer library of all the different amino acids. The rotamer library has the collection of all possible rotational isomers of each amino acid.

Step 3B: For each of the 19 mutations, the structure was minimized, and the interaction energy of the mutated-type complex was calculated (ΔG_{MT}).

Step 3C: The change in binding energy ($\Delta\Delta G_{bind}$) is calculated for that mutation. The percentage change in binding energy ($\Delta\Delta G_{bind} * 100/\Delta G_{WT}$) was added to the dataset for that type of mutation.

The database created in Chapter 2 contains 384 structures of antibody-antigen complexes and each of these complexes are of different sizes and, their interface characteristics are also different. This makes their wild-type binding energy different from one another. Comparing the exact numerical values of the change in binding energy for each type of mutation would distort the spread of the dataset, thus it is important to scale the data based on the wildtype binding

energy for each complex. Scaling ensures that the values for change in binding energy for each complex is comparable to one another

Statistical Analysis of the Changes upon Mutation

Statistical procedures based on the assumption that the data follow a Gaussian distribution can make drawing accurate and reliable conclusions difficult [20]. The results for the mutational analysis are sets of data of 380 types of mutations for each force field. For each type of mutation, the outliers were removed for each category of data using the Inter-Quartile Range method. The Shapiro Wilk's test was applied to the data for each type of mutation. The test was conducted before and after removing the outliers. The percentage of datasets that displayed a normal behavior increased from 30% to 77% upon refining the data. The Shapiro Wilk's test was necessary to determine the normality of each type of mutation and how best to incorporate the means and standard deviations into the similarity matrix. It was expected that the data for each mutation will follow the same pattern, that is, a certain mutation should have a similar effect on the binding interactions of an antigen-antibody complex regardless of the structure of the complex.

Building the Matrices for Protein Binding

The mean may be used to describe the numerical value given to each mutation to describe the effects on binding due to that mutation of the significant residues in the 384 protein antigen sequences. The variance for each type of mutation describes the spread of the data set in comparison to the mean; the nature of the spread depends on several factors (such as, the frequency of usage of the wild-type amino acid and the size of the complexes involved). The

purpose of creating the matrices was for their usability as protein design tools in similar ways to similarity matrices.

The PAM and BLOSSUM similarity matrices for protein structures were developed through different methods, but share several critical similarities: they are symmetrical, their values are integers, and they have values for all entries in the matrices, including conserving the current amino acid rather than changing it. The reason the PAM and BLOSSUM matrices are symmetrical arises from their comparison of known protein sequences. If protein A has amino acid X_1 and protein B has amino acid X_2 at equivalent positions, then it is equally valid to say that the mutation is $X_1 \rightarrow X_2$ as it is to say that the mutation is $X_2 \rightarrow X_1$. Thus, the number of times X_1 mutates to X_2 in a set of protein sequences is identical to the number of times X_2 mutates to X_1. For the interface mutations being studied here, that is not the case. These mutations have an evolutionary direction: from an existing complex to a putative complex.

A consequence of this is that the effects of mutating $X_1 \rightarrow X_2$ may be very different than mutating $X_2 \rightarrow X_1$. An example of this from Table 1 is that on average mutating ALA → ARG improved the predicted binding energy by 4.304% while mutating ARG → ALA worsened the predicted binding energy by 10.265%. This is to be expected: when ARG is important in a binding interface it is likely to be part of a salt bridge while ALA's contributions are likely to come from its backbone. Mutating ALA to ARG at that position could still contribute the backbone interactions while creating the potential for a salt bridge whereas mutating ARG to ALA is much more likely to remove a beneficial interaction. As these effects and magnitudes are not equal and similarity matrices for interface mutations should not be symmetrical.

While similarity matrices for interface mutations should not be symmetrical, it is possible to generate versions that share the other features of PAM and BLOSSUM. The first step in doing

so is to determine appropriate numerical scores for retaining a given amino acid rather than mutating it. In PAM and BLOSSUM, the scores were the percentage occurrence of each amino acid and as a result each row summed to one. Here, we chose to have the percentage change in binding energy for each amino acid sum to zero. In other words, the percentage change for retaining a given amino acid was equal to the negative of the sum of all the percentage changes for mutating it (e.g., the score for retaining CYS in Table 1 was 29.708). With those values calculated, the scores for each mutation were calculated using Equation 1.

$$S_{i,j} = \frac{P_{i,j}}{|P_{i,j}|}\log_2\left(|P_{i,j}| + 1\right) \quad (1)$$

where $S_{i,j}$ is the score for mutating amino acid *i* to amino acid *j* (e.g., a value in Table 3) and $P_{i,j}$ is the percentage change for mutating amino acid *i* to amino acid *j* (e.g., a value in Table 1), with the scores rounded to the nearest integer. The magnitude of each score is determined by the $\log_2$ of the absolute value of its percentage change plus one, where the one is added so that all logarithmic values are positive. The use of logarithmic scaling of the scores was based on its use in the PAM and BLOSSUM matrices. The fraction multiplied by the logarithmic value ensures that the scores have the same signs as the percentage changes (i.e., a percentage change that indicates a worsening of binding energy will have a negative score and one that indicates an improved binding energy will have a positive score).

Results and Discussions

The results for the mutational analysis are sets of data of 380 types of mutations for each force field. Tables 1, 2, and 3 are the matrices for the mutational analyses of the protein antigens residues using the CHARMM force field; Table 1 is the matrix created with the mean values for the data sets for each type of mutation, Table 2 is the matrix of the variance for each type of

mutation, and Table 3 is the matrix created by converting the mean values to integer values. Tables 4, 5, and 6 and Tables 7, 8, and 9 are the corresponding tables for the mutational analyses of protein antigen residues using the Amber, and Rosetta force fields, respectively. As mentioned earlier, the analyses were made for both antigen and antibody residues. Tables 10, 11, and 12, Tables 13, 14, and 15, and Tables 16, 17, and 18 are the corresponding tables for the mutational analyses of antibody residues using the CHARMM, Amber, and Rosetta force fields, respectively. Tables 1, 4, 7, 10, 13, and 16 are made of the percentage change in binding energy.

There were 384 complexes used for this study, thus, the nature of the distribution of individual values of the changes is dependent on the total count of each residue. Table 19 shows the total count for each type of residue in the paratopes and the epitopes, as predicted by the different force fields. Given each set for each type of mutation was refined to remove outliers, 79.75%, 79.5%, and 86.5% of the sets showed Gaussian behavior for the mutations in the antigen epitopes using the CHARMM, Amber and Rosetta force fields respectively. Similarly, 74%, 86.5%, and 93% of the sets showed Gaussian behavior for the mutations in the antibody paratopes using the CHARMM, Amber and Rosetta force fields respectively.

Tables 3, 6, 9, 12, 15, and 18 were all converted from the mean values of percentage changes in binding energy for each type of mutation, as observed in Tables 1, 4, 7, 11, 14 and 16 respectively. In the tables with the mean values, the diagonal values are zero, in fact, that mutation was not performed. It is assumed that mutation of the same residue would show no change in binding energy for the complex. In the latter tables with the integer values, the diagonal values are non-zero. The calculations were made to give numerical values to the general trend of mutating the amino acid in question. In these tables, the negative numbers in the diagonals mean it is more preferrable to change the amino acid while the positive numbers mean

it is not preferred to change the amino acid. Higher the number in either direction, the more preferrable or not preferrable the mutation of that residue is. For the rest of the numbers in the matrices, the negative numbers mean a loss of binding energy and the positive numbers mean a gain of binding energy, similar to the matrices with the mean values.

The pattern in how the binding energy changed with different mutations could be classified into trends. Comparing Tables 1 and 3, a difference in amino acid preference can be observed when only means are taken into consideration than when the means are converted to integer numbers. When Table 1 is observed, there is a trend for favorable mutations of Alanine, Glycine and Valine to other amino acids. Favorable changes in energy are observed for Alanine mutating to Cysteine, Phenylalanine, Histidine, Isoleucine, Asparagine, Glutamine, Arginine, Serine, Threonine, Valine, Tryptophan, and Tyrosine. In Table 3, The diagonal numbers are indicative of the general trend to mutate the residues. Mutations of Alanine and Valine are favored, but the mutations of Glycine are not. For Tables 4 and 6, the number show that there is no trend of favoring any mutation, as may be misinterpreted by observing the means only. For Tables 7 and 9, the trends observed in favoring the mutations of Glycine and Serine from the means cannot be observed in the integer values. For Tables 10 and 12, the mutations of Glycine show an interesting trend, where a value of zero would mean that the effect can be unpredictable. Similar changes in trends can be observed for Tables 13 and 15 and Tables 16 and 18.

The force fields currently available do not come without limitations, the observations are evidence of the limitations and biases that exist in force fields. The use of three different force fields helps to identify the different biases that exist in the different potentials used. Comparing Tables 1, 4, and 7, high penalties for mutations of Aspartic Acid, Glutamic Acid, Lysine, and Arginine are observed in the diagonals of the CHARMM and Amber force field results, while

high penalties for mutations of Phenylalanine, Methionine, and Tryptophan are observed in the Rosetta results. The mutation of Tyrosine is very similarly penalized across all the force fields. Tyrosine has both charged and hydrophobic properties and is a dominant amino acid in antibody recognition [21]. The additive force fields are biased towards charged residues, compared to statistical force fields; while the statistical force field is presumably biased towards hydrophobic residues, the high significance of the polar residues in antibody interactions minimizes the bias in the calculations.

The type of force field and the different methods used to predict the energy is reflected in the total count of the type of residues included in the significant residues. For the antigen epitopes, the CHARMM and Amber force fields show significantly higher counts of the polar residues, Aspartic Acid, Glutamic Acid, Lysine, and Arginine, compared to all the other residues, while the Rosetta force field shows higher counts for the polar residues, yet the count is more distributed among all the types of residues. For the antibody paratopes, all the three force fields show the highest count for Tyrosine than any of the other residues. The force fields show higher counts for Aspartic Acid and Arginine, and significant counts for Serine. The total counts for each type of residue can reflect their significance to binding interactions as predicted by each force field, this significance is also reflected in the matrices.

The results show that the trends of mutating the residues follow similar trends whether the residues are in the epitope or the paratope. This directs to the importance of interactions rather than specific residues in antibody binding mechanisms. Thus, the residues that contribute most to certain interactions have eventually become more abundant on the protein surfaces.

The most important feature of antibodies is their ability to bind to their targets with high affinity and high specificity. Antibody affinity maturation is one of the fundamental processes in

the immune defense against pathogens and is extensively studied as a part of antibody design and engineering. Antibodies that are found in this database are assumed to have already undergone extensive affinity maturation. Antigen are assumed to be always evolving. The affinity maturation of antibodies can be observed in the results when antibody mutations are observed to show a general trend of disfavoring any mutation in the important residues.

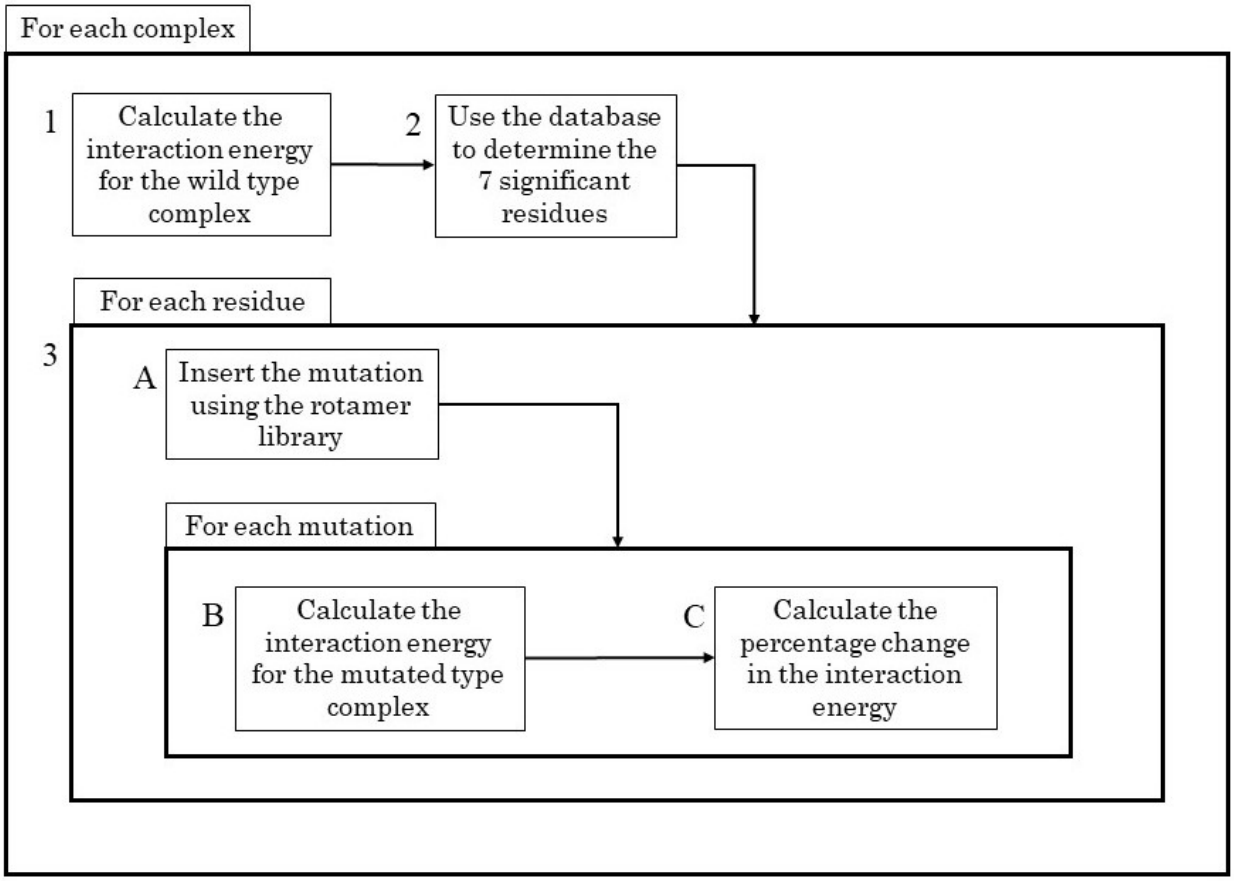
Knowledge of the antibody-antigen complex structure can provide good insights into the antibody-antigen binding mechanism, yet structure alone is not sufficient. An essential part of designing antibodies is developing a thorough understanding of the interactions between the antibodies and their targets and understanding the role of specific residues in these interactions. Protein design tools can now predict structure with high accuracy, though the functional properties of a protein are yet to be successfully designed. The protein interactions take place at the atomic levels and characterizing the properties of the binding site is important to the process of a successfully working protein design. Statistical understanding of the structural features of proteins can help to bring this success in accuracy of structure prediction of proteins. In a similar manner, the statistical understanding of protein interactions can help build an understanding about interface PPIs and help towards better design of interactions.

Experimental techniques were used for studying structures, but these techniques may not be the most feasible to study the mutational changes of antibodies. Experimental techniques do not have the same evolutionary ability as computational techniques, to be able to create these mutations and minimize the structures and calculate the binding affinity at a much shorter time. The results have been converted into matrices to mirror the idea of studying the evolutionary changes in protein structure. It must be noted that the matrices are not symmetrical as other similarity matrices. When studying structure, it is assumed that mutation from one amino acid to

another or vice versa is occurring at the same rate in either direction. The same assumption cannot be made here since different amino acids interact in different ways. It is assumed that the effects of mutation of one amino acid to another will be different vice versa.

The idea of building these matrices was for building knowledge and using this knowledge in designing PPIs. Improving the affinity of protein-protein interactions has always been a challenging problem that has practical applications in the development of therapeutic proteins for diagnostics. Improvement in antigen binding affinity boosts the biological activity of the antibody and can reduce the therapeutic dose of antibody, lowering toxicity and cost [22] – [24].

Figure 9: Algorithm for the calculations required to make the matrix. The steps are divided into two sections. Parts of the algorithm were made to generate data for the mutations made on the individual residues of the individual complexes. The latter part of the algorithm was for analyzing the data and constructing the matrix.



	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	1.936	-5.109	-1.521	1.250	-1.054	1.640	0.841	-0.096	1.016	-0.025	0.640	-0.133	1.977	4.304	0.336	0.528	0.769	2.509	1.873
CYS	-0.407	0.000	-4.614	-4.340	-1.717	-0.912	-1.463	-0.805	-2.567	-1.244	-0.453	-0.451	-1.676	-0.425	-3.497	-0.744	-0.645	-1.357	-1.807	-0.584
ASP	-6.922	-6.019	0.000	-2.999	-5.524	-7.118	-5.884	-6.235	-7.573	-6.515	-5.657	-5.951	-6.395	-4.814	-6.847	-6.240	-6.281	-6.249	-5.508	-5.014
GLU	-8.142	-7.406	-4.830	0.000	-7.076	-8.428	-6.435	-7.525	-8.631	-7.543	-6.901	-6.819	-7.736	-6.010	-7.274	-7.736	-7.420	-7.882	-7.066	-6.404
PHE	-5.225	-3.024	-11.354	-8.470	0.000	-6.143	-3.921	-3.108	-7.057	-3.256	-2.497	-4.474	-4.829	-3.448	-4.306	-4.870	-4.182	-4.024	-1.854	-2.133
GLY	-0.534	0.999	-4.599	-3.183	0.463	0.000	-0.690	-0.396	-1.276	0.205	1.429	-0.193	-0.880	0.339	1.945	-0.105	0.290	-0.853	1.188	0.723
HIS	-4.125	-2.020	-6.560	-4.508	-1.837	-4.448	0.000	-3.145	-2.912	-3.077	-1.783	-3.513	-2.600	-2.078	-2.068	-3.728	-3.253	-3.607	-2.507	-2.008
ILE	-3.845	-1.113	-8.104	-9.321	-0.971	-4.662	-3.208	0.000	-5.155	-1.118	-0.493	-2.659	-3.426	-0.363	-0.777	-2.894	-2.691	-1.772	-0.354	-1.777
LYS	-6.220	-5.078	-8.844	-8.344	-5.448	-6.407	-5.742	-5.667	0.000	-5.871	-5.221	-5.683	-5.904	-5.028	-1.243	-6.022	-5.827	-6.046	-5.359	-4.899
LEU	-2.474	-1.317	-7.023	-6.119	-0.495	-3.055	-1.322	-1.047	-2.768	0.000	0.068	-2.034	-3.395	-1.005	-0.655	-1.883	-1.678	-1.489	-0.272	-1.190
MET	-4.482	-3.128	-6.796	-4.345	-2.466	-5.174	-4.225	-2.827	-2.949	-2.361	0.000	-4.824	-4.724	-4.644	-1.223	-4.044	-4.037	-3.461	-1.951	-2.625
ASN	-4.143	-2.293	-6.138	-4.936	-2.389	-4.825	-2.555	-2.922	-3.721	-2.845	-1.936	0.000	-3.526	-1.815	-0.780	-3.028	-3.062	-3.287	-1.704	-2.113
PRO	-1.463	-0.718	-5.384	-3.400	-0.556	-2.679	-0.935	-0.465	-2.511	-0.787	0.109	-1.885	0.000	-0.221	-0.221	-1.307	-1.235	-0.735	1.749	0.955
GLN	-5.070	-3.767	-7.353	-6.518	-3.756	-5.489	-2.957	-4.098	-5.508	-3.712	-2.958	-3.335	-4.571	0.000	-2.812	-4.674	-4.108	-4.392	-3.580	-3.665
ARG	-10.265	-9.398	-13.330	-12.857	-9.401	-10.602	-9.629	-9.987	-8.528	-10.019	-9.043	-9.760	-10.394	-9.547	0.000	-10.027	-10.129	-10.318	-9.435	-8.554
SER	-3.767	-1.391	-6.093	-4.542	-2.303	-4.829	-2.017	-2.855	-4.690	-3.486	-1.964	-2.151	-3.546	-2.101	-0.505	0.000	-1.032	-3.307	-2.044	-2.112
THR	-3.621	-1.185	-8.602	-6.681	-1.522	-4.462	-2.038	-2.830	-4.708	-2.656	-1.414	-1.990	-3.113	-1.366	-0.550	-1.569	0.000	-3.121	-0.892	-0.461
VAL	-0.169	0.950	-2.828	-0.986	0.965	-1.523	0.631	1.106	1.309	0.752	0.816	0.431	-0.751	0.964	1.483	0.414	0.715	0.000	2.141	2.587
TRP	-6.229	-5.331	-11.480	-11.032	-3.281	-7.690	-4.145	-4.614	-8.968	-4.215	-3.833	-5.025	-5.451	-4.543	-4.205	-5.745	-6.123	-5.313	0.000	-3.084
TYR	-6.218	-5.156	-9.096	-7.876	-4.644	-6.493	-5.625	-5.977	-6.973	-5.681	-5.084	-6.126	-6.310	-5.101	-5.222	-6.294	-6.419	-6.029	-4.435	0.000

Table 1: Matrix showing the means of percentage of changes in binding energy due to point mutations in protein antigens using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	8.558	44.232	25.635	15.772	1.483	20.022	2.184	24.409	6.201	9.987	10.870	3.658	22.215	34.203	5.320	9.806	2.266	13.567	20.632
CYS	6.211	0	83.570	22.644	11.976	9.206	4.271	9.659	7.463	10.974	4.613	10.509	6.453	7.027	10.945	9.300	9.542	7.515	20.828	16.234
ASP	26.614	26.453	0	33.691	33.633	25.749	28.502	30.095	40.266	29.994	29.059	25.212	30.369	28.617	43.484	24.852	28.336	27.095	35.214	34.299
GLU	35.338	30.993	30.814	0	33.708	36.204	32.086	34.210	46.963	37.384	35.541	34.496	35.964	29.338	52.561	32.529	33.231	35.159	40.908	35.898
PHE	16.659	19.259	80.936	43.861	0	20.784	21.736	13.222	54.350	15.018	23.315	27.043	19.254	28.921	58.497	21.123	19.384	16.666	22.338	17.108
GLY	4.897	8.090	47.191	53.986	20.510	0	13.000	13.850	18.366	11.201	19.885	17.077	20.809	15.108	35.844	8.784	11.955	11.001	25.349	26.823
HIS	14.670	19.726	58.188	50.790	17.033	14.484	0	20.419	34.537	17.936	18.130	16.189	13.222	28.663	51.156	16.079	19.949	18.285	25.625	28.618
ILE	4.610	7.216	93.046	134.324	9.291	7.787	12.539	0	35.026	1.498	5.427	17.203	9.457	13.374	24.090	7.844	9.950	3.406	31.350	6.832
LYS	25.301	25.132	44.418	46.669	33.608	27.344	30.468	29.590	0	30.839	29.855	28.972	28.503	29.759	44.079	27.570	27.110	29.462	33.638	34.993
LEU	7.463	6.396	52.400	47.170	6.955	11.102	10.857	4.867	19.399	0	6.140	11.623	11.333	11.642	29.906	14.207	8.296	5.815	21.418	16.885
MET	11.976	8.817	63.631	59.795	2.034	15.892	18.509	8.338	22.571	13.739	0	9.132	11.642	9.801	20.177	13.170	8.605	11.277	9.453	11.296
ASN	9.558	11.564	38.366	33.609	17.793	12.574	18.713	12.206	24.295	14.468	13.578	0	14.444	14.418	24.312	12.757	13.272	11.642	24.183	22.606
PRO	3.544	5.145	88.357	38.530	7.731	8.789	8.466	3.432	13.439	7.129	7.185	8.581	0	7.190	21.505	6.741	6.643	2.232	16.894	32.671
GLN	12.379	10.624	68.962	55.833	18.367	13.628	13.642	12.632	37.773	11.440	10.336	13.419	12.171	0	35.947	16.368	13.967	14.186	21.048	21.384
ARG	47.773	46.882	76.562	73.681	52.922	50.490	52.931	52.793	43.734	57.085	46.475	47.938	52.909	51.719	0	48.613	48.689	53.095	59.814	45.091
SER	15.204	17.047	60.363	59.435	22.317	14.080	17.644	23.662	33.625	21.392	19.610	20.290	26.680	24.083	48.598	0	13.481	19.180	38.909	26.321
THR	16.652	12.435	71.286	78.439	27.281	20.633	24.637	16.354	48.132	14.680	18.095	22.764	22.687	21.035	46.161	16.648	0	18.888	38.579	36.467
VAL	7.924	6.294	123.118	44.676	15.960	16.573	19.020	4.977	11.320	3.310	6.004	15.518	9.074	16.702	14.316	12.851	12.728	0	20.440	18.079
TRP	21.470	23.630	72.074	89.880	20.602	12.467	30.452	22.901	74.667	22.523	16.212	27.240	26.954	25.804	28.895	25.137	19.781	26.195	0	24.009
TYR	12.223	16.106	29.507	34.074	28.639	13.145	19.645	11.828	22.383	12.779	13.842	15.449	13.426	20.590	30.077	16.749	15.171	10.841	24.358	0

Table 2: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-4	2	-3	-1	1	-1	2	1	-1	1	0	1	0	2	2	0	1	1	2	2
C	-1	5	-3	-3	-1	-1	-1	-1	-2	-1	0	-1	-2	-1	-2	-1	-1	-1	-1	0
D	-3	-3	7	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
E	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
F	-3	-2	-4	-3	6	-3	-2	-2	-3	-2	-2	-2	-3	-2	-2	-3	-2	-2	-2	-2
G	-1	1	-2	-2	1	3	-1	0	-1	0	1	0	-1	0	2	0	0	-1	1	1
H	-2	-2	-3	-3	-1	-2	6	-2	-2	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2
I	-2	-1	-3	-3	-1	-2	-2	6	-2	-1	-1	-2	-2	-1	-1	-2	-2	-1	-2	-2
K	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-1	-3	-3	-3	-3	-3
L	-2	-1	-3	-3	0	-2	-1	-1	-2	5	0	-2	-2	-1	-1	-2	-1	-1	0	-1
M	-2	-2	-3	-2	-1	-3	-2	-2	-2	-2	6	-3	-2	-3	-1	-2	-2	-2	-1	-2
N	-2	-2	-3	-3	-2	-3	-2	-2	-2	-2	-2	6	-2	-2	-1	-2	-2	-2	-1	-2
P	-1	-1	-3	-2	-1	-2	-1	-1	-2	-1	0	-2	4	0	0	-1	-1	-1	2	1
Q	-3	-2	-3	-3	-2	-3	-2	-2	-3	-2	-2	-2	-2	6	-2	-3	-2	-2	-2	-2
R	-3	-3	-4	-4	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	8	-3	-3	-3	-3	-3
S	-2	-1	-3	-3	-2	-3	-2	-2	-3	-2	-2	-2	-2	-2	-1	6	-1	-2	-2	-2
T	-2	-1	-3	-3	-1	-2	-2	-2	-3	-2	-1	-2	-2	-1	-1	-1	6	-2	-1	-1
V	0	1	-2	-2	1	-1	1	1	1	1	1	0	-1	1	1	0	1	-3	2	2
W	-3	-3	-4	-4	-2	-3	-2	-2	-3	-2	-2	-3	-3	-2	-2	-3	-3	-3	7	-2
Y	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-2	7

Table 3: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	-3.171	-3.598	-1.597	-2.378	-4.641	-2.178	-4.039	-1.004	-4.471	-3.712	-2.130	-4.603	-2.076	1.616	-3.085	-1.953	-3.401	-1.737	-2.376
CYS	-3.747	0.000	-3.076	-1.852	-2.968	-5.077	-2.409	-2.613	-1.724	-3.144	-2.469	-1.866	-4.026	-2.260	-2.674	-4.195	-3.523	-4.391	-2.974	-3.878
ASP	-12.290	-11.945	0.000	-8.003	-11.708	-12.659	-11.695	-11.836	-12.100	-12.119	-12.092	-10.784	-12.391	-10.783	-11.159	-11.991	-11.613	-12.016	-11.708	-11.558
GLU	-14.140	-13.526	-9.396	0.000	-12.755	-14.324	-12.574	-13.283	-13.150	-13.378	-12.690	-12.525	-13.558	-11.789	-11.697	-13.333	-13.278	-13.468	-12.261	-12.436
PHE	-9.677	-9.331	-8.391	-8.395	0.000	-10.713	-8.432	-9.210	-7.512	-9.653	-8.693	-8.165	-9.973	-7.193	-6.690	-8.682	-9.060	-9.905	-8.382	-8.489
GLY	-6.864	-5.701	-5.295	-3.503	-6.249	0.000	-5.491	-7.445	-2.978	-6.232	-5.421	-4.572	-8.336	-2.959	-0.085	-5.293	-5.780	-6.231	-4.830	-5.132
HIS	-3.281	-2.537	-1.450	-1.570	-1.798	-3.360	0.000	-2.294	-1.275	-2.261	-1.647	-2.365	-3.406	-2.018	-0.007	-2.266	-2.165	-2.865	-0.971	-1.869
ILE	-5.945	-5.424	-5.188	-4.198	-5.434	-6.240	-5.189	0.000	-2.433	-5.451	-4.506	-4.487	-5.989	-2.712	-0.198	-4.567	-4.944	-4.218	-5.330	-4.792
LYS	-13.642	-13.332	-13.855	-13.020	-12.959	-13.930	-12.991	-13.023	0.000	-13.231	-12.628	-12.512	-13.758	-11.720	-9.000	-13.211	-12.931	-13.392	-12.615	-12.544
LEU	-6.680	-6.143	-5.491	-3.070	-5.957	-7.540	-5.005	-5.686	-3.498	0.000	-5.345	-3.975	-7.569	-4.094	-2.805	-6.437	-5.505	-6.191	-6.680	-4.707
MET	-8.231	-8.016	-8.338	-6.254	-4.988	-9.384	-6.794	-6.966	-2.723	-6.386	0.000	-6.268	-7.669	-5.566	-3.687	-7.376	-6.839	-6.617	-6.420	-5.150
ASN	-8.173	-7.329	-6.471	-5.744	-7.396	-8.577	-7.098	-7.650	-5.023	-7.710	-7.575	0.000	-8.031	-5.513	-3.011	-6.815	-6.702	-7.874	-7.415	-7.403
PRO	-4.888	-3.561	-2.633	-3.790	-2.713	-5.395	-3.350	-4.258	0.575	-3.300	-3.530	-0.958	0.000	-1.473	1.861	-2.039	-3.096	-4.169	-2.010	-2.859
GLN	-9.177	-8.926	-6.993	-5.700	-8.372	-9.692	-8.110	-7.958	-5.277	-8.588	-7.765	-6.976	-9.428	0.000	-5.249	-8.567	-8.125	-8.739	-8.240	-7.952
ARG	-15.296	-14.650	-15.187	-14.214	-14.352	-15.473	-14.675	-14.719	-11.136	-14.522	-14.322	-14.054	-15.260	-13.740	0.000	-14.493	-14.283	-15.086	-14.163	-14.052
SER	-6.566	-6.095	-4.183	-3.819	-6.554	-6.940	-5.640	-6.851	-2.897	-6.776	-5.552	-3.443	-7.435	-4.675	-1.595	0.000	-4.245	-6.463	-5.062	-6.056
THR	-8.613	-7.414	-7.048	-4.930	-7.003	-9.449	-6.865	-7.591	-4.808	-7.599	-6.537	-6.067	-8.270	-5.540	-2.853	-6.127	0.000	-8.174	-7.104	-6.553
VAL	-3.681	-3.306	-1.916	-3.366	-3.959	-4.458	-3.372	-2.997	-1.643	-3.959	-2.764	-2.586	-3.947	-1.308	-1.680	-2.690	-3.039	0.000	-2.661	-3.978
TRP	-8.966	-8.571	-8.728	-6.542	-6.926	-8.987	-7.002	-8.257	-5.892	-7.599	-7.317	-6.999	-9.302	-6.766	-4.840	-8.653	-8.054	-8.959	0.000	-6.547
TYR	-7.094	-6.993	-7.156	-5.699	-6.444	-7.725	-6.688	-6.724	-4.969	-6.968	-5.997	-6.286	-7.440	-6.345	-3.933	-6.112	-6.163	-6.890	-5.662	0.000

Table 4: Matrix showing the percentage of changes in binding energy due to point mutations in protein antigens using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	18.897	49.253	58.528	41.601	36.994	24.750	26.272	39.297	28.224	41.512	37.420	31.196	21.579	51.259	33.466	20.385	27.977	59.568	35.888
CYS	5.149	0	5.800	41.599	8.205	3.636	3.719	3.780	26.958	6.216	3.306	1.980	5.126	4.745	59.834	1.801	5.584	7.248	4.504	6.353
ASP	37.931	38.966	0	36.867	41.976	35.149	41.198	36.463	46.712	34.671	38.176	32.686	40.184	37.286	52.217	38.595	39.082	40.623	42.118	44.102
GLU	44.921	42.613	45.669	0	43.142	47.274	42.955	41.827	49.664	46.499	39.480	43.602	45.002	37.058	53.595	44.070	45.496	43.350	42.014	42.481
PHE	32.011	21.670	38.262	38.868	0	28.894	22.009	19.284	40.519	17.704	32.158	23.317	24.069	29.162	55.223	26.556	20.980	19.023	35.497	39.357
GLY	26.185	23.759	41.479	38.126	28.739	0	20.785	23.706	44.601	36.071	31.288	38.139	29.925	31.482	86.804	24.098	32.777	17.900	14.933	25.905
HIS	26.964	21.268	11.373	11.759	11.735	24.874	32.577	15.648	10.081	16.250	13.371	15.330	24.515	14.712	0.651	15.347	13.375	21.221	5.488	13.015
ILE	30.222	27.876	31.445	32.902	52.418	15.449	32.452	0	77.837	30.472	22.313	35.896	34.770	34.195	52.738	40.315	34.671	14.921	32.366	48.148
LYS	42.481	38.829	45.052	43.415	46.007	39.295	41.136	41.994	0	40.334	40.283	37.357	43.151	38.649	47.387	40.405	38.636	41.972	41.874	44.068
LEU	35.969	46.568	73.390	81.250	57.677	38.737	39.149	38.910	61.677	0	37.996	36.036	48.018	36.391	54.740	40.765	36.915	31.936	52.964	69.001
MET	23.574	18.923	13.716	24.087	26.101	26.863	42.019	17.875	51.911	17.753	0	15.274	20.858	21.814	49.070	38.973	40.984	14.640	59.594	30.206
ASN	34.104	37.159	48.697	52.818	38.511	37.579	41.571	38.490	50.098	35.123	37.478	0	33.630	33.113	63.485	36.683	35.777	36.817	44.359	46.259
PRO	32.592	34.072	63.941	28.911	52.700	31.512	54.217	41.864	73.915	46.519	35.568	37.544	0	39.621	90.085	42.209	32.401	31.282	59.786	43.425
GLN	36.435	39.031	53.905	56.118	41.256	36.850	39.317	33.377	48.738	30.999	32.745	37.289	35.067	0	44.682	33.472	39.228	29.147	37.641	39.221
ARG	54.326	51.059	59.022	49.453	59.263	53.001	62.057	59.393	48.561	54.380	56.066	54.376	61.198	55.763	0	45.182	51.628	56.046	63.254	58.571
SER	32.861	32.182	54.447	64.136	41.888	32.560	41.451	40.951	56.757	37.146	33.683	33.337	48.695	41.549	65.673	0	32.457	32.767	42.345	50.802
THR	31.564	32.219	45.210	47.900	40.858	35.819	35.370	37.527	62.568	40.039	40.350	32.449	31.594	30.459	69.349	24.165	0	33.981	58.334	51.908
VAL	18.102	18.503	39.808	26.207	32.206	24.405	31.312	27.739	20.088	31.920	26.436	20.869	21.839	36.618	28.361	17.934	20.958	0	26.347	27.861
TRP	35.446	33.330	32.340	45.531	33.218	33.321	28.546	36.790	55.086	28.473	34.742	33.876	31.589	32.701	46.706	33.971	35.199	37.716	0	33.949
TYR	40.389	35.458	39.302	55.225	36.027	34.886	40.529	31.924	69.398	41.853	41.778	42.728	40.132	45.360	52.642	43.006	37.040	36.770	50.067	0

Table 5: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6	-2	-2	-1	-2	-2	-2	-2	-1	-2	-2	-2	-2	-2	1	-2	-2	-2	-1	-2
C	-2	6	-2	-2	-2	-3	-2	-2	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
D	-4	-4	8	-3	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-3	8	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
F	-3	-3	-3	-3	7	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
G	-3	-3	-3	-2	-3	7	-3	-3	-2	-3	-3	-2	-3	-2	0	-3	-3	-3	-3	-3
H	-2	-2	-1	-1	-1	-2	5	-2	-1	-2	-1	-2	-2	-2	0	-2	-2	-2	-1	-2
I	-3	-3	-3	-2	-3	-3	-3	6	-2	-3	-2	-2	-3	-2	0	-2	-3	-2	-3	-3
K	-4	-4	-4	-4	-4	-4	-4	-4	8	-4	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4
L	-3	-3	-3	-2	-3	-3	-3	-3	-2	7	-3	-2	-3	-2	-2	-3	-3	-3	-3	-3
M	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	7	-3	-3	-3	-2	-3	-3	-3	-3	-3
N	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-2	-3	-3	-3	-3	-3
P	-3	-2	-2	-2	-2	-3	-2	-2	1	-2	-2	-1	6	-1	2	-2	-2	-2	-2	-2
Q	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3
R	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	8	-4	-4	-4	-4	-4
S	-3	-3	-2	-2	-3	-3	-3	-3	-2	-3	-3	-2	-3	-3	-1	7	-2	-3	-3	-3
T	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	7	-3	-3	-3
V	-2	-2	-2	-2	-2	-2	-2	-2	-1	-2	-2	-2	-2	-1	-1	-2	-2	6	-2	-2
W	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3
Y	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	-3	-3	-3	7

Table 6: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	-0.191	-6.265	-5.863	-1.854	-4.994	-7.132	-0.342	-7.767	-0.612	-0.842	-4.358	-6.393	-3.084	-5.678	-6.095	-3.769	-0.807	0.841	-1.401
CYS	-6.612	0.000	-11.249	-10.504	-3.554	-9.780	-6.551	-6.179	-12.263	-4.465	-8.554	-6.403	-11.367	-12.237	-12.301	-9.579	-7.219	-5.377	-10.872	-8.919
ASP	-6.892	-6.590	0.000	-6.937	-6.560	-8.575	-9.849	-7.294	-10.284	-6.468	-7.371	-8.063	-9.088	-7.976	-8.819	-8.175	-6.747	-6.356	-7.564	-7.298
GLU	-7.580	-6.825	-5.857	0.000	-5.662	-8.983	-7.945	-5.630	-10.054	-5.439	-5.336	-7.926	-9.173	-6.793	-10.430	-9.915	-8.295	-6.322	-5.330	-6.983
PHE	-9.122	-9.916	-11.937	-12.793	0.000	-10.585	-10.229	-7.769	-11.522	-6.981	-8.001	-10.792	-12.764	-10.700	-11.617	-12.498	-10.198	-8.964	-8.446	-9.044
GLY	1.578	1.397	-5.429	-2.097	2.244	0.000	-2.147	0.977	-2.441	1.394	2.427	-1.413	-6.221	-0.841	-1.696	-4.059	-1.947	1.742	2.143	-0.714
HIS	-0.355	-0.322	-0.812	-1.427	-0.078	-0.833	0.000	-0.055	-0.310	-0.289	-0.084	-0.990	-0.744	-1.324	-0.049	-1.205	-1.175	-0.220	0.003	-0.128
ILE	-6.900	-6.136	-10.744	-10.114	-4.129	-8.838	-7.668	0.000	-9.007	-3.336	-5.275	-8.572	-8.040	-7.683	-7.443	-9.428	-8.410	-5.015	-5.533	-5.994
LYS	-7.571	-7.358	-10.340	-10.359	-6.325	-8.532	-9.103	-6.787	0.000	-6.377	-6.289	-8.826	-9.061	-8.678	-5.990	-9.557	-8.257	-6.190	-7.303	-8.109
LEU	-6.082	-5.765	-8.348	-8.841	-3.582	-9.776	-6.852	-4.025	-10.640	0.000	-3.563	-7.118	-8.607	-7.461	-9.122	-10.195	-7.593	-4.871	-7.900	-5.083
MET	-10.432	-10.262	-11.961	-10.246	-8.023	-12.612	-10.953	-9.400	-13.442	-5.751	0.000	-10.780	-12.803	-9.783	-12.590	-13.356	-10.590	-10.324	-8.625	-10.591
ASN	-3.465	-2.374	-5.686	-5.965	-0.997	-4.740	-5.760	-1.292	-5.350	-1.360	-2.291	0.000	-5.616	-3.519	-4.770	-6.716	-4.249	-2.011	-1.432	-3.468
PRO	-2.690	-1.725	-7.155	-8.667	-2.428	-6.329	-6.068	-0.984	-8.152	-0.900	-1.305	-6.555	0.000	-5.108	-5.391	-7.863	-5.166	-0.002	0.915	-4.740
GLN	-6.905	-6.548	-8.285	-8.747	-3.855	-8.869	-9.160	-5.507	-10.641	-4.658	-3.829	-7.312	-10.389	0.000	-7.676	-9.709	-9.101	-6.495	-6.172	-7.039
ARG	-7.075	-6.178	-9.800	-9.189	-5.561	-8.329	-7.363	-5.510	-6.253	-6.010	-4.798	-7.611	-9.196	-6.739	0.000	-9.273	-8.195	-6.445	-5.166	-6.061
SER	-1.478	0.872	-3.448	-4.876	1.289	-1.966	-3.960	0.083	-6.292	-0.208	0.148	-3.486	-5.497	-3.237	-4.274	0.000	-2.440	0.790	0.648	1.062
THR	-4.200	-4.408	-7.494	-9.235	-2.214	-7.258	-8.184	-3.640	-9.884	-3.511	-2.868	-8.646	-8.349	-5.837	-7.924	-8.661	0.000	-3.349	-1.814	-4.288
VAL	-4.696	-4.350	-7.153	-8.848	-2.986	-7.698	-7.268	-1.526	-8.624	-3.143	-1.861	-6.982	-7.845	-5.996	-6.302	-8.657	-6.141	0.000	-4.647	-4.966
TRP	-13.163	-13.392	-13.398	-13.437	-9.749	-13.117	-10.954	-11.554	-13.252	-11.893	-10.156	-12.880	-12.585	-11.890	-15.060	-13.247	-12.083	-11.132	0.000	-11.056
TYR	-8.006	-6.724	-9.686	-8.191	-5.402	-9.722	-8.887	-6.018	-9.946	-6.892	-5.535	-8.176	-9.200	-8.925	-8.713	-9.031	-7.570	-7.652	-5.483	0.000

Table 7: Matrix showing the percentage of changes in binding energy due to point mutations in protein antigens using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	143.932	159.943	185.269	153.463	125.736	175.707	141.007	167.381	142.302	175.909	137.444	125.170	130.741	149.752	112.863	118.263	132.690	177.103	175.368
CYS	183.129	0	182.750	139.158	64.266	148.521	67.290	115.521	124.368	90.041	136.377	148.094	112.931	95.180	53.133	166.174	217.186	147.180	139.097	128.218
ASP	141.084	155.561	0	124.558	176.046	142.202	140.966	130.475	157.430	155.586	151.397	153.162	144.005	154.580	145.505	156.860	171.269	146.552	163.198	172.449
GLU	108.884	123.794	156.703	0	124.266	102.623	115.677	121.776	144.824	117.954	115.646	131.056	113.569	135.443	123.044	112.441	117.273	108.552	134.777	145.925
PHE	106.299	96.168	76.499	99.637	0	111.708	105.171	101.950	133.747	80.933	92.008	116.433	123.534	97.660	122.481	102.728	106.033	124.839	134.593	91.876
GLY	92.665	103.321	124.718	124.883	121.505	0	93.663	101.731	93.326	104.422	120.099	103.405	113.796	106.689	101.552	84.608	100.777	118.377	149.849	144.392
HIS	1.846	1.818	8.182	14.529	0.125	4.964	107.641	0.236	3.725	0.814	0.114	8.560	6.873	8.384	1.405	8.936	6.292	0.898	0.012	0.314
ILE	109.077	106.585	91.670	107.452	83.014	106.687	123.226	0	112.129	94.226	78.777	99.202	112.770	113.547	139.093	96.903	99.589	93.044	100.664	89.274
LYS	93.137	72.712	88.799	128.426	105.747	98.198	85.200	87.849	0	82.454	90.414	87.902	90.094	83.201	128.273	87.874	99.235	82.717	105.839	111.336
LEU	156.197	136.021	206.578	189.112	159.798	145.083	158.080	141.329	151.613	0	121.166	137.116	170.417	156.351	163.712	150.355	133.656	135.756	166.544	181.739
MET	143.224	83.932	107.590	130.225	179.806	143.021	137.433	145.624	180.688	157.636	0	116.350	155.242	160.783	136.772	129.653	130.390	98.519	174.358	154.759
ASN	78.793	73.918	80.964	84.111	97.123	85.011	96.181	66.242	127.042	102.043	82.168	0	92.562	93.005	145.103	64.296	87.547	68.104	122.048	93.840
PRO	78.959	108.316	112.924	145.850	135.211	77.194	176.943	132.206	126.302	105.834	123.141	111.944	0	130.008	110.557	99.954	87.140	84.416	168.098	161.546
GLN	114.696	112.057	110.813	134.899	115.471	103.760	131.876	144.954	135.200	135.887	101.540	158.825	122.105	0	141.528	87.152	131.870	122.997	161.865	131.690
ARG	106.662	117.021	123.790	124.006	123.796	118.923	112.311	122.489	132.149	103.015	99.663	108.740	108.517	141.165	0	121.602	112.971	109.893	156.390	120.773
SER	113.563	112.134	222.926	246.236	144.651	108.919	91.713	131.618	244.229	177.734	127.448	203.642	117.680	128.168	123.509	0	141.473	187.527	136.023	146.211
THR	118.457	132.786	104.092	104.759	176.146	94.162	138.753	134.577	172.461	121.773	131.924	112.893	112.346	144.215	122.576	129.322	0	125.586	245.594	147.420
VAL	120.679	122.791	137.557	150.613	168.735	121.051	144.102	147.871	199.957	146.193	158.282	166.832	111.377	144.002	219.722	111.469	131.042	0	158.503	131.512
TRP	135.827	135.470	132.914	122.491	154.821	189.800	180.038	110.828	191.971	117.600	113.083	120.608	154.028	148.346	122.555	136.057	172.120	173.856	0	194.358
TYR	118.482	161.656	114.104	146.084	128.968	123.221	126.648	126.455	161.936	114.010	149.282	169.330	154.340	119.670	155.649	155.304	148.572	127.643	151.447	0

Table 8: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antigens using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6	0	-3	-3	-2	-3	-3	0	-3	-1	-1	-2	-3	-2	-3	-3	-2	-1	1	-1
C	-3	7	-4	-4	-2	-3	-3	-3	-4	-2	-3	-3	-4	-4	-4	-3	-3	-3	-4	-3
D	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
E	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-4	-3	-3	-3	-3	-3
F	-3	-3	-4	-4	8	-4	-3	-3	-4	-3	-3	-4	-4	-4	-4	-4	-3	-3	-3	-3
G	1	1	-3	-2	2	4	-2	1	-2	1	2	-1	-3	-1	-1	-2	-2	1	2	-1
H	0	0	-1	-1	0	-1	4	0	0	0	0	-1	-1	-1	0	-1	-1	0	0	0
I	-3	-3	-4	-3	-2	-3	-3	7	-3	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
K	-3	-3	-4	-4	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
L	-3	-3	-3	-3	-2	-3	-3	-2	-4	7	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3
M	-4	-3	-4	-3	-3	-4	-4	-3	-4	-3	8	-4	-4	-3	-4	-4	-4	-4	-3	-4
N	-2	-2	-3	-3	-1	-3	-3	-1	-3	-1	-2	6	-3	-2	-3	-3	-2	-2	-1	-2
P	-2	-1	-3	-3	-2	-3	-3	-1	-3	-1	-1	-3	6	-3	-3	-3	-3	0	1	-3
Q	-3	-3	-3	-3	-2	-3	-3	-3	-4	-3	-2	-3	-4	7	-3	-3	-3	-3	-3	-3
R	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3
S	-1	1	-2	-3	1	-2	-2	0	-3	0	0	-2	-3	-2	-2	5	-2	1	1	1
T	-2	-2	-3	-3	-2	-3	-3	-2	-3	-2	-2	-3	-3	-3	-3	-3	7	-2	-1	-2
V	-3	-2	-3	-3	-2	-3	-3	-1	-3	-2	-2	-3	-3	-3	-3	-3	-3	7	-2	-3
W	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4	-4	-4	8	-4
Y	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7

Table 9: Matrix showing the effects of the changes in binding energy due to point mutations in protein antigens using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	2.954	-6.755	-2.340	1.780	0.045	1.528	0.478	-1.092	0.073	1.214	0.457	-0.709	0.350	-1.339	2.139	0.889	0.723	1.621	2.739
CYS	-2.442	0.000	-12.689	-5.938	-2.464	-3.512	0.499	-5.986	-2.697	-6.521	-6.452	-4.193	-4.095	-4.614	-2.478	1.442	-4.096	-2.127	1.584	-1.192
ASP	-7.834	-7.541	0.000	-2.466	-7.011	-7.920	-6.851	-7.311	-8.708	-7.425	-7.274	-6.488	-7.773	-6.275	-8.009	-7.305	-7.220	-7.353	-7.108	-6.891
GLU	-6.995	-7.065	-3.509	0.000	-6.586	-7.226	-6.546	-7.387	-8.653	-6.891	-6.422	-6.446	-7.132	-6.044	-7.701	-7.194	-7.092	-6.839	-6.912	-6.484
PHE	-5.649	-3.767	-13.328	-10.934	0.000	-6.270	-4.318	-3.369	-7.649	-3.900	-3.146	-4.626	-5.985	-4.259	-4.493	-5.150	-4.641	-4.097	-3.969	-5.315
GLY	-0.600	0.476	-1.879	-2.128	0.514	0.000	0.412	-0.536	-0.533	-1.013	1.812	0.076	-2.455	0.650	2.264	-0.279	-0.088	-1.100	1.983	1.964
HIS	-5.820	-4.507	-7.812	-5.690	-4.681	-6.534	0.000	-4.923	-3.709	-5.437	-4.331	-4.898	-6.587	-5.484	-5.147	-5.588	-5.243	-5.014	-3.953	-3.751
ILE	-1.310	-0.577	-9.842	-9.989	0.264	-3.140	-1.125	0.000	-3.103	0.499	0.504	-1.855	-1.928	0.347	-0.878	-1.019	-1.822	-0.307	-0.514	1.651
LYS	-4.568	-3.893	-6.348	-5.924	-4.103	-4.501	-4.298	-4.583	0.000	-4.694	-4.160	-4.184	-4.845	-4.054	-1.510	-4.329	-4.207	-4.767	-4.143	-3.783
LEU	-2.127	-0.764	-9.661	-6.589	0.211	-2.917	-0.824	-0.604	-2.444	0.000	0.872	-2.188	-1.932	-0.490	-0.533	-2.980	-2.548	-1.024	0.311	0.865
MET	-5.236	-2.797	-8.004	-9.965	-1.834	-5.697	-2.429	-2.286	-3.091	-2.878	0.000	-4.653	-3.936	-2.793	-3.902	-3.830	-2.668	-1.174	-2.376	1.748
ASN	-4.154	-2.359	-5.226	-5.057	-2.684	-4.763	-2.588	-2.579	-3.275	-3.026	-2.282	0.000	-2.973	-2.542	-2.213	-2.788	-3.193	-3.199	-2.745	-1.991
PRO	-2.520	-0.665	-9.116	-3.128	-0.746	-3.366	-2.503	-1.686	-6.430	-1.510	-1.931	-2.254	0.000	-4.263	-3.464	-1.932	-2.859	-1.516	-1.189	-1.885
GLN	-4.050	-3.087	-5.359	-5.242	-3.356	-4.044	-3.010	-3.876	-3.703	-3.914	-2.845	-3.488	-3.701	0.000	-2.923	-3.900	-3.623	-4.429	-3.008	-3.276
ARG	-10.072	-9.138	-12.304	-12.272	-9.266	-10.014	-9.781	-9.780	-8.400	-10.137	-9.364	-9.286	-10.001	-9.318	0.000	-9.638	-9.595	-9.866	-9.380	-8.884
SER	-4.390	-2.165	-6.894	-6.342	-3.374	-4.992	-2.727	-3.746	-3.852	-3.862	-3.399	-2.492	-4.785	-3.052	-1.488	0.000	-1.868	-3.639	-3.102	-3.052
THR	-3.790	-2.017	-5.755	-5.568	-2.231	-4.908	-2.584	-3.382	-3.701	-3.425	-2.649	-2.481	-4.121	-2.731	-2.123	-1.549	0.000	-3.212	-2.888	-2.179
VAL	-2.872	-0.786	-7.934	-2.280	0.087	-4.345	-2.072	0.229	-1.091	-0.986	0.562	-2.644	-2.153	-1.049	-1.869	-3.013	-1.474	0.000	0.864	-0.023
TRP	-6.881	-6.183	-11.034	-10.139	-4.835	-7.361	-6.300	-5.787	-7.993	-5.853	-5.158	-6.938	-7.388	-5.838	-6.394	-6.913	-6.506	-6.519	0.000	-6.022
TYR	-7.409	-6.566	-11.254	-10.247	-5.854	-7.911	-6.877	-6.543	-8.373	-6.525	-6.168	-7.336	-7.297	-6.579	-5.933	-7.450	-7.291	-6.949	-6.296	0.000

Table 10: Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	8.661	55.818	58.436	8.462	4.831	13.817	4.388	12.591	6.290	7.999	14.224	11.240	20.746	15.266	11.533	11.668	4.492	31.619	28.987
CYS	1.439	0	197.585	48.421	0.000	4.347	28.212	11.259	11.834	11.55	17.961	0.582	1.066	5.956	0.000	0.396	3.314	1.209	0.000	23.8
ASP	40.597	39.843	0	27.614	42.342	36.691	36.688	43.443	45.504	40.859	42.648	34.867	38.989	34.967	53.263	36.834	41.030	40.757	46.901	45.132
GLU	37.812	38.116	29.757	0	45.127	34.773	32.973	40.018	55.429	38.948	37.303	36.149	42.577	30.933	53.238	39.298	38.951	34.882	45.258	41.386
PHE	7.821	11.196	35.070	28.387	0	9.551	22.864	7.994	34.983	9.839	9.065	14.405	7.125	11.205	55.380	9.506	8.694	8.752	25.268	23.456
GLY	6.660	11.781	56.662	68.069	29.423	0	14.727	20.885	41.810	16.146	22.779	18.378	24.770	23.502	27.960	6.105	10.831	19.212	20.254	28.391
HIS	8.171	8.725	28.425	28.844	13.021	9.178	0	10.599	10.521	13.072	14.418	13.705	12.661	14.641	24.681	10.430	10.673	10.803	14.522	30.080
ILE	22.999	18.278	65.258	90.655	22.896	26.387	17.061	0	69.733	17.350	21.439	21.641	8.318	31.461	31.674	16.068	9.210	14.374	30.001	52.590
LYS	16.518	16.375	19.814	21.104	19.311	16.544	15.966	18.432	0	17.806	19.382	16.816	18.262	17.557	35.046	14.940	15.550	20.530	21.320	26.765
LEU	7.012	6.858	121.469	62.874	9.234	11.032	10.543	5.626	10.553	0	5.952	3.524	14.392	13.107	25.075	14.788	7.419	6.065	17.055	12.862
MET	13.312	10.725	45.240	79.953	3.933	15.040	33.102	5.428	55.794	3.006	0	5.131	6.952	7.639	31.275	7.577	7.272	7.718	41.140	10.517
ASN	12.081	12.673	43.699	43.343	21.761	11.210	21.753	15.142	36.132	17.610	18.259	0	14.526	18.442	33.932	16.432	14.045	14.405	26.292	28.556
PRO	6.037	2.777	119.178	16.459	8.363	12.002	7.212	12.379	47.228	7.073	16.265	1.388	0	11.127	10.102	0.821	11.710	9.011	32.606	4.524
GLN	12.900	8.746	31.052	23.858	12.987	13.933	9.288	9.005	23.090	10.528	9.387	9.375	8.565	0	15.629	14.516	14.019	11.486	10.247	11.879
ARG	38.888	35.106	45.750	48.215	40.295	37.765	38.248	38.523	33.064	39.257	38.056	37.004	36.619	34.149	0	36.818	38.065	38.740	38.654	44.546
SER	11.338	8.949	41.698	43.015	24.153	12.511	17.353	15.536	35.395	16.666	18.910	14.383	17.104	17.459	38.198	0	11.705	14.720	30.215	29.249
THR	10.242	10.705	42.952	49.483	20.792	11.239	18.511	12.246	30.043	11.615	12.180	11.507	17.603	16.071	43.217	11.041	0	11.541	30.084	21.804
VAL	3.037	1.981	70.432	17.286	7.983	8.165	6.182	3.280	2.129	1.487	2.179	4.035	2.448	4.551	6.731	5.503	5.004	0	8.422	8.404
TRP	13.170	10.659	49.737	40.966	14.530	12.620	14.033	11.091	36.691	13.510	14.194	12.759	14.844	12.139	27.865	14.347	13.473	13.331	0	28.123
TYR	15.776	16.008	36.461	35.693	21.453	16.232	20.700	16.435	31.161	16.440	18.728	19.154	16.573	17.658	34.926	20.308	18.870	17.073	31.621	0

Table 11: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-2	2	-3	-2	2	0	1	0	-2	0	1	1	-1	1	-1	2	1	1	2	2
C	-2	6	-4	-3	-2	-2	1	-3	-2	-3	-3	-2	-2	-2	-2	1	-2	-2	1	-1
D	-3	-3	7	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
E	-3	-3	-2	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
F	-3	-2	-4	-4	7	-3	-3	-2	-3	-2	-2	-3	-3	-3	-3	-3	-3	-2	-2	-3
G	-1	0	-2	-2	1	0	1	-1	0	-1	1	0	-2	1	2	0	0	-1	2	2
H	-3	-2	-3	-3	-2	-3	7	-2	-2	-2	-2	-2	-3	-3	-2	-3	-3	-2	-2	-2
I	-1	-1	-4	-3	0	-2	-1	5	-2	0	1	-2	-2	0	-1	-1	-2	-1	0	2
K	-3	-2	-3	-3	-2	-3	-2	-3	6	-3	-2	-2	-3	-2	-2	-2	-2	-3	-2	-2
L	-2	-1	-3	-3	0	-2	-1	-1	-2	5	1	-2	-2	0	-1	-2	-2	-1	0	1
M	-3	-2	-3	-4	-1	-3	-3	-2	-2	-2	6	-2	-2	-2	-2	-2	-2	-2	-2	1
N	-2	-2	-3	-3	-2	-3	-2	-2	-2	-2	-2	6	-2	-2	-2	-2	-2	-2	-2	-2
P	-1	0	-3	-2	0	-2	-1	-1	-3	-1	0	-2	5	0	-2	-1	-1	-1	1	-1
Q	-2	-2	-3	-3	-2	-2	-2	-2	-2	-2	-2	-2	-2	6	-2	-2	-2	-2	-2	-2
R	-3	-3	-4	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	8	-3	-3	-3	-3	-3
S	-2	-2	-3	-3	-2	-3	-2	-2	-2	-2	-2	-2	-2	-2	-1	6	-2	-2	-2	-2
T	-2	-2	-3	-3	-2	-3	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	6	-2	-2	-2
V	-2	-1	-3	-2	0	-2	-2	0	-1	-1	1	-2	-2	-1	-1	-2	-1	5	0	0
W	-3	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3
Y	-3	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7

Table 12: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the CHARMM force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	-1.699	-5.272	-4.644	-3.284	-4.678	-4.234	-4.507	-1.868	-4.574	-5.794	-1.798	-6.194	-4.073	0.768	-3.361	-1.344	-2.656	-4.522	-6.816
CYS	-5.043	0.000	-8.137	-5.238	-8.751	-4.394	-4.164	-6.825	-7.985	-4.856	-7.106	-5.026	-4.996	-6.628	-3.169	-4.272	-5.588	-4.397	-7.329	-12.145
ASP	-13.545	-13.453	0.000	-7.546	-12.956	-13.309	-12.740	-13.450	-13.545	-13.253	-13.085	-11.991	-13.908	-11.983	-13.152	-12.812	-13.009	-13.480	-12.545	-13.045
GLU	-12.843	-12.334	-8.110	0.000	-12.498	-13.427	-12.341	-12.490	-12.982	-12.560	-12.934	-11.683	-12.955	-11.647	-12.014	-12.340	-12.885	-13.393	-12.643	-12.479
PHE	-6.777	-6.309	-5.775	-5.137	0.000	-7.211	-4.726	-5.700	-4.934	-6.669	-4.766	-5.640	-6.967	-5.169	-5.417	-6.217	-5.749	-5.554	-6.537	-3.938
GLY	-4.762	-5.183	-3.928	-3.195	-3.179	0.000	-3.478	-3.714	-1.323	-5.052	-3.864	-2.968	-7.622	-0.830	1.672	-3.133	-4.704	-6.371	-2.595	-3.807
HIS	-2.486	-2.299	-0.152	-0.794	-0.827	-2.845	0.000	-2.022	0.096	-1.767	-1.475	-0.120	-2.789	-0.003	-0.162	-1.055	-1.838	-2.489	-1.579	-6.727
ILE	-3.429	-3.130	0.581	1.352	-1.299	-4.054	-1.211	0.000	1.577	-3.495	-1.529	-1.347	-2.892	1.686	3.254	-0.530	-1.576	-2.392	-3.822	-2.320
LYS	-13.125	-12.797	-13.637	-13.848	-13.147	-12.737	-12.752	-13.342	0.000	-13.041	-12.817	-12.747	-12.836	-12.298	-9.242	-11.946	-12.390	-12.694	-12.838	-12.732
LEU	-3.834	-4.435	-3.488	-2.674	-3.915	-4.620	-3.028	-3.996	0.722	0.000	-3.658	-2.213	-4.822	-1.839	-0.451	-3.575	-3.663	-3.436	-3.203	-4.071
MET	-6.329	-4.197	-5.421	-2.545	-3.633	-6.426	-3.890	-2.675	-4.991	-3.808	0.000	-5.476	-4.904	-4.696	-1.487	-3.906	-4.237	-3.632	-3.175	-2.916
ASN	-9.130	-8.419	-8.548	-8.046	-8.465	-9.673	-7.838	-8.832	-6.111	-8.693	-7.991	0.000	-8.702	-6.965	-7.079	-7.377	-7.916	-8.654	-7.795	-8.516
PRO	-6.486	-5.178	-2.653	-1.643	-7.688	-6.398	-9.598	-5.820	-0.974	-5.741	-6.179	-5.273	0.000	-1.760	-3.648	-3.382	-4.228	-5.065	-4.463	-8.062
GLN	-7.131	-6.401	-5.858	-5.137	-5.084	-8.212	-6.269	-6.277	-4.304	-6.761	-5.524	-5.461	-7.177	0.000	-3.857	-5.743	-5.666	-6.262	-4.559	-5.361
ARG	-15.181	-14.209	-15.314	-14.528	-14.107	-15.373	-14.341	-14.639	-11.530	-14.912	-14.367	-14.215	-15.197	-14.009	0.000	-14.488	-14.219	-14.802	-14.272	-14.307
SER	-7.129	-6.070	-6.079	-5.477	-7.277	-7.347	-6.879	-6.952	-4.647	-7.226	-7.081	-4.636	-7.497	-5.232	-3.140	0.000	-4.163	-7.195	-6.543	-7.248
THR	-7.329	-6.201	-5.730	-6.346	-7.232	-8.139	-5.957	-7.021	-3.995	-6.813	-7.189	-4.628	-7.285	-4.785	-3.034	-4.556	0.000	-6.894	-6.774	-6.462
VAL	-5.327	-4.321	-1.187	-4.609	-5.220	-7.109	-4.526	-4.852	0.590	-5.040	-1.698	-2.413	-5.001	-1.810	-0.076	-2.936	-4.022	0.000	-1.581	-3.923
TRP	-9.543	-8.649	-9.015	-8.227	-8.259	-9.831	-8.650	-8.717	-8.191	-9.077	-8.316	-8.226	-9.811	-7.391	-7.289	-9.060	-8.958	-8.929	0.000	-8.664
TYR	-8.943	-8.486	-7.843	-7.245	-7.750	-9.440	-7.281	-8.152	-6.564	-8.262	-7.868	-8.132	-8.907	-7.663	-5.963	-8.466	-8.396	-8.487	-7.356	0.000

Table 13:Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	21.601	75.981	79.171	47.341	22.888	52.763	23.455	25.066	42.155	38.004	42.445	48.050	44.047	72.007	41.877	38.815	17.510	92.921	23.757
CYS	0.000	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ASP	52.555	55.056	0	44.661	47.910	45.969	45.780	50.035	61.910	48.127	53.736	53.335	51.893	48.255	58.327	46.740	54.277	52.869	50.959	55.980
GLU	39.079	37.788	32.756	0	37.105	37.762	42.873	42.468	46.122	39.163	46.365	34.790	37.168	39.343	53.327	34.222	38.332	48.512	52.522	44.408
PHE	17.348	25.876	20.242	28.964	0	17.808	22.641	24.376	52.036	18.315	33.244	11.156	28.270	15.262	43.235	9.862	27.985	23.845	24.782	31.801
GLY	45.384	40.343	115.861	134.194	49.876	0	64.288	45.891	80.222	46.171	56.749	56.431	57.383	47.324	108.494	55.077	56.597	59.631	64.805	59.922
HIS	54.058	30.331	35.270	56.501	34.676	60.676	0	43.998	106.624	39.065	67.109	28.997	45.155	35.262	43.329	44.190	51.857	44.453	11.321	83.725
ILE	12.248	20.935	65.152	31.604	18.195	13.053	31.039	0	47.037	20.425	25.163	10.145	12.845	19.255	40.133	5.622	17.500	16.133	106.481	86.023
LYS	34.798	35.767	43.641	36.377	34.303	45.789	45.317	31.495	0	37.114	34.741	35.229	44.158	40.726	27.943	41.536	42.012	37.588	44.589	44.955
LEU	39.577	22.116	50.766	26.095	33.988	43.805	12.526	31.154	78.010	0	33.055	41.042	37.903	58.514	99.578	46.072	38.584	28.115	40.677	30.008
MET	29.191	27.098	17.802	10.511	14.860	29.439	27.369	17.179	39.307	29.850	0	13.613	34.799	3.782	18.753	9.962	18.981	18.859	15.585	17.637
ASN	32.937	35.238	48.769	50.269	42.913	36.415	42.506	33.525	48.715	37.956	33.291	0	41.879	37.360	52.415	44.247	43.701	30.630	48.381	40.868
PRO	31.919	35.307	15.292	35.673	26.105	39.552	44.003	21.983	100.083	29.409	43.096	21.457	0	45.359	72.976	25.955	26.098	22.148	19.030	15.163
GLN	33.583	38.437	39.103	33.886	22.640	38.521	30.321	33.355	26.049	29.330	24.697	34.150	33.098	0	31.719	23.815	15.438	31.970	36.703	29.949
ARG	42.253	49.224	44.352	55.458	43.636	41.434	39.677	38.170	41.320	42.225	46.567	43.083	41.702	41.312	0	54.321	46.169	44.524	38.666	41.789
SER	24.333	19.716	51.171	40.203	29.330	29.077	34.135	36.772	47.588	30.335	29.219	33.830	32.694	38.668	61.374	0	21.811	22.662	46.336	44.146
THR	43.603	42.135	59.368	62.506	49.391	46.104	44.865	43.567	54.529	34.964	48.892	41.210	44.232	43.261	52.404	37.437	0	36.455	51.988	45.156
VAL	5.283	30.130	33.469	43.132	50.586	6.228	46.854	52.505	79.215	40.437	21.348	27.018	44.302	31.336	69.414	22.248	29.303	0	13.846	33.793
TRP	30.426	29.357	38.049	42.979	40.449	28.495	36.518	28.797	51.323	34.901	31.616	23.656	30.752	39.897	47.440	32.132	33.894	36.765	0	52.894
TYR	35.226	30.127	54.600	49.387	37.652	34.530	40.879	37.151	49.046	34.631	37.014	33.483	30.793	38.635	64.345	35.992	35.552	31.660	45.734	0

Table 14: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6	-1	-3	-2	-2	-3	-2	-2	-2	-2	-3	-1	-3	-2	1	-2	-1	-2	-2	-3
C	-3	7	-3	-3	-3	-2	-2	-3	-3	-3	-3	-3	-3	-3	-2	-2	-3	-2	-3	-4
D	-4	-4	8	-3	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
E	-4	-4	-3	8	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
F	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-2
G	-3	-3	-2	-2	-2	6	-2	-2	-1	-3	-2	-2	-3	-1	1	-2	-3	-3	-2	-2
H	-2	-2	-1	-1	-2	-2	6	-2	1	-2	-1	-2	-3	-2	-2	-2	-2	-2	-2	-3
I	-2	-2	1	1	-1	-2	-1	5	1	-2	-1	-1	-2	1	2	-1	-1	-2	-2	-2
K	-4	-4	-4	-4	-4	-4	-4	-4	8	-4	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4
L	-2	-2	-2	-2	-2	-2	-2	-2	1	6	-2	-2	-3	-2	-1	-2	-2	-2	-2	-2
M	-3	-2	-3	-2	-2	-3	-2	-2	-3	-2	6	-3	-3	-3	-1	-2	-2	-2	-2	-2
N	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3
P	-3	-3	-2	-1	-3	-3	-3	-3	-1	-3	-3	-3	7	-1	-2	-2	-2	-3	-2	-3
Q	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	-3	-3	-3	7	-2	-3	-3	-3	-2	-3
R	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	8	-4	-4	-4	-4	-4
S	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	-3	-2	-3	-3	-2	7	-2	-3	-3	-3
T	-3	-3	-3	-3	-3	-3	-3	-3	-2	-3	-3	-2	-3	-3	-2	-2	7	-3	-3	-3
V	-3	-2	-1	-2	-3	-3	-2	-3	1	-3	-1	-2	-3	-1	0	-2	-2	6	-1	-2
W	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3
Y	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7

Table 15: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the Amber force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0.000	-0.635	-4.901	-3.539	-1.255	-2.957	-4.082	0.042	-4.354	-1.187	-0.807	-3.273	-4.919	-2.161	-4.624	-3.626	-1.006	0.767	-1.713	-2.062
CYS	-6.324	0.000	-8.977	-5.060	-10.850	-4.087	-10.613	-9.347	-9.415	-9.633	-3.640	-8.600	-8.819	-6.445	-12.369	-8.848	-6.694	-12.200	-9.358	-9.176
ASP	-8.174	-8.279	0.000	-6.407	-7.084	-8.620	-9.168	-7.980	-10.590	-7.854	-7.971	-8.742	-9.232	-7.293	-10.418	-8.654	-8.163	-7.332	-8.023	-8.593
GLU	-8.121	-7.373	-4.657	0.000	-7.276	-9.006	-9.647	-7.076	-10.230	-7.807	-6.894	-7.552	-8.340	-6.639	-9.363	-8.298	-8.792	-7.117	-7.179	-8.343
PHE	-6.652	-6.641	-9.447	-8.061	0.000	-8.315	-7.850	-5.252	-7.831	-5.153	-5.298	-8.038	-8.503	-7.278	-8.258	-8.468	-7.348	-5.309	-6.174	-6.520
GLY	-1.048	-1.063	-4.345	-4.266	-0.858	0.000	-3.331	-2.508	-3.980	-0.374	-0.602	-2.320	-8.729	-2.984	-2.592	-4.085	-3.552	-2.583	0.322	-1.379
HIS	-3.575	-4.394	-8.890	-9.103	-7.385	-7.285	0.000	-2.389	-4.606	-4.736	-6.213	-3.599	-8.066	-3.185	-7.426	-6.013	-6.534	-4.800	-6.019	-4.612
ILE	-6.235	-5.487	-10.292	-10.501	-4.571	-8.029	-8.792	0.000	-9.767	-5.378	-5.261	-8.935	-9.562	-5.109	-10.582	-9.164	-6.525	-4.769	-5.167	-7.301
LYS	-8.532	-8.863	-8.586	-8.901	-8.918	-8.448	-9.228	-7.579	0.000	-7.559	-8.369	-9.463	-8.927	-8.875	-7.262	-9.416	-8.760	-6.079	-8.178	-9.048
LEU	-4.520	-3.880	-7.036	-7.186	-2.796	-6.134	-5.074	-3.043	-7.257	0.000	-3.543	-7.113	-6.920	-5.403	-5.815	-6.784	-5.545	-3.551	-4.527	-4.031
MET	-8.170	-6.984	-14.214	-11.947	-7.872	-12.289	-7.104	-4.999	-8.833	-4.472	0.000	-9.968	-8.692	-8.202	-8.537	-11.172	-7.254	-5.154	-6.642	-9.572
ASN	-5.102	-5.760	-5.980	-7.003	-6.599	-5.289	-6.810	-5.097	-7.750	-5.139	-4.760	0.000	-7.275	-4.955	-6.865	-7.996	-6.020	-6.014	-4.764	-6.531
PRO	-1.642	-0.760	-8.283	-7.866	-2.282	-5.134	-8.248	-4.730	-7.594	-0.093	-2.145	-6.295	0.000	-6.098	-7.122	-6.802	-2.691	-0.366	3.397	0.051
GLN	-6.465	-8.125	-7.707	-6.410	-4.367	-7.255	-6.277	-5.450	-6.709	-4.184	-5.083	-7.040	-8.331	0.000	-5.120	-7.186	-7.324	-5.363	-4.532	-4.008
ARG	-8.540	-8.590	-11.099	-10.660	-7.377	-9.315	-9.275	-8.039	-7.055	-7.912	-7.291	-8.669	-9.572	-8.083	0.000	-9.560	-9.562	-8.192	-7.671	-8.474
SER	-0.435	0.159	-1.664	-2.260	-0.041	-2.064	-2.909	0.740	-2.068	0.670	-0.151	-1.327	-2.781	-0.632	-2.426	0.000	-2.491	-0.051	-0.589	-1.156
THR	-3.810	-3.050	-6.149	-6.440	-2.429	-4.749	-5.613	-2.497	-4.323	-1.909	-2.272	-5.681	-5.490	-4.000	-4.609	-6.405	0.000	-2.994	-4.278	-3.288
VAL	-4.426	-3.881	-7.253	-7.407	-4.131	-6.027	-7.758	-4.187	-10.265	-3.424	-3.646	-6.812	-7.730	-6.158	-7.416	-6.436	-5.541	0.000	-5.016	-6.703
TRP	-9.884	-9.925	-11.438	-11.474	-8.508	-10.289	-11.579	-8.217	-12.676	-9.563	-9.109	-10.795	-11.128	-11.139	-12.451	-11.059	-10.002	-9.248	0.000	-10.096
TYR	-7.246	-6.855	-9.301	-8.721	-5.529	-8.604	-8.091	-5.294	-9.023	-5.919	-6.209	-8.425	-8.929	-7.950	-8.411	-8.733	-8.007	-6.142	-6.737	0.000

Table 16: Matrix showing the percentage of changes in binding energy due to point mutations in protein antibodies using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue. The pigmentation of the cell indicates the positive (beneficial), neutral or negative (detrimental) nature of the mutation ranging from green, yellow to red respectively

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	0	120.211	134.321	115.984	127.362	172.431	95.731	95.254	145.858	121.528	137.948	115.124	142.781	117.299	130.832	138.240	142.335	98.741	143.962	114.844
CYS	129.039	0	110.972	141.549	73.664	26.663	163.548	169.615	288.114	214.918	113.395	120.040	147.221	83.351	106.637	188.638	190.980	31.405	168.368	91.745
ASP	118.106	119.586	0	89.982	108.264	119.743	114.455	102.958	109.956	104.725	112.378	114.155	118.115	110.993	106.520	110.585	109.125	117.950	127.680	117.424
GLU	113.253	120.113	97.306	0	111.278	100.024	122.014	119.884	89.321	105.997	127.288	109.963	103.235	121.143	110.446	110.710	128.243	125.470	145.965	106.116
PHE	104.724	110.535	119.190	84.793	0	108.765	111.173	93.204	109.093	100.690	90.975	90.188	122.500	104.016	102.842	101.828	107.693	91.477	133.809	119.418
GLY	79.946	92.810	71.648	82.483	115.055	0	71.751	122.182	142.782	128.495	121.570	103.176	115.467	107.299	142.440	85.769	105.881	117.560	129.699	105.389
HIS	42.954	143.227	127.656	103.068	141.881	83.047	0	123.000	99.787	162.156	108.930	77.168	214.350	186.592	104.859	77.642	130.971	113.011	126.389	122.613
ILE	150.784	139.924	125.047	180.994	144.546	170.500	92.995	0	170.772	144.425	141.473	168.546	122.821	188.169	122.179	165.124	146.495	140.627	147.340	175.199
LYS	83.283	105.475	62.120	83.865	99.317	96.158	81.776	89.361	0	73.096	96.303	88.563	56.103	111.294	86.533	97.752	74.449	82.858	96.896	90.413
LEU	102.957	127.365	116.252	136.559	112.869	104.810	107.978	128.343	161.019	0	120.603	130.331	125.832	108.569	145.470	143.537	102.361	118.889	109.133	134.923
MET	61.923	55.645	90.411	183.642	111.651	168.822	51.153	83.467	63.201	36.644	0	136.019	50.401	77.623	45.350	136.561	76.068	50.700	135.284	117.961
ASN	86.998	92.407	146.882	146.100	162.231	123.477	84.240	122.242	103.249	123.800	128.113	0	135.814	177.475	111.883	78.859	117.498	104.785	171.177	135.466
PRO	172.707	140.037	300.348	223.668	185.042	190.142	268.718	95.804	154.275	161.216	256.184	178.772	0	206.642	188.492	156.932	191.348	129.211	99.394	56.415
GLN	60.790	88.731	118.973	133.290	119.876	92.436	110.629	118.586	112.270	77.968	110.982	114.797	89.548	0	65.868	77.883	109.488	65.408	128.772	93.718
ARG	100.883	101.791	119.739	116.142	106.428	104.212	108.525	114.808	111.691	116.042	126.346	111.771	111.647	102.803	0	104.234	113.314	104.928	113.977	104.175
SER	106.933	114.858	171.332	189.699	140.882	95.614	114.207	152.285	148.716	140.555	148.312	132.386	123.733	174.079	128.880	0	105.435	124.265	168.389	139.476
THR	88.780	84.965	95.009	110.344	99.949	82.062	101.718	92.549	96.143	120.853	88.102	99.519	62.243	75.410	105.437	86.381	0	124.797	143.079	95.267
VAL	94.656	79.025	146.746	143.643	165.546	97.858	109.119	110.409	109.814	71.910	102.775	80.663	123.522	104.922	132.633	79.746	85.602	0	129.154	139.583
TRP	109.859	99.879	108.449	103.575	136.373	118.218	132.583	123.248	127.445	113.767	108.090	102.816	110.806	112.355	114.458	118.680	100.402	103.343	0	129.538
TYR	104.129	114.413	134.493	120.435	124.726	118.441	115.994	121.366	122.685	115.520	112.274	109.930	116.135	111.095	129.142	111.469	117.358	115.012	130.798	0

Table 17: Matrix showing the variances in percentage of changes in binding energy due to point mutations in protein antibodies using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6	-1	-3	-2	-1	-2	-2	0	-2	-1	-1	-2	-3	-2	-2	-2	-1	1	-1	-2
C	-3	7	-3	-3	-4	-2	-4	-3	-3	-3	-2	-3	-3	-3	-4	-3	-3	-4	-3	-3
D	-3	-3	7	-3	-3	-3	-3	-3	-4	-3	-3	-3	-3	-3	-4	-3	-3	-3	-3	-3
E	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
F	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
G	-1	-1	-2	-2	-1	6	-2	-2	-2	0	-1	-2	-3	-2	-2	-2	-2	-2	0	-1
H	-2	-2	-3	-3	-3	-3	7	-2	-2	-3	-3	-2	-3	-2	-3	-3	-3	-3	-3	-2
I	-3	-3	-3	-4	-2	-3	-3	7	-3	-3	-3	-3	-3	-3	-4	-3	-3	-3	-3	-3
K	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
L	-2	-2	-3	-3	-2	-3	-3	-2	-3	7	-2	-3	-3	-3	-3	-3	-3	-2	-2	-2
M	-3	-3	-4	-4	-3	-4	-3	-3	-3	-2	7	-3	-3	-3	-3	-4	-3	-3	-3	-3
N	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3	-3	-3	-3
P	-1	-1	-3	-3	-2	-3	-3	-3	-3	0	-2	-3	6	-3	-3	-3	-2	0	2	0
Q	-3	-3	-3	-3	-2	-3	-3	-3	-3	-2	-3	-3	-3	7	-3	-3	-3	-3	-2	-2
R	-3	-3	-4	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7	-3	-3	-3	-3	-3
S	-1	0	-1	-2	0	-2	-2	1	-2	1	0	-1	-2	-1	-2	4	-2	0	-1	-1
T	-2	-2	-3	-3	-2	-3	-3	-2	-2	-2	-2	-3	-3	-2	-2	-3	6	-2	-2	-2
V	-2	-2	-3	-3	-2	-3	-3	-2	-3	-2	-2	-3	-3	-3	-3	-3	-3	7	-3	-3
W	-3	-3	-4	-4	-3	-3	-4	-3	-4	-3	-3	-4	-4	-4	-4	-4	-3	-3	8	-3
Y	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	7

Table 18: Matrix showing the effects of the changes in binding energy due to point mutations in protein antibodies using the Rosetta force field. The rows indicate the native residue in the structure and the columns indicate the mutated residue.

	Antigens			Antibodies		
	CHARMM	Amber	Rosetta	CHARMM	Amber	Rosetta
ALA	34	31	70	24	16	45
CYS	13	12	12	2	1	4
ASP	380	324	163	346	208	257
GLU	399	341	173	176	102	108
PHE	46	49	68	37	34	112
GLY	66	62	82	59	38	108
HIS	75	104	98	27	9	23
ILE	32	34	72	23	16	69
LYS	371	345	195	111	58	70
LEU	58	63	108	28	21	73
MET	19	18	29	10	8	11
ASN	191	193	88	119	89	108
PRO	43	53	57	10	7	19
GLN	134	138	87	42	27	52
ARG	338	314	171	310	171	213
SER	152	126	57	228	143	145
THR	112	103	88	102	61	90
VAL	32	38	70	11	15	49
TRP	47	49	48	82	94	182
TYR	89	71	60	402	264	482

Table 19: Total count of each type of residue in the epitopes of the antigens and the paratopes of the antibodies, for each force fields.

References

[1] H. C. Jubb, A. P. Pandurangan, M. A. Turner, B. Ochoa-Montaño, T. L. Blundell, and D. B. Ascher, "Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health," *Progress in Biophysics and Molecular Biology*, vol. 128, pp. 3–13, 2017, doi: https://doi.org/10.1016/j.pbiomolbio.2016.10.002.

[2] A. J. Bordner and R. Abagyan, "Statistical analysis and prediction of protein–protein interfaces," *Proteins: Structure, Function, and Bioinformatics*, vol. 60, no. 3, pp. 353–366, Aug. 2005, doi: 10.1002/prot.20433.

[3] L. Lo Conte, C. Chothia, and J. Janin, "The atomic structure of protein– protein recognition sites," *J Mol Biol*, vol. 285, pp. 2177–2198, 1999, doi: 10.1006/jmbi.1998.2439.

[4] M. Naghavi *et al.*, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013," *The Lancet*, vol. 385, no. 9963, pp. 117–171, 2015, doi: 10.1016/S0140-6736(14)61682-2.

[5] L. Lo Conte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *Journal of Molecular Biology*, vol. 285, no. 5, pp. 2177–2198, Feb. 1999, doi: 10.1006/JMBI.1998.2439.

[6] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, "Hot spots—A review of the protein–protein interface determinant amino-acid residues," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 803–812, Sep. 2007, doi: 10.1002/prot.21396.

[7] R. B. Russell *et al.*, “A structural perspective on protein–protein interactions,” *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 313–324, 2004, doi: https://doi.org/10.1016/j.sbi.2004.04.006.

[8] G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. T. Freer, and P. W. Rose, “Computational detection of the binding-site hot spot at the remodeled human growth hormone–receptor interface,” *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. 2, pp. 201–219, Nov. 2003, doi: 10.1002/prot.10456.

[9] T. Kortemme and D. Baker, “Computational design of protein–protein interactions,” *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 91–97, 2004, doi: https://doi.org/10.1016/j.cbpa.2003.12.008.

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

[11] J. Hu, J. Li, N. Chen, and X. Zhang, “Conservation of hot regions in protein–protein interaction in evolution,” *Methods*, vol. 110, no. 4, pp. 73–80, Nov. 2016, doi: 10.1016/j.ymeth.2016.06.020.

[12] C. Yan, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, “Characterization of Protein–Protein Interfaces,” *The Protein Journal*, vol. 27, no. 1, pp. 59–70, 2008, doi: 10.1007/s10930-007-9108-x.

[13] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, "Hot spots—A review of the protein–protein interface determinant amino-acid residues," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 803–812, Sep. 2007, doi: 10.1002/prot.21396.

[14] B. R. Brooks *et al.*, "CHARMM: the biomolecular simulation program," *Journal of computational chemistry*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[15] D. A. Case *et al.*, "AMBER 2017." University of California, San Francisco, 2017.

[16] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," *Methods in Enzymology*, vol. 383, no. 2003, pp. 66–93, 2004, doi: 10.1016/S0076-6879(04)83004-0.

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

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

[19] M. Li, A. Goncearenco, and A. R. Panchenko, "Annotating Mutational Effects on Proteins and Protein Interactions: Designing Novel and Revisiting Existing Protocols," *Methods in molecular biology (Clifton, N.J.)*, vol. 1550, pp. 235–260, 2017, doi: 10.1007/978-1-4939-6747-6_17.

[20] A. Ghasemi and S. Zahediasl, “Normality tests for statistical analysis: A guide for non-statisticians,” *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, 2012, doi: 10.5812/ijem.3505.

[21] F. A. Fellouse, C. Wiesmann, and S. S. Sidhu, “Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12467–72, 2004, doi: 10.1073/pnas.0401786101.

[22] L. G. Presta, “Molecular engineering and design of therapeutic antibodies,” *Current Opinion in Immunology*, vol. 20, no. 4, pp. 460–470, 2008, doi: https://doi.org/10.1016/j.coi.2008.06.012.

[23] KIM and SJ, “Antibody engineering for the development of therapeutic antibodies,” *Mol Cells*, vol. 20, pp. 17–29, 2005.

[24] T. Igawa *et al.*, “Reduced elimination of IgG antibodies by engineering the variable region,” *Protein Engineering, Design and Selection*, vol. 23, no. 5, pp. 385–392, Feb. 2010, doi: 10.1093/protein/gzq009.

Chapter 4 Analysis of How Mutations Disrupt Hotspot Binding Interactions

Introduction

Early during the COVID-19 pandemic, it was experimentally observed that the anti-SARS antibodies, M396[1], S230[2], and 80R[3] failed to bind to the receptor-binding domain (RBD) of SARS-CoV-2 [4], despite its high degree of similarity to that of SARS-CoV. A study was conducted using computational calculations to understand why this was the case. There were two plausible hypotheses for why the mutations from the SARS-CoV RBD to that of SARS-CoV-2 would disrupt binding: introduction of detrimental interactions or disruption of hotspot interactions, where a hotspot is a residue that is essential to antibody-antigen binding [1], [5] – [8]. It was revealed that disruption of two or three significant interactions at the interface leads to the loss of binding for all three antibodies, i.e., loss of binding was caused by the disruption of hotspot interactions rather than the introduction of detrimental contacts. Although these hypotheses are not exclusive to one another and it was *a priori* anticipated that loss of binding would occur for both reasons, it was somewhat surprising that the calculations revealed that binding was disrupted strictly due to the loss of hotspot interactions for all three antibodies. Collectively, the antibodies lost a total of seven hotspot interactions: two each in M396 and S230 and three in 80R. The disrupted hotspots in M396 are shown in Figure 9, S230 and 80R are shown in Figures 10 and 11, respectively. This part of the analysis was done in collaboration with colleague, Varun Chauhan.

This revelation motivated the study of the deleterious effects of antigen point mutations in antibody-antigen interactions in general. With protein antigens repeatedly evolving to evade antibodies in the immune system, it is important to know how antibody binding is affected by antigen mutations. A selected set of antibody-protein complexes was used for this mutational analysis study. This study was different from the previous mutational analyses taking a more in-

depth look into the interactions of all contact residues in the antigenic interfaces. The findings from this study can help understand which phenomena contribute towards the loss of binding for antibodies with mutated antigens and be directed towards developing strategies to prepare in advance for and respond rapidly to future emerging pandemics.

Methods

Selecting the Complexes

A set of complexes were selected from the non-redundant antibody-antigen database [8], described in the previous chapters using a proportional stratified probabilistic strategy [9], where different metrics that describe the protein-protein interfaces are divided into strata and complexes are selected within each stratum, proportional to the size of the stratum. The study is extensive in regard to the number of residues selected for mutation. Running mutational analyses for the interface residues of all the complexes in the database would take years of computational time. Simultaneously, this study is built on understanding interactions on interfaces, thus metrics particularly describing the interfaces were used to select complexes for this study. The four metrics used were antibody contact residues, antigen contact residues, shape complementarity, and buried surface area.

Contact Residues (CRs) were defined as antibody amino acids with at least one heavy atom (i.e., non-hydrogen) within 4.5Å of an antigen heavy atom, as used in previous studies [10], [11]. For antibodies, it is important to identify the specific residue positions that create contact with the antigen structure, given that they do not necessarily coincide with the residues that participate in binding [12]. When using a threshold of 4Å, the frequency distribution of the number of CR ranges ~ 18 to 19 for both antibodies and antigens [12]. Different threshold

definitions have been used for identifying CRs like 4Å [9] – [13] and 6Å [18], [19]. For this study, a threshold of 4.5Å was used to identify the CRs for both the antibodies and the antigens.

Shape Complementarity (SC) plays a role in complex formation and the intensity of binding and is a key factor in protein interactions [20]. SC is a statistical measure of the "geometric surface complementarity" of protein-protein interfaces. The two factors that affect the measurement are the relative shape of the surfaces with respect to each other and the interactions bringing individual elements of the contrasting surfaces into proximity [21]. Previous studies on protein-protein interfaces show different classes of proteins have different ranges for SC [22], with antibody-antigen interfaces showing poorer shape complementarity than any other type of protein-protein interfaces [21].

Buried Surface Area (BSA) is an estimated measure of the size of the interface between two molecules[23] using the coordinates of the complex, given that areas of the protein surfaces are buried upon association[24]. There is also no correlation between BSA and Gibbs free energy of dissociation [24], as understood from the study of protein-protein interfaces. Yet, BSA has been demonstrated to show correlations to the overall flexibility of the proteins [25].

Figures 1 to 4 show the distribution of the antibody CRs, antigen CRs, SC, and BSA for the 384 antibody-antigen complexes in the database. The set of complexes used for this study is only a subset of the complexes used in the previous study. The vast extent of residues selected for each complex for the mutational analysis makes it not feasible to run the mutations for all the contact residues in all the 384 complexes; thus, 70 complexes were selected for this study.

Assessing the Protein-Protein Interactions

For the selected complexes, the frequency distribution of the types of amino acids interacting at the interfaces was visualized with heat maps. Further, the frequency distributions of the amino acids for the best five interactions and the worst three interactions were visualized. The study of the antibody-antigen interfaces in Chapter 2 revealed that the best five interactions are enough to describe around 50% of the interaction energy and the worst three interactions are enough to describe the largest detrimental interactions on the binding interface. Since this work is based on the same database, the best five and worst three interactions were observed.

Mutational Analysis of the Antigen Residues

The following calculations were independently repeated for the antigen contact residues of the selected complexes using three molecular mechanics force fields: CHARMM[26], Amber[27], and Rosetta[28], each. For each residue, a mutation of the native amino acid to each of the other 19 common amino acids was created, the energy of the mutated complex was minimized, and the pairwise interaction energies for the contact residues were calculated. The mutations were made using a rotamer library used in previous mutational analysis studies. The energies of the interaction of the top five and worst three residues in each complex were calculated and compared to the antibody database average. The means of the top five interactions of the mutated complexes were compared to the corresponding means in the antibody database complexes. The comparisons were done using unpaired t-test analyses. Similar comparisons were made for the worst three interactions.

Results and Discussions

The question of how antibody interactions are affected by the introduction of mutated antigens is the focal point of this project. The 70 complexes selected for this project were 1A14, 1ADQ, 1AFV, 1DQJ, 1FBI, 1FSK, 1G9M, 1HYS, 1OP9, 1OTS, 1UJ3, 2ADF, 2OZ4, 2P4A,

2Q8A, 2YBR, 2YPV, 2ZCH, 3AB0, 3DVG, 3EFF, 3EZJ, 3G04, 3J70, 3J8Z, 3JBQ, 3LD8, 3NH7, 3NPS, 3SQO, 3V0A, 3W14, 3WKM, 4EDW, 4FFY, 4GRW, 4HGK, 4I18, 4IDJ, 4K4M, 4K94, 4KRP, 4KUC, 4LEO, 4LSU, 4QNP, 4RWY, 4UV7, 4V1D, 4WEM, 4XMM, 4XMN, 4XNY, 4YDL, 4YJZ, 4ZS6, 5BV7, 5C6T, 5C7X, 5D72, 5D8J, 5D93, 5DUM, 5F45, 5FV2, 5HVG, 5IKC, 5K59, 5KVD, and 5KVF.

Tables 20 and 21 are heat maps of the frequency of the antigen residues interacting with antibody residues. The mutational analysis results provide residue-residue interaction information. This map was created with the frequency of each interaction of one type of amino acid with another. It is to be noted that the size (thus, significance) of the interaction was not a factor in this frequency distribution, if there was an interaction, however small, it was counted. The frequency of use for the different amino acids in the antigen residues is distributed, but the antibody residues show more preferred use of certain amino acids. Similar trends were noticed in the analysis of the interfaces in Chapter 2. Antibody interfaces have preferences for Tyrosine, Arginine, Serine [29] – [32] , with more polar residues at the interface[33].

Tables 3 and 4 are heat maps of the best five interactions at the interfaces. It can be observed that interactions between specific types of amino acids being more prevalent in the hotspot residues. The most frequent interactions are between Aspartic Acid, Lysine, Arginine, Glutamic Acid, Serine, Tyrosine. This demonstrates that specific types of residues are more important to binding and contribute the most to binding energy [32]. Indeed, hotspot residues are known to rely on certain defined geometric and chemical complementary properties [29]. The results from the tables show the relative importance of electrostatic interactions [35], hydrophobic interactions [36], hydrogen bonds, and salt bridges [37] at the interfaces.

Tables 24 and 25 are heat maps of the worst three interactions at the interfaces. It can be observed that specific amino acids are more prone to clashes with each other. This distribution in frequency may be more contributed to the non-complementarity of the geometric and chemical properties of amino acids at the interfaces. It is observed that similar types of polar residues are clashing with each other and residues with larger surface areas are clashing more with each other and with other residues. The frequency of these types of residues appearing more on the interface may play a role in the distributions seen for both the best and worst interactions. The more they appear, the more they are going to interact, in both beneficial or detrimental ways.

Table 26, 27, and 28 compares the best five interaction energies and the worst three interaction energies from the mutational analysis to those of the database. The means of all the five best energies in the mutated complexes were compared to the corresponding means in the antibody database used in Chapter 2. The unpaired t-test did not show any pattern for the comparison of the means. The null hypothesis was that there was no significant difference in the means, the hypothesis was accepted and rejected at different levels of interactions. This can be interpreted as no one phenomena that works in disrupting binding when the antigen residues mutate at the interface. The case of the anti-SARS-CoV antibodies losing binding to the SARS-CoV-2 RBD was a coincidence.

This project was built to study the changes in PPIs at the antibody-antigen interfaces with point mutations. PPIs are important for all fundamental biological processes and are of much higher concern at the mutation enriched pathogen interfaces. Pre-pandemic, our understanding of antibody-antigen interactions was that when an antigen mutates, the binding to the antibody was lost either due to the loss of hotspot interactions or to the introduction of detrimental interactions. When the anti-SARS-CoV antibodies could not bind to the SARS-CoV-2 RBD, the expectation

was that both phenomena would be observed at the interface. Yet, the computational study revealed that binding was lost solely due to the loss of hotspot interactions. As unexpected as this was, it was also reason enough to question whether there is a single phenomenon that occurs more often in antibody binding when antigens mutate.

The results were consistent with what was understood about antibodies binding to mutated antigens, that there is no consistent pattern that is followed in general. The conclusion interpreted from the results was that both phenomena worked at different degrees in all cases. More precisely, when a specific antigen mutates, the loss of binding to antibodies will need to be studied individually, instead of making a bulk assumption. While the conclusions were consistent with what was understood pre-pandemic, this project added to deepen our understanding of the effect of mutations on binding interfaces and gives more aspects of the interfaces that may need to be studied.

The heat maps that were built from residue-residue interactions show consistency with the results from Chapter 2. This was expected as the same force fields and parameters were used, and the complexes selected were a subset of the complexes in the same database. Yet, the heat maps also give additional ideas to where this project could be further developed. One such idea would be to study specific interactions on interfaces and observe their changes with mutations. The binding mechanisms of PPIs has more avenues to be studied in terms of therapeutic developments. Studying the relationship between the structures and binding mechanisms of PPIs is crucial to understanding how mutations can lead to novel pathogens that can evade the immune systems.

Figure 10: Disrupted Hotspot Interactions in Antibody M396 with the SARS-CoV-2 RBD. Calculations revealed that M396 losses binding to the RBD of SARS-CoV-2 due to the disruption of two hotspot interactions. In all panels, the SARS-CoV RBD is shown in green, the SARS-CoV-2 RBD is shown in orange, and the M396 antibody is shown in cyan. Antibody residues are numbered according to the International Immunogenetics Information System (IMGT®) numbering scheme. A shows how mutation K403R orients the oxygen atoms of D405 away from R408, causing R408 to make multiple hydrogen bonds with Q414 and leading to the loss of a salt bridge with light chain. In addition, B illustrates how mutation I503V weakens the strong hydrophobic interaction with the light chain W107.

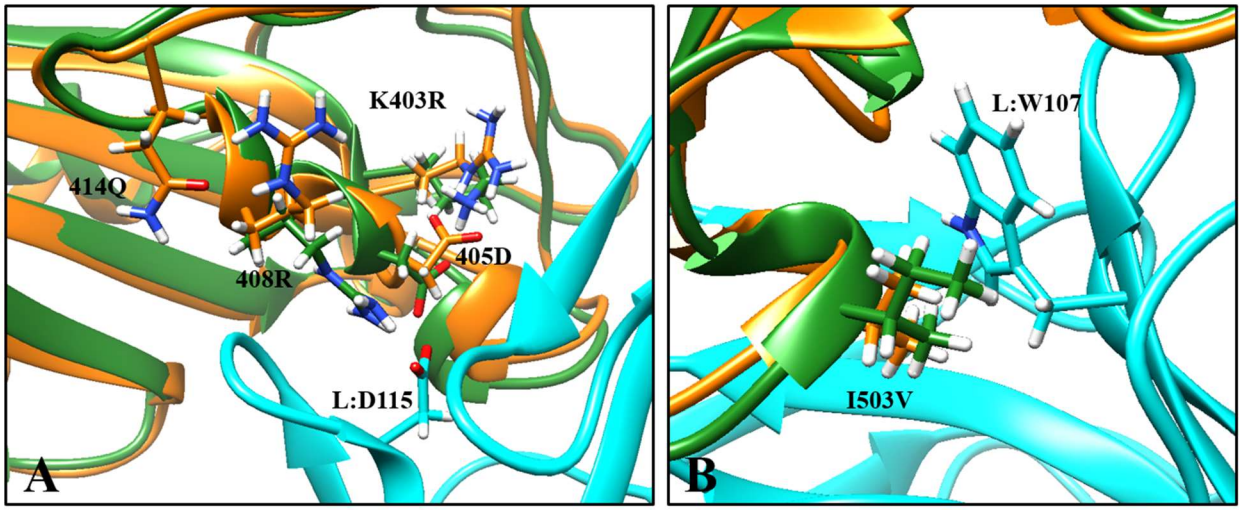


Figure 11: Disrupted Hotspot Interactions in Antibody S230 with the SARS-CoV-2 RBD. Calculations revealed that S230 losses binding to the RBD of SARS-CoV-2 due to the disruption of two hotspot interactions. In all panels, the SARS-CoV RBD is shown in green, the SARS-CoV-2 RBD is shown in orange, and the S230 antibody is shown in cyan. Antibody residues are numbered according to the International Immunogenetics Information System (IMGT®) numbering scheme. In A, mutation K478T breaks the salt bridge with H: 69D. In addition, B illustrates how mutation K397N leads to the loss of a salt bridge with light chain D34.

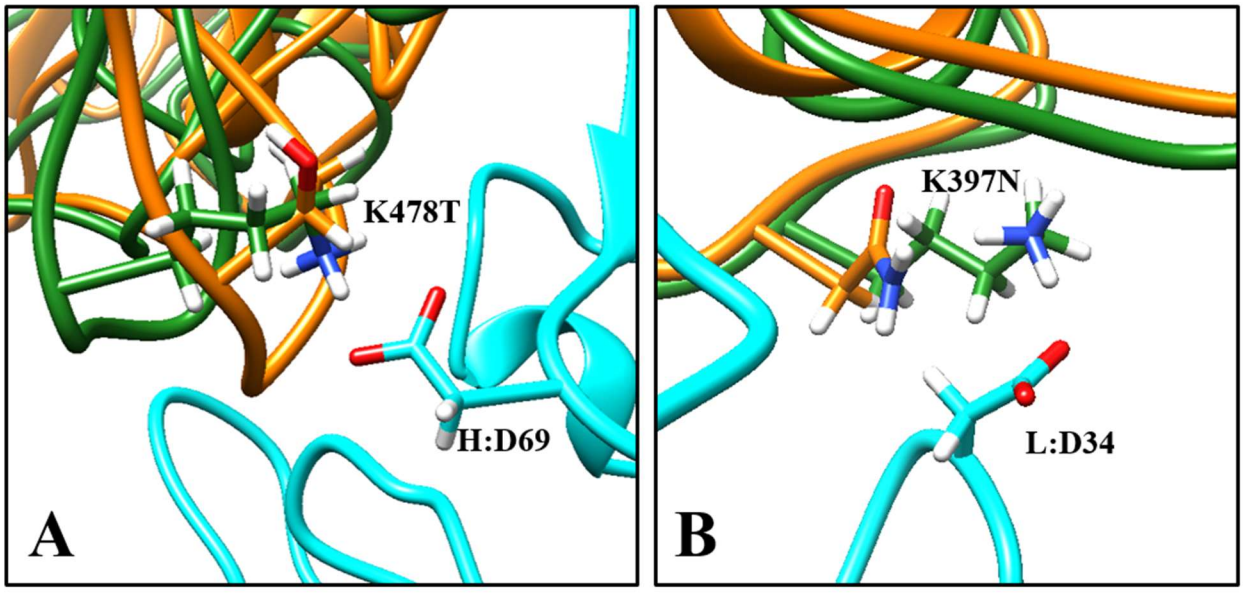


Figure 12: Disrupted Hotspot Interactions in Antibody 80R with the SARS-CoV-2 RBD. Calculations revealed that 80R loses binding to the SARS-CoV-2 RBD due to the disruption of three hotspot interactions. In all panels, the SARS-CoV RBD is shown in green, the SARS-CoV-2 RBD is shown in orange, and the 80R antibody is shown in cyan. In **A**, mutation D494S breaks the salt bridge with L: R36. **B** shows how mutation Y498Q removes the strong π-π interaction with H: Y113. Finally, **C** illustrates how mutation R439N leads to the loss of a salt bridge with

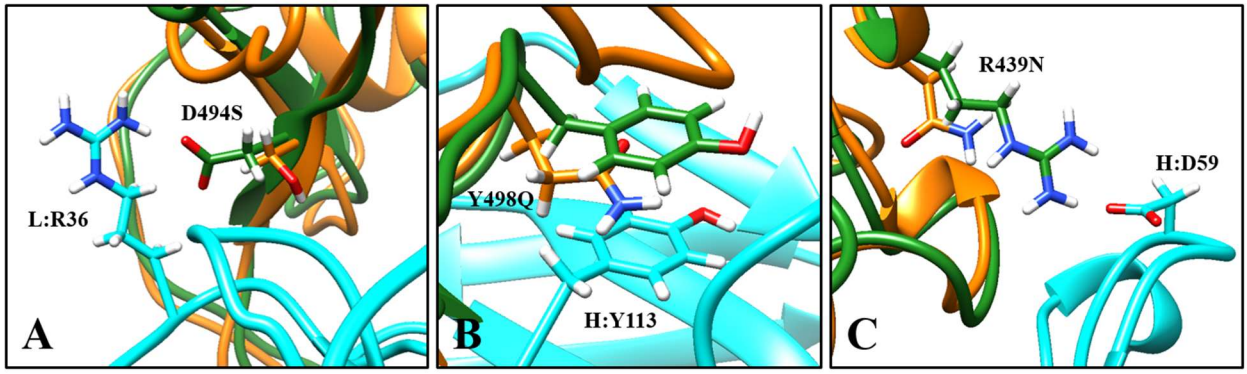


Figure 13: Distribution of the antibody contact residues in all the structures. The mean was 31.26 with a standard deviation of 7.91

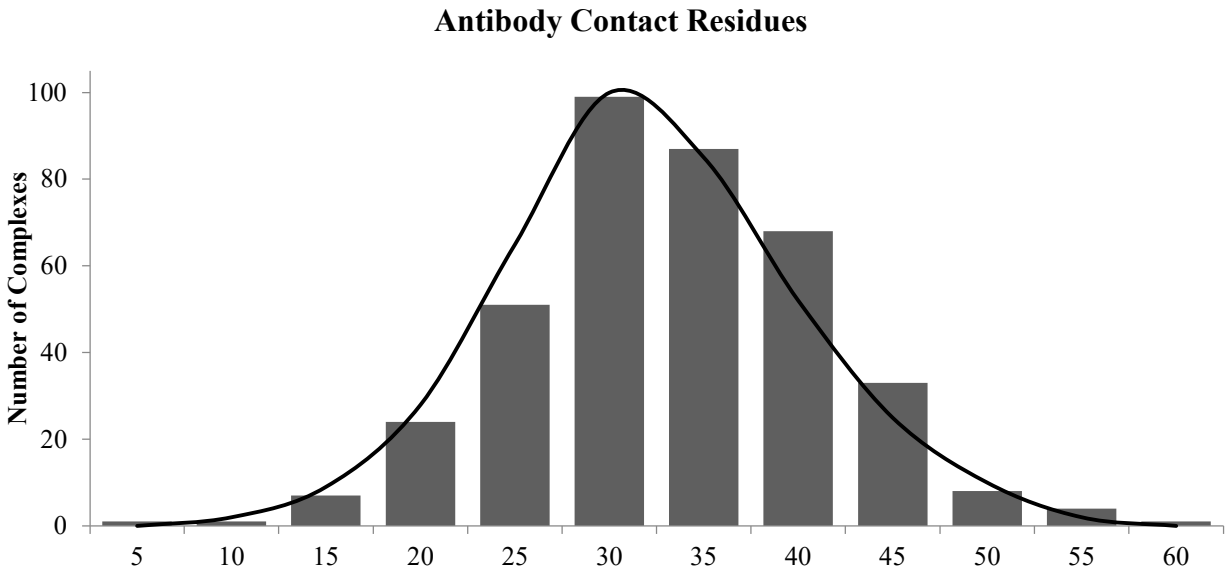


Figure 14: Distribution of the antigen contact residues in all the structures. The mean was 30.24 with a standard deviation of 9.42.

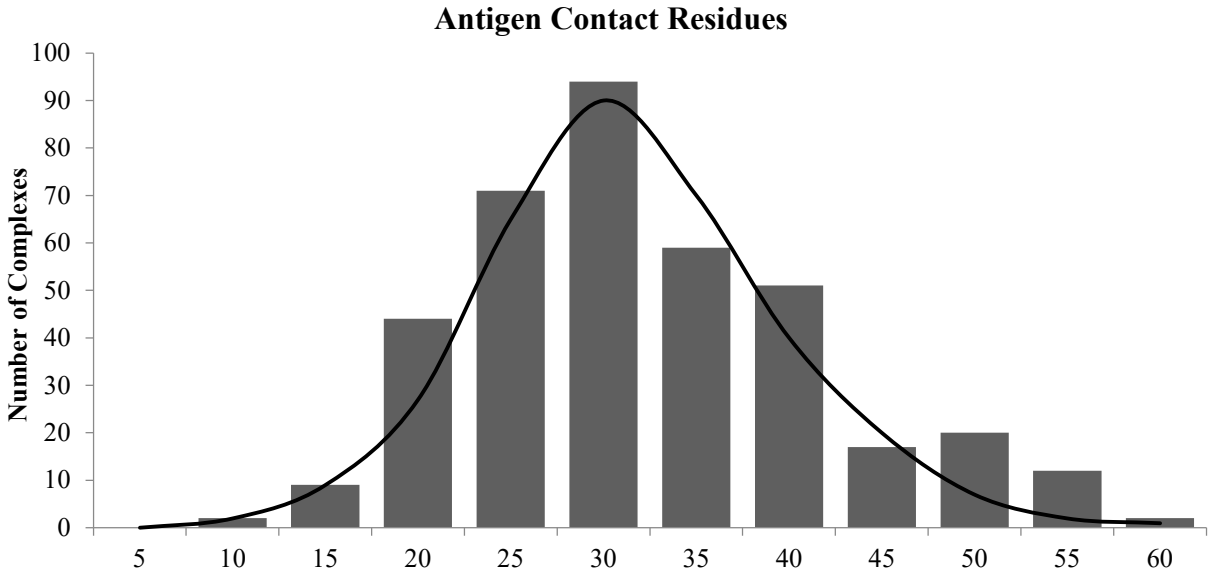


Figure 15: Distribution of the shape complementarity in all the structures. The mean was 0.672 with a standard deviation of 0.07.

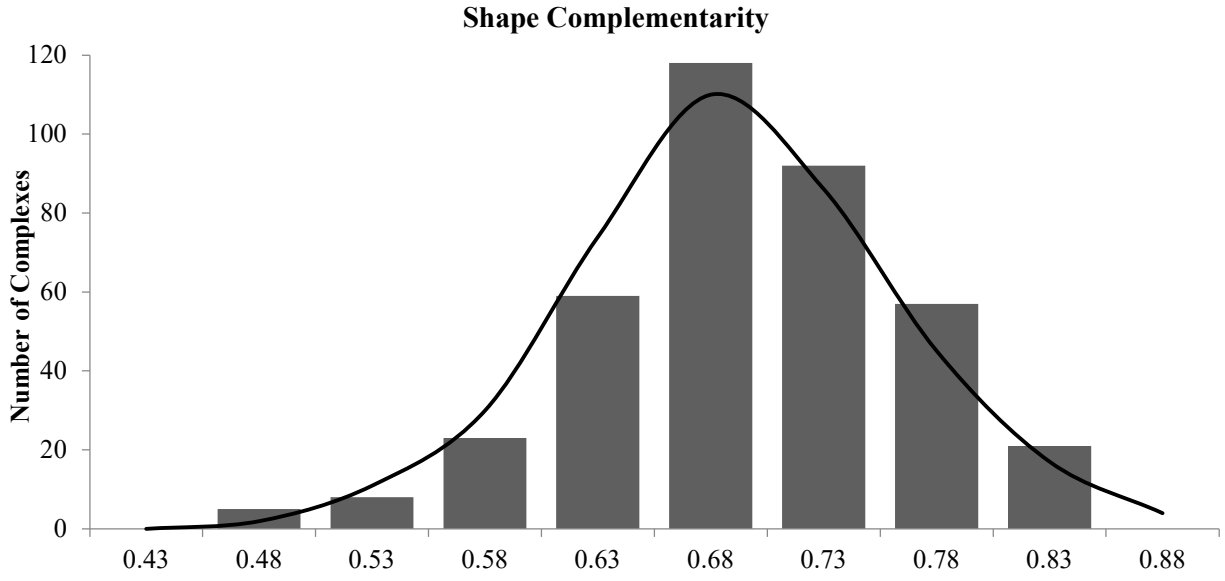


Figure 16: Distribution of the Buried Surface Area in all the structures. The mean was 1859.44 with a standard deviation of 495.25

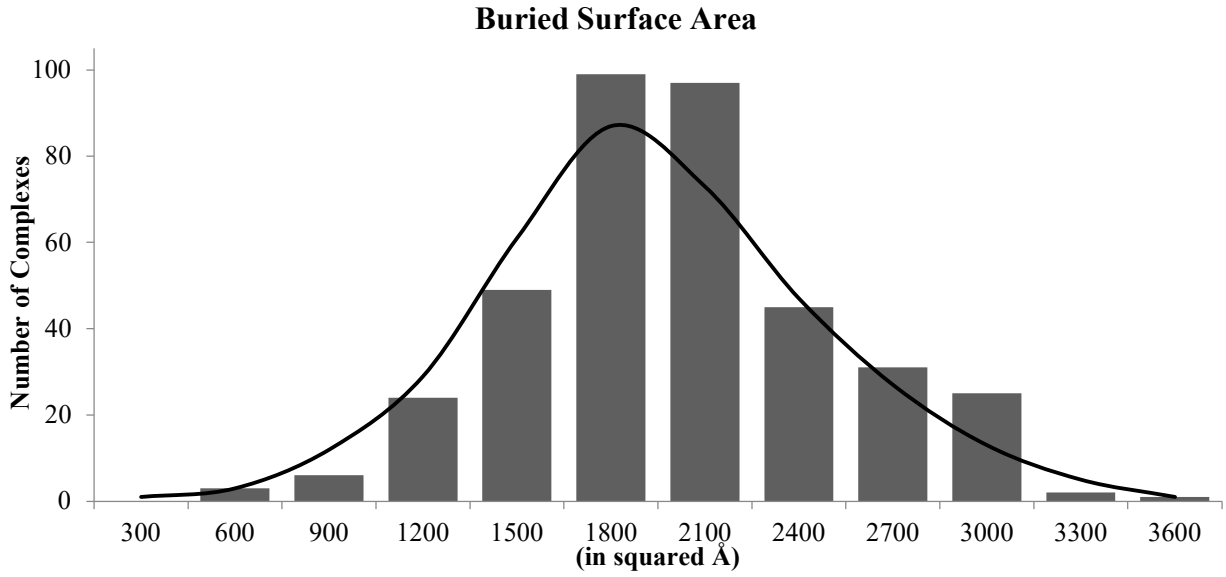


Table 20: Heat Map showing all interactions of antibody-antigen interfaces in the selected complexes with the CHARMM force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. The amino acids most frequently interacting for antibody residues are Tyrosine, Serine, Arginine, Aspartic Acid, and Glycine. There are no distinct amino acids that are frequently interacting in the antigen residues.

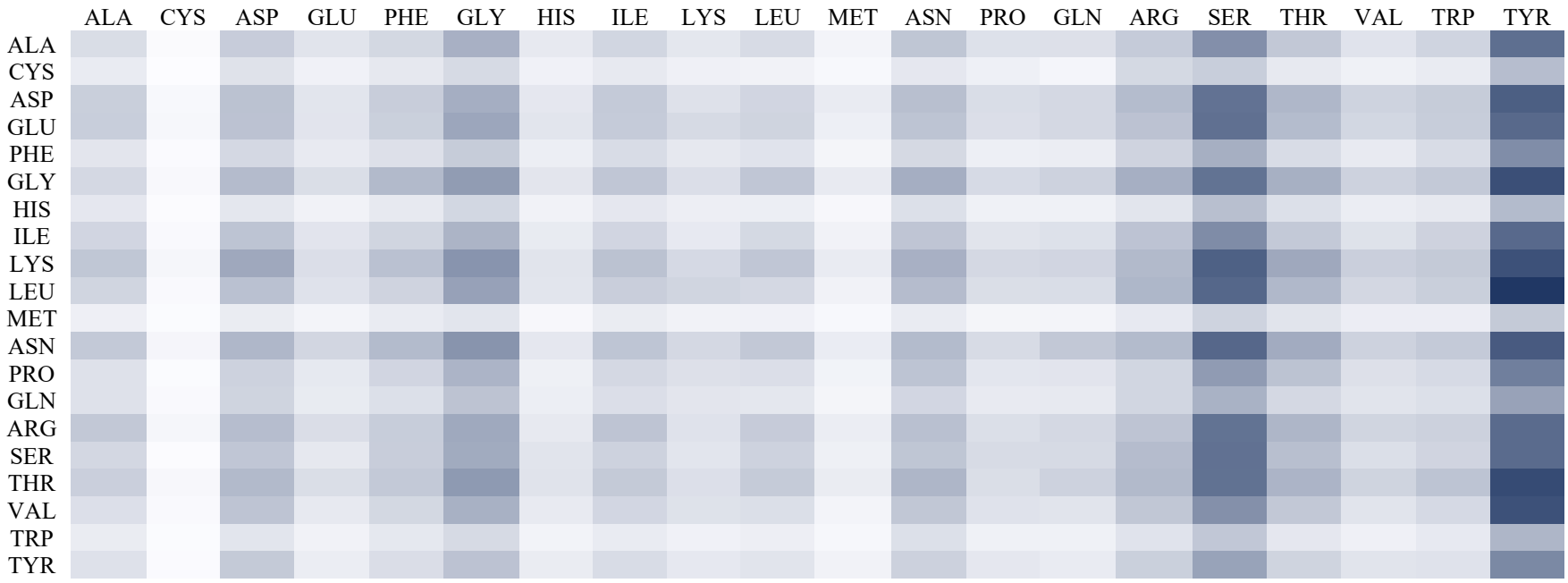


Table 21: Heat Map showing all interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. The amino acids most frequently interacting for antibody residues are Tyrosine, Serine, Arginine, Asparagine, and Glycine. There are no distinct amino acids that are frequently interacting in the antigen residues.

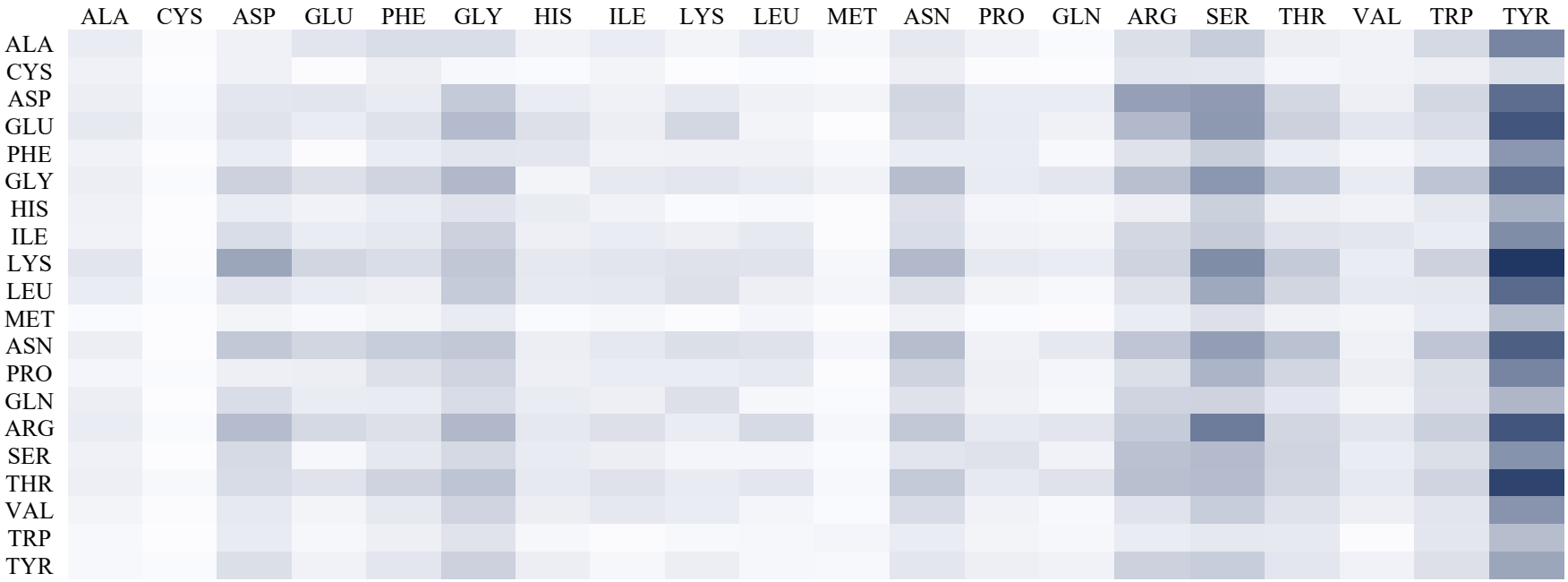


Table 22: Heat Map showing the best five interactions of antibody-antigen interfaces in the selected complexes with a CHARMM force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. The most prominent interactions are between Aspartic Acid, Glutamic Acid, Lysine, Arginine, Serine, and Tyrosine.

Table 23: Heat Map showing the best five interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. The most prominent interactions are between Aspartic Acid, Glutamic Acid, Lysine, Arginine, and Tyrosine.

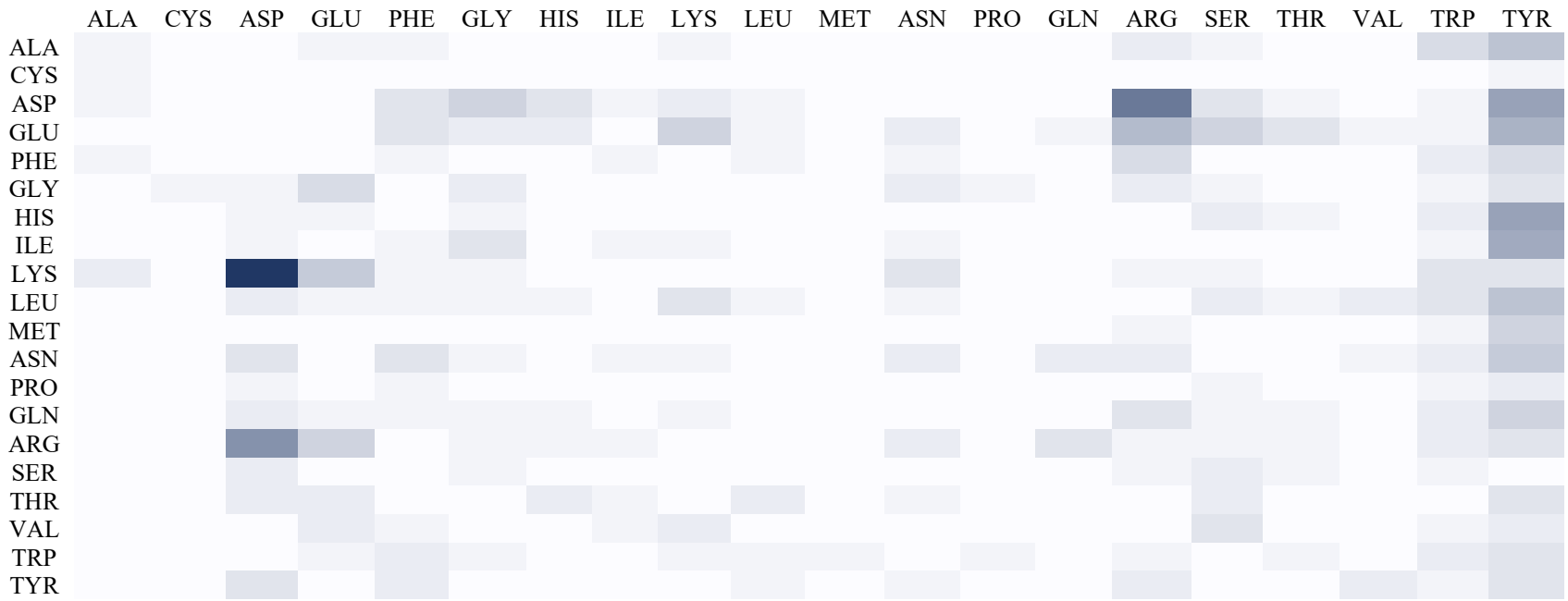


Table 24: Heat Map showing the worst three interactions of antibody-antigen interfaces in the selected complexes with the CHARMM force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. The worst clashes are between Aspartic Acid, Glutamic Acid, Lysine, Arginine, and Tyrosine.

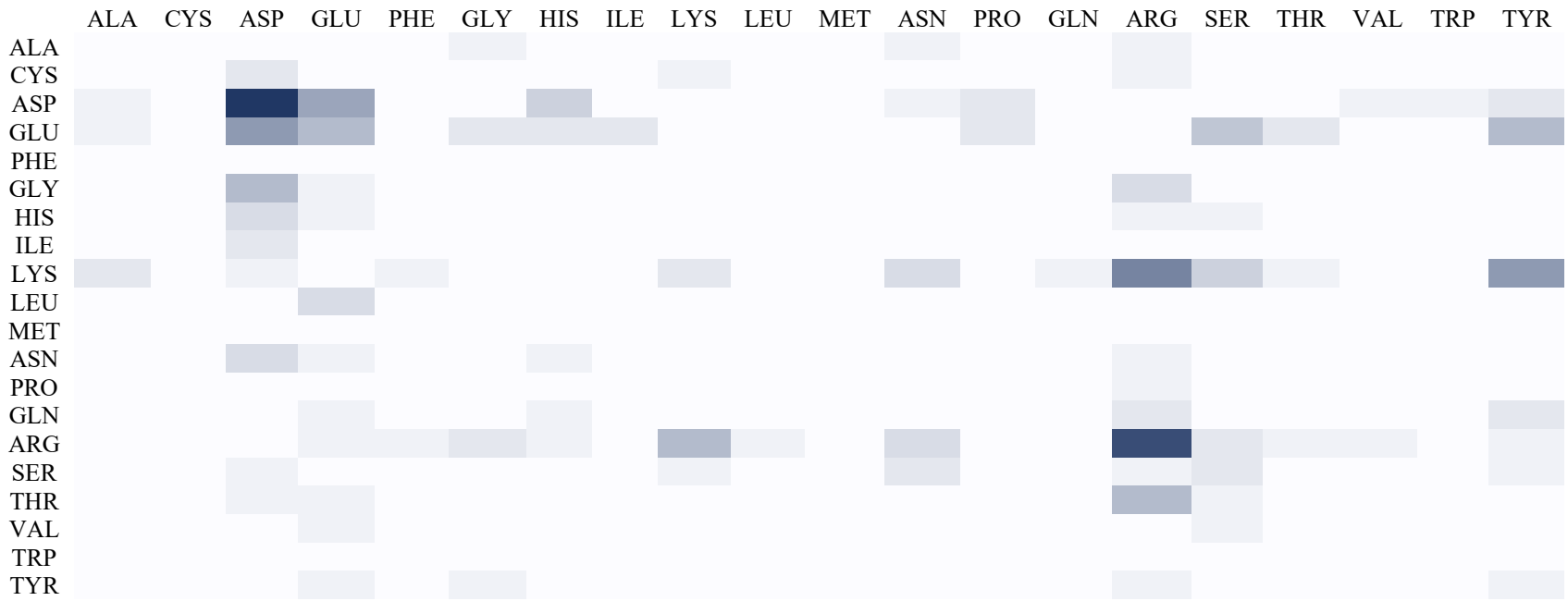


Table 25: Heat Map showing the worst three interactions of antibody-antigen interfaces in the selected complexes with Rosetta force field. The rows are the antigen residues and the columns are antibody residues. The pigmentation of the cell indicates no interactions to high number of interactions from white to navy blue. The worst clashes are between Aspartic Acid, Glutamic Acid, Lysine, Arginine, Serine, and Tyrosine.

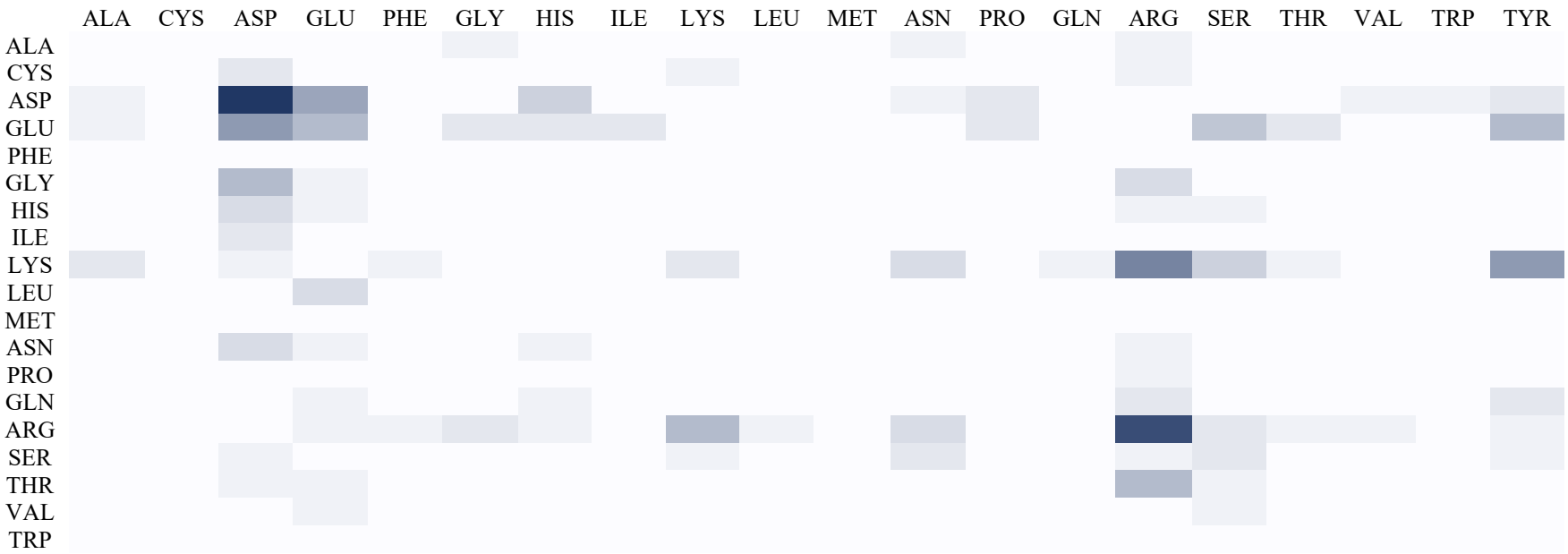


Table 26: Mutational analysis results compared to the corresponding values in the antibody-antigen database for the CHARMM force field. The top five interaction energies and the worst three interactions are compared using unpaired t-test analysis. The value of significance (α) is 0.05 for the test and the null hypothesis (H_0) is that there is no significant difference in the means. The sample size for the antibody database is 384 and for the mutational average is 42,200.

			t-Test Analysis		
	Antibody Database Average	**Mutational Analysis Average**	**t-Statistic**	**t-Critical**	**p-value**
Top five interaction energies (kcal/mol)	-41.75 ± 13.63	-42.41 ± 13.08	-0.956	1.966	0.340 (H_0 is not rejected)
	-30.81 ± 12.60	-32.41 ± 12.59	-2.468	1.966	0.014 (H_0 is rejected)
	-24.12 ± 10.50	-26.03 ± 11.64	-3.533	1.966	4.600E-04 (H_0 is rejected)
	-20.28 ± 9.41	-21.70 ± 10.40	-2.933	1.966	0.004 (H_0 is rejected)
	-16.97 ± 7.80	-18.30 ± 8.94	-3.313	1.966	0.001 (H_0 is rejected)
Worst three interaction energies (kcal/mol)	5.34 ± 4.64	4.38 ± 3.50	-1.436	1.966	0.152 (H_0 is not rejected)
	3.12 ± 2.68	2.45 ± 2.78	-2.766	1.966	0.006 (H_0 is rejected)
	2.18 ± 1.91	1.48 ± 1.80	-1.337	1.966	0.182 (H_0 is not rejected)

Table 27: Mutational analysis results compared to the corresponding values in the antibody-antigen database for the Amber force field. The top five interaction energies and the worst three interactions are compared using unpaired t-test analysis. The value of significance (α) is 0.05 for the test and the null hypothesis (H_0) is that there is no significant difference between the two means. The sample size for the antibody database is 384 and for the mutational average is 42,200.

			t-Test Analysis		
	Antibody Database Average	**Mutational Analysis Average**	**t-Statistic**	**t-Critical**	**p-value**
Top five interaction energies (kcal/mol)	-13.99 ± 4.27	-14.24 ± 4.51	-1.118	1.966	0.264 (H_0 is not rejected)
	-10.62 ± 3.79	-10.50 ± 3.88	0.646	1.966	0.519 (H_0 is not rejected)
	-8.50 ± 3.20	-8.73 ± 3.54	-1.404	1.966	0.161 (H_0 is not rejected)
	-7.24 ± 2.82	-7.62 ± 3.29	-2.644	1.966	0.009 (H_0 is rejected)
	-6.17 ± 2.40	-6.35 ± 2.64	-1.446	1.966	0.149 (H_0 is not rejected)
Worst three interaction energies (kcal/mol)	1.41 ± 1.33	1.73 ± 1.66	4.420	1.966	1.281E-05 (H_0 is rejected)
	0.83 ± 0.78	1.03 ± 1.25	4.473	1.966	1.012E-05 (H_0 is rejected)
	0.58 ± 0.51	0.70 ± 0.95	3.627	1.966	3.243E-04 (H_0 is rejected)

Table 28 Mutational analysis results compared to the corresponding values in the antibody-antigen database for the Rosetta force field. The top five interaction energies and the worst three interactions are compared using unpaired t-test analysis. The value of significance (α) is 0.05 for the test and the null hypothesis (H_0) is that there is no significant difference between the two means. The sample size for the antibody database is 384 and for the mutational average is 42,200.

			t-Test Analysis		
	Antibody Database Average	**Mutational Analysis Average**	**t-Statistic**	**t-Critical**	**p-value**
Top five interaction energies (kcal/mol)	-4.45 ± 0.83	-4.35 ± 0.96	2.211	1.966	0.028 (H_0 is rejected)
	-3.77 ± 0.71	-3.68 ± 0.81	2.513	1.966	0.012 (H_0 is rejected)
	-2.91 ± 0.61	-3.26 ± 0.71	1.562	1.966	0.119 (H_0 is not rejected)
	-2.98 ± 0.56	-2.91 ± 0.61	2.512	1.966	0.012 (H_0 is rejected)
	-2.73 ± 0.52	-2.66 ± 0.58	2.562	1.966	0.010 (H_0 is rejected)
Worst three interaction energies (kcal/mol)	0.91 ± 0.54	1.15 ± 0.64	2.420	1.966	0.128 (H_0 is not rejected)
	0.60 ± 0.32	0.87 ± 0.56	2.473	1.966	0.015 (H_0 is rejected)
	0.45 ± 0.25	0.26 ± 0.23	3.270	1.966	0.243 (H_0 is not rejected)

References

[1] P. Prabakaran *et al.*, "Structure of Severe Acute Respiratory Syndrome Coronavirus Receptor-binding Domain Complexed with Neutralizing Antibody," *Journal of Biological Chemistry* , vol. 281, no. 23, pp. 15829–15836, Jun. 2006, doi: 10.1074/jbc.M600697200 .

[2] Z. Zhu *et al.*, "Potent cross-reactive neutralization of SARS coronavirus isolates by human monoclonal antibodies," *Proceedings of the National Academy of Sciences*, vol. 104, no. 29, pp. 12123 LP – 12128, Jul. 2007, doi: 10.1073/pnas.0701000104.

[3] W. C. Hwang *et al.*, "Structural Basis of Neutralization by a Human Anti-severe Acute Respiratory Syndrome Spike Protein Antibody, 80R," *Journal of Biological Chemistry* , vol. 281, no. 45, pp. 34610–34616, Nov. 2006, doi: 10.1074/jbc.M603275200 .

[4] D. Wrapp *et al.*, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation," *Science (1979)*, p. eabb2507, Feb. 2020, doi: 10.1126/science.abb2507.

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

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

[7] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of Molecular Biology*, vol. 280, no. 1, pp. 1–9, 1998, doi: 10.1006/jmbi.1998.1843.

[8] V. M. Chauhan, S. Islam, A. Vroom, and R. Pantazes, “Development and Analyses of a Database of Antibody – Antigen Complexes,” *Computer Aided Chemical Engineering*, vol. 44, pp. 2113–2118, Jan. 2018, doi: 10.1016/B978-0-444-64241-7.50347-5.

[9] J. Martínez-Mesa, D. A. González-Chica, R. P. Duquia, R. R. Bonamigo, and J. L. Bastos, “Sampling: how to select participants in my research study?,” *An Bras Dermatol*, vol. 91, no. 3, pp. 326–330, 2016, doi: 10.1590/abd1806-4841.20165254.

[10] M. M. Meyer *et al.*, “Library analysis of SCHEMA-guided protein recombination,” *Protein Sci*, vol. 12, no. 8, pp. 1686–1693, Aug. 2003, doi: 10.1110/ps.0306603.

[11] R. J. Pantazes, M. C. Saraf, and C. D. Maranas, “Optimal protein library design using recombination or point mutations based on sequence-based scoring functions,” *Protein Engineering, Design and Selection*, vol. 20, no. 8, pp. 361–373, 2007, doi: 10.1093/protein/gzm030.

[12] J. W. Stave and K. Lindpaintner, “Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure,” *The Journal of Immunology*, vol. 191, no. 3, pp. 1428 LP – 1435, Aug. 2013, doi: 10.4049/jimmunol.1203198.

[13] J. Novotný *et al.*, “Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains),” *Proceedings of the National Academy of Sciences*, vol. 83, no. 2, pp. 226 LP – 230, Jan. 1986, doi: 10.1073/pnas.83.2.226.

[14] P. Haste Andersen, M. Nielsen, and O. Lund, “Prediction of residues in discontinuous B-cell epitopes using protein 3D structures,” *Protein Science*, vol. 15, no. 11, pp. 2558–2567, Nov. 2006, doi: 10.1110/ps.062405906.

[15] J. V Ponomarenko and P. E. Bourne, “Antibody-protein interactions: benchmark datasets and prediction tools evaluation,” *BMC Structural Biology*, vol. 7, no. 1, p. 64, 2007, doi: 10.1186/1472-6807-7-64.

[16] H. R. Ansari and G. P. S. Raghava, “Identification of conformational B-cell Epitopes in an antigen from its primary sequence,” *Immunome Research*, vol. 6, no. 1, p. 6, 2010, doi: 10.1186/1745-7580-6-6.

[17] L. Zhao and J. Li, “Mining for the antibody-antigen interacting associations that predict the B cell epitopes,” *Structural Biology*, vol. 10, no. Suppl 1, pp. 1–13, 2010, doi: 10.1186/1472-6807-10-S1-S6.

[18] A. Schlessinger, Y. Ofran, G. Yachdav, and B. Rost, “Epitome: database of structure-inferred antigenic epitopes,” *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D777–D780, Jan. 2006, doi: 10.1093/nar/gkj053.

[19] V. Kunik, B. Peters, and Y. Ofran, “Structural consensus among antibodies defines the antigen binding site,” *PLoS Comput Biol*, vol. 8, no. 2, pp. e1002388–e1002388, 2012, doi: 10.1371/journal.pcbi.1002388.

[20] Y. Li, X. Zhang, and D. Cao, “The Role of Shape Complementarity in the Protein-Protein Interactions,” *Scientific Reports*, vol. 3, p. 3271, Nov. 2013.

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

[22] 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.

[23] C. Chothia and J. Janin, “Principles of protein–protein recognition,” *Nature*, vol. 256, no. 5520, pp. 705–708, 1975, doi: 10.1038/256705a0.

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

[25] J. A. Marsh, “Buried and Accessible Surface Area Control Intrinsic Protein Flexibility,” *Journal of Molecular Biology*, vol. 425, no. 17, pp. 3250–3263, 2013, doi: https://doi.org/10.1016/j.jmb.2013.06.019.

[26] 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.

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

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

[29] C. Yan, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, “Characterization of Protein–Protein Interfaces,” *The Protein Journal*, vol. 27, no. 1, pp. 59–70, 2008, doi: 10.1007/s10930-007-9108-x.

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

[31] J. Hu, J. Li, N. Chen, and X. Zhang, “Conservation of hot regions in protein–protein interaction in evolution,” *Methods*, vol. 110, no. 4, pp. 73–80, Nov. 2016, doi: 10.1016/j.ymeth.2016.06.020.

[32] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, “Hot spots—A review of the protein–protein interface determinant amino-acid residues,” *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 803–812, Sep. 2007, doi: 10.1002/prot.21396.

[33] E. D. Levy, “A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution,” *Journal of Molecular Biology*, vol. 403, no. 4, pp. 660–670, 2010, doi: https://doi.org/10.1016/j.jmb.2010.09.028.

[34] C. Tim and W. J. A., “A Hot Spot of Binding Energy in a Hormone-Receptor Interface,” *Science (1979)*, vol. 267, no. 5196, pp. 383–386, Jan. 1995, doi: 10.1126/science.7529940.

[35] H. B. M. Shashikala, A. Chakravorty, and E. Alexov, “Modeling Electrostatic Force in Protein-Protein Recognition ,” *Frontiers in Molecular Biosciences* , vol. 6. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmolb.2019.00094

[36] C.-J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, “Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect,” *Protein Science*, vol. 6, no. 1, pp. 53–64, Jan. 1997, doi: 10.1002/pro.5560060106.

[37] D. Xu, C. J. Tsai, and R. Nussinov, “Hydrogen bonds and salt bridges across protein-protein interfaces.,” *Protein Engineering, Design and Selection*, vol. 10, no. 9, pp. 999–1012, Sep. 1997, doi: 10.1093/protein/10.9.999.

Chapter 5 The Effects of SARS-CoV-2 Spike Protein Mutations on the Immune System

Introduction

SARS-CoV-2 has proved to be a fast-mutating virus and further changes to its proteins are likely. Predicting how the immune system would be able to protect the human body when newer variants evolve is a key step to "getting ourselves ahead of the game". In the previous three chapters, the effects of mutations on protein binding and interactions for antibodies have been studied. Mutations in antigens affect other proteins in the immune system beyond antibodies. This chapter studies the effects of viral mutations on the binding for other therapeutic proteins in the immune system.

Antigens that are circulating in the bloodstream are processed by antigen presenting cells (APCs) into short peptides and presented to the appropriate receptors. APCs have specific Major histocompatibility complex (MHC) molecules that interact with T cells [1]. MHC proteins, also known as human leukocyte antigens (HLA), are glycoproteins that bind with peptides (epitopes) derived from pathogen proteins and present them for inspection by T-cells. Epitope recognition by T cells is fundamental to the adaptive immune system for the host to identify and respond to antigens [2]. There are two classes of MHC proteins: Class I and Class II. MHC Class I molecules bind to CD8+ cytotoxic T cells and MHC Class II bind to CD4+ helper T cells. MHC molecules are known to be polymorphic thus binding to different connectors. MHC Class I proteins are encoded by three loci: HLA-A, HLA-B, and HLA-C. MHC Class II proteins also are encoded by three loci: HLA-DR, HLA-DQ, and HLA-DP. The peptide binding site of Class I proteins has a closed cleft allowing only short peptides (8 to 11 residues) to bind in an extended conformation. In contrast, the cleft of Class II proteins is open-ended, allowing much longer peptides to bind, with the caveat that only 9 residues can occupy the site. The SARS-CoV-2 viral

peptides show more interaction with helper T cells than cytotoxic T cells [3], making MHC Class II proteins an object of interest in the therapeutic developments against the virus.

Apart from T-cells, another important part of the immune system is B-cells. Similar to T-cells, B-cells have receptors that will connect to an antigen shape and each B cell produces a single species of antibody with a unique antigen-binding site. B-cells can connect to antigens right on the surface of the invading virus or bacteria. The B-cell epitopes are generally discontinuous, and any residue capable of being in contact with an antibody can be a part of a B-cell epitope. The types of B cells in COVID patients are naive non-isotype-switch, memory, and antibody-secreting [3].

The structure of the Spike protein is vital for understanding the interaction between the host cell and SARS-CoV-2. The Spike protein is made up of two subunits, S1 and S2, that control the binding to and fusion into the cell, respectively[4]. The S1 subunit, shown in Figure 16, controls the binding to the host cell receptors. This portion of the Spike protein has been targeted for vaccine and targeted drug design[5]. The S1 subunit is split into three main parts: signal peptide (SP), N terminal domain (NTD), and the receptor-binding domain (RBD). The SP is a short hydrophobic peptide that transport the protein to the membrane destination[6]. The RBD binds to ACE2 receptor on the host cells; this binding triggers the viral fusion event in the S2 subunit. The S2 subunit contains the fusion peptides (FP), heptapeptide repeat domains (HR-1 and HR-2), transmembrane (TM) domain, and cytoplasmic tail (CT)[6]. Entry into the cell is initiated by three cleavage events. The multiple cleavage points increase infectability due to the increased probability of being cleaved. The two primary cleavage events are splitting the S1 and S2 subunit and cleaving S2 into Fusion Peptide (FP) and S2'[7]. The cleaving of S1 and S2 is

necessary for viral fusion to occur. The cleavage point that is unique to SARS-CoV-2 is an additional point between the S1 and S2 subunit[5].

Understanding the effect of mutations on the B-cell and T-cell epitopes can provide knowledge on how the different variants of the virus evade the immune system. In this chapter, a mutational analysis is made on the changes in binding of peptides in the MHC Class II binding grooves. Simultaneously, a relationship between the HLA allele distribution in ethnicities to the binding property changes of the MHC Class II proteins to the mutated Spike protein is built. To study the B-cell epitopes, the Solvent Accessible Surface Area (SASA) of the residues in the Spike protein is used. Residues most capable of disrupting B-cell binding can be identified using their SASA, with the assumption that any residue with an exposed surface can be part of a B-cell epitope. This project is unique in relating the existing variant mutations with immune system evasion in different ethnicities predicting how further mutations will affect different ethnicities.

Methods

RBD Structure Prediction

The computational approaches utilized here require atomic level models of the SARS-CoV-2 Spike protein. The structure of the spike protein was obtained from PDB 6XEY [8]. The FASTA sequence was complete and was used for the MHC Class II binding peptide prediction in the next step. In contrast, the protein structure was not complete, with residues 1-26, 177-186, 621-639, 677-689, 829-853, and 1147-1288 missing. The protein structure prediction software i-Tasser [9] was used to complete the structure of the Spike protein. The complete structure of the single strand of the Spike protein was converted to a trimer using UCSF Chimera [10].

MHC Class II Binding Peptide Prediction

Numerous bioinformatics tools are available to generate T-cell peptide binding predictions, including aRB [11], SMM-align [12], ProPred [13], MHCPred [14], SVRMHC [15], and others. A prior study [16] compared the performance of these tools and identified ProPred as the best overall tool. Therefore, it was used in this study. ProPred is a web tool for predicting the promiscuous MHC class II binding regions for several HLA-DR alleles. ProPred is based on TEPITOPE's pocket profile [17] and has one of the broadest coverages with 51 types of alleles. The TEPITOPE matrices are well-reputed for their predictive power . In contrast, ProPred only returns the top 10% of the predicted binders [18].

ProPred uses FASTA sequences as the input, so the mutations were made to the FASTA sequence of the Spike protein. Every residue in the protein was mutated to the 19 other common amino acids and each of the mutated protein sequences was evaluated with ProPred to identify the predicted binding proteins for each allele. The Docker container, GPSRdocker (http://webs.iiitd.edu.in/gpsrdocker/), was used to run the predictions in parallel. The binding prediction calculations generate scores for each nonamer in the input sequence and rank the peptides based on the binding scores. There is a maximum possible binding score for each of the 51 alleles, ranging from 11.6 to 6.

HLA Allele Distribution Analysis

HLA allele frequency was determined using the Allele Frequency Net Database [19]. The database has information, like allele frequency and sample size, for different alleles in different populations around the world. The database is built from literature, proceedings of International Histocompatibility Workshop (IHW), and unpublished data. The data was collected for 50 of the 51 alleles in the ProPred predictions, because there was not enough information regarding the allele DRB5*01:05 in the database for the required calculations. The allele frequency (AF) and

sample size (n) for a population in the database was used to calculate the number of copies of the allele for that data set, shown in Equation 1. When data was not available for allele frequency, Hardy-Weinberg proportions was able to calculate the allele frequency using the phenotype frequency (PF), i.e., percent of individuals that carry the allele, shown in Equation 2. These calculations were provided from the Allele Frequency Net Database.

$$Number\ of\ copies\ of\ alleles = 2 * n * AF \qquad [1]$$

$$AF = 1 - \sqrt{1 - PF} \qquad [2]$$

Ethnicity was determined using population data given from the source, by comparing original data against the population to ethnicity data. The number of copies of the allele was determined for each population data point was determine to distribution of the ethnicities and alleles in different populations. It must be noted that ethnicities and alleles have separate distribution data for the populations. To understand the overall distribution of HLA alleles with respect to the ethnicities, the percentage of an ethnicity in a specific allele and the percentage of an allele in a specific ethnicity were both determined. This section of the project was done in collaboration with a colleague, Mercedes Haley.

Binding Peptide Mutational Analysis

For each of the mutated protein sequences, the predicted binding peptides were compared to that of the wild type Spike protein. When a FASTA sequence is used as an input in ProPred, the prediction calculations are returned as sets of nonamers with binding scores comparative to the highest score for the alleles. These nonamers are referred as the binders or binging peptides, as these nine amino acids are referred as the core of the peptides that fit in the binding pockets. The addition or removal of possible binders and their possible effect is calculated as a percentage

of the highest possible score of any binder to the allele. Using the percentage of the highest possible score for each allele to define the binding capacity of the peptide for that allele, the overall effect of that single mutation on the binding capacity of the MHC Class II proteins from that allele was determined. For example, loss of a binder with a very high prediction score with simultaneous addition of a binder with a very low prediction score would be considered an overall loss in the binding capacity of the MHC Class II proteins from that allele.

Analysis of Mutations in the Spike Protein

Following the identification of the mutations that were predicted to change the binding capacity of the MHC Class II proteins, the Spike mutations that appear in the current variants were identified. The information on the distribution of each variant in the different countries was derived from the COV Lineage website (at https://cov-lineages.org/lineage.html?lineage=B.1.1.7) and their references[14] – [16]. In addition, the distribution of ethnicity for each allele and the distribution of alleles for each ethnicity was assorted. The two sets of data were compared to understand how the current variants may have evaded the immune system in specific ethnicities affecting the rate of infection of the virus. The percentage change in binding may be the summation of the percentage binding score of two or more peptides that were lost due to the mutation. The mutations that were predicted to cause a loss of binding were categorized based on the percentage of the binding score of the lost peptide to the highest possible score for each allele.

Solvent Accessible Surface Area (SASA) Calculation

SASA is a geometric measure to determine the exposure of the residues in a protein structure to the solvent. This measure can be an indicator of the capability of individual residues

in the Spike protein to interact with antibodies. The SASA of the residues of the Spike protein must be determined with the protein trimer, instead of a single chain. The Spike protein is a trimer in the viral structure and the interactions of the residues in each chain affect the exposed surface area for each residue in each chain. The SASA calculation was carried out using the previously developed principles (https://www.ccp4.ac.uk/html/areaimol.html) [23], [24].

Results and Discussions

An important consideration is that while a single mutation may have minimal impact on protein function, accumulated mutations are likely to have significant impact on the function of the protein. Understanding the effects of SARS-CoV-2 mutations on the efficacy of vaccines and immune responses is imperative to the development of therapeutic measures against the fast-evolving virus. The most recent variant of concern (VOC) detected in November 2021, Omicron, has more mutations in the Spike protein of the virus than any of the previous VOCs. Identifying structural and functional impacts of the different mutations in the variants are vital aspects to exploring how these mutations affect the pathogen's ability to evade the immune system, infect host cells, and transfer between cells. Tables 32 to 36 show the results from the different analyses carried out for each VOCs. Each table shows the percentages of the changes in binding predicted for the mutations in each variant for each allele, and the distribution of the alleles in the different ethnicities. Each of these tables also show the domain position of the mutation on the Spike protein.

Starting from early 2021, the Delta variant was able to spread in 98 countries in a few months, with a high rate of infectivity, becoming the dominant variant in 12 countries for the year. The Delta variant has 15 mutations in the Spike protein while the Alpha variant has 13 mutations in the Spike protein. The Delta variant also showed high rates of re-infection in

convalescent patients and infection in vaccinated individuals [25]. Tables 32 and 35 show that the position of the mutations in the Delta variant were more distributed in the NTD than in the RBD as compared to the Alpha variant. The spread of the Delta variant is claimed to be associated with antibody evasion for both non-RBD (e.g., NTD) and RBD epitopes of the Spike protein [26]. The role of the NTD is not completely understood, yet there has been evidence that shows the NTD and RBD have a critical role due to the appearance of neutralizing antibodies binding to both the NTD and RBD [27] – [29], thereby making NTD an important target in therapeutic techniques [30]. The latest VOC of 2021, Omicron, has 30 mutations in the Spike protein and Table 36 shows the mutations distributed in parts of both the S1 and S2 subunits.

The binding of the RBD to the ACE-2 receptor triggers the viral fusion event in the S2 subunit. HR-1 and HR-2 in the S2 subunit interact to create a viral envelope that allows the virus to enter the host cell and the cysteine-rich portion and N-terminal of S2 creates an anchor to the target cell when cell-to-cell infection occurs [6]. There is potential for drug target treatments aimed at HR-1 and HR-2 to prevent viral fusion into the cell [31]. Figure 17 shows the distribution of the mutations in the Spike protein between Alpha, Delta, and Omicron variants. In the figure, a change in distribution of the mutations on the Spike protein is observed from the Alpha, Delta, to the Omicron variant. With change in the distribution of the mutations in the Spike protein between the variants, more mutations are observed in the S2 subunit aiding in viral transmissibility in Omicron than the previous variants. This fact is concerning as Omicron variant may presumably be more transmissible than the previous variants.

Table 29 shows the distribution of alleles in each ethnicity, while Table 30 shows the distribution of the ethnicities for each allele. The most prevalent ethnicities throughout all the alleles are Asian, Black, Caucasoid, Hispanic, and Oriental. The Arab ethnicity is most prevalent

in alleles DRB1*03:01 and DRB1*13:21, while the Amerindian group is most prevalent in the DRB1*04:21 allele. The most prominent (or top) alleles, defined by their prevalence in the top 50% of each ethnicity, as shown in Table 3, are DRB1*07:01 and DRB1*15:01. Some alleles appear in only one ethnicity, for example, DRB1*01:02 in Berber and DRB1*04:02 in Jew.

The Alpha variant (B.1.1.7) first appeared in the United Kingdom, where the more widely found ethnicities with DRB alleles are Caucasoid and Mixed [19]. The Caucasoid and Mixed groups have the top alleles of DRB1*03:01, DRB1*07:01, and DRB1*15:01. Jews are also a commonly found ethnicity in United Kingdom and their top alleles are DRB1*04:02, DRB1*07:01, and DRB1*11:04. The mutation D1118H show detrimental changes in the binding capacity against the DRB1*03:01 allele. However, the mutation, D1118H, shows favorable changes towards the DRB1*04:02, DRB1*04:05, DRB1*13:01, and DRB1*13:02 alleles. The mutation S494P shows detrimental changes against the DRB1*07:01 and DRB1*15:01 alleles and the mutation A570D for the DRB1*15:02 allele. The countries with the highest number of cases of the Alpha variant besides the United Kingdom were United States, Germany, Sweden, and Denmark. These countries are predominantly Caucasoids and their alleles show detrimental changes in binding against the mutations in the Alpha variant. The DRB5*01:01 allele was the only allele in the Caucasoid group that had no detrimental effects in binding. The other alleles did have beneficial changes on some mutation points.

The Beta variant originated in South Africa, where the commonly found ethnicities with DRB alleles are Black and Caucasoid. None of these ethnicities showed adverse changes in binding for the top alleles. The countries that were most affected besides South Africa were Philippines, United States, Sweden, and Germany. The ethnicities commonly found in these countries are Austronesian, Caucasoid, and Oriental ethnicities. Only Austronesian group could

have suffered from an adverse change from the mutation A570D against the DRB1*15:02 allele. The Beta variant was less severe than other variants, likely due to the lack of immune-evading mutations.

The Gamma variant first appeared in Japan and Brazil. The most common ethnic groups in Japan are Oriental and Caucasoid, while in Brazil, the common groups are Amerindian, Caucasoid, Mixed and Mestizo. The most adverse changes possible from the mutation R190S against the DRB1*15:01 allele in the ethnic groups from the two locations. The Mestizo ethnic group may have possibly been affected by the mutation N501Y against the DRB1*04:04 allele. The other countries affected by this variant had the same ethnicities as the countries of origin.

The Delta variant showed the most adverse changes among the ethnicity groups in the region of origin, India, where Asian and Caucasoid ethnicities are more commonly found with DRB alleles. The top alleles for these ethnic groups are DRB1*15:01, DRB1*07:01, DRB1*15:02, DRB1*3:01, and DRB1*01:01. The top alleles for each of these ethnicities had the most adverse changes from the mutations. Significant detrimental changes were observed by several mutations in this variant, the mutation L452R against DRB1*11:04, the mutation R158G against both the DRB1*15:01 and the DRB1*15:02 alleles, the mutations T19R and T95I against the DRB1*07:01 allele, and the mutation T95I against the DRB1*04:04 allele. The countries most effected besides the India were United States, United Kingdom, Denmark, and Germany. The most common ethnicities in all the countries are Caucasoid, Oriental, Asian, and Mixed listed, all of whom are affected by these mutations to some scale. Apart from these countries, the Delta variant affected Middle Eastern countries and the Southeast Asian countries. Arab, Kurd, and Oriental ethnic groups were all affected by the mutations in the Delta variant.

Omicron (B.1.1.529) is the most recent VOC that originated in South Africa. The Black and Caucasoid ethnic groups appear in South Africa, with the Black ethnic group appearing more frequently. The top alleles for this ethnic group are DRB5*01:01, DRB1*07:01, DRB1*11:01, and DRB1*13:02. The Black ethnic group is the only group where the DRB1*13:02 appears as a prominent allele, for which the mutations on the RBD show adverse changes while the mutations on the S2 subunit show favorable changes. The DRB1*07:01 and DRB5*01:01 alleles appear in both the Black and Caucasoid groups. The mutation Y505H shows extreme detrimental effects against the DRB1*07:01 allele, with the loss of two prominent binders; such extreme effects have not been seen in any of the previous variants. The countries most affected by Omicron, besides South Africa, were Botswana, United States, Nepal, and Mexico. These countries predominantly have Black, Caucasoid, and Amerindian ethnicities, with the DRB1*07:01 and DRB5*01:05 alleles commonly present in these ethnicities. The mutations in the Omicron variant show adverse changes against the DRB1*03:01, DRB1*04:01, DRB1*04:04, DRB1*04:05, DRB1*13:02, DRB1*07:01, DRB1*15:01, DRB1*15:02, and DRB5*01:01 alleles. This could mean that there are more vulnerable ethnicities against this new variant, specifically groups that do not share a diverse range of alleles, like the Mestizo and Oriental ethnicities.

Apart from looking at the previous and current VOCs, mutations that are predicted to cause adverse interactions in the binding pockets were identified. Table 37 shows the mutations causing extreme changes in binding for almost all the alleles. For all the positions shown in the table, the mutation of either an aliphatic or a basic residue to aspartic acid are observed. Introducing a large negative charge in the place of a relatively uncharged or positively charged sidechain would be disruptive to any protein binding, and these mutations could be disruptive for

the activities of the Spike protein as well. One of the important Spike protein mutations in the Omicron variant is Y505H, one of the 12 mutations in the 505 position that were predicted to cause adverse effects on MHC Class II protein binding. The mutation Y505H shows extreme adverse changes in binding for some alleles with the loss of two binding peptides. This position was known to be one of the hotspot residues for the spike protein [32], [33], and not expected to mutate with regards to protein stability and binding to ACE-2. The mutation Y505H was previously proven to increase stability but decrease binding capacity to ACE-2 [34], [35]. Thus, the predicted changes of mutation Y505H and the previous studies are evidence that the viral evolution may be directed by the virus's ability to evade the immune system. Beyond the benefits of immune evasion, there were mutations observed that proved other evolutionary benefits such as the mutation D614G, which was observed in all the previous and current VOCs, is known to enhance viral replication in airway tissue [36].

The predictions emphasize the importance of understanding the impact of future mutations on the human immune. The mutations that are predicted to have extreme adverse effects on binding in the same positions may need to be scrutinized further to assume consequences of their appearance in future mutants. Table 37 shows the mutations causing extreme changes in binding for almost all the alleles. The mutations showing such extreme changes are all in the S1 subunit, with only two of them in the RBD. Table 38 shows the predicted changes from the mutations in the position Y505 compared to the predicted changes from the mutations in the position L455.

Figure 18 shows the surface exposed residues based on their SASA. The hypothesis is that any residue that is exposed to the surface, however small that exposure may be, can make an interaction with other proteins, thus, can contribute to binding to antibodies. This hypothesis was

made from examples of antibody-antigen structures found in nature and is a conservative assumption to err on the side of identifying escape mutations. Figure 19 shows an example of PDB 1BZQ, where a Tyrosine in the antigen makes a hydrogen bond with another Tyrosine in the antibody despite the low exposed surface area of the Tyrosine in the antigen. Mutation of any of the exposed residues can influence B-cell epitopes, thus making them more capable of evading antibodies. Furthermore, upon observation of Figure 3, more intermediately exposed residues can be identified, which can mean that they have a higher probability of being part of the B-cell epitopes in the Spike protein. Comparing Figure 17 and Figure 18, the distribution of the mutations in the variants is evidence that natural selection is directed towards changing B-cell and T-cell epitopes, and thus evading the immune system. The mutations in the position 454 and 455 show adverse changes in binding peptides for the MHC Class II peptides as mentioned in the results, while they are also included in the larger set of residues that have a solvent exposure below 50 squared Å. This information is enough to require the monitoring of any mutations observed in these two positions in future variants.

To validate the ProPred predictions, results from other prediction tools that have worked on the mutations observed in the variants. T-CoV is a comprehensive tool to predict binding for different epitopes to the mutations of SARS-CoV-2[37]. This tool uses NetMHCpan-4.1 (available at http://www.cbs.dtu.dk/services/NetMHCpan-4.1/) and NetMHCIIpan-4.0 (available at http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/) to predict binding affinity. T-CoV also found that the mutation D1118H in the Alpha variant reduced the T cell response for HLA-DRB1*03:01 allele. One of the most severe mutations for the Omicron variant is Y505H. Our results showed that there was loss of binding peptides at positions 496 and 504. The binder sequence that includes both positions is SFRPTYGVGHQPYRVVV, the same binder for which

T-CoV showed weak binding to epitopes in the Omicron variant. T-CoV only identifies the binders using the prediction method, it does not work on relating the binding changes to the ethnic distribution in regions.

Analysis of the T cell epitopes from convalescent Covid-19 patients have shown that around 30% of the epitopes target the Spike protein, the rest of them target different regions of the virus [38], making Spike protein the main target of neutralizing antibodies. Immune evasion is a major factor for viral mutations. Attempting to evade the immune system, the virus will mutate the outer-most region, the Spike protein, more than the other proteins in the structure. Therefore, the Spike protein mutations were the focus of this study. The Omicron variant shows multiple mutations in the other proteins of the virus and one further path for this project may be to study the mutations on the other outer proteins of the virus.

The SARS-CoV-2 virus has caused devastating effects to millions of lives in a span of a few years. Since early 2020, countries with less developed healthcare sectors like India, Brazil, Peru, and Mexico have seen thousands of deaths each day. Countries with more developed healthcare sectors, like the USA, the UK, Italy, did not suffer any less. Even with a worldwide vaccination regime, the virus just mutated to better evade the immune system. A question to be asked is if the vaccines can generate the same level of protection for everyone around the world. It is understood that the HLA allele distribution is not the same for people of different ethnic backgrounds. One goal of this project was to bring this information forward, that different mutations in the Spike protein do not affect different people in the same way. There are more vulnerable ethnicities that may require targeted or specified vaccines than others.

There is a change in the distribution of the mutations in the Spike protein between the variants from Alpha to Delta to Omicron, there are more mutations in the S2 subunit in Omicron

than before. The S2 subunit works with host cell transmissions, leading to assumption that the Omicron variant will be more transmissible than the previous variants, showing higher rates of infectivity. The mutations also indicate less severity of infection in the Omicron variant than Delta. After the severity of the Delta variant, the Omicron variant is under closer observations from scientists. It is true that not enough is known about the effects of the Omicron variant to make proper conclusions.

When Omicron variant was first observed, 30 mutations in the Spike protein was not completely surprising. Viral evolution is driven by its ability to survive and spread through the host cells. The mutations were more clustered in the RBD, which is the part of the structure that comes in first contact with the antibodies. The virus will evolve towards mutating the B-cell epitopes to disrupt binding to the B-cell antibodies. The other goal of this project was to be prepared for further mutations in the virus. More than two mutation sites were identified as severe consequence mutation site in this project. It may be possible that these specific sites do not mutate anytime soon, but this project has more future paths in predicting the effects of mutations. One such path may be to conduct a similar study for MHC Class I protein binding predictions. If the mutations predicted for both MHC proteins are the same, the mutations can be examined further experimentally.

This project can be built further beyond just SARS-CoV-2, working on other viruses. There are other fast-mutating viruses that need to be observed and their mutations require studying. This workflow can be used for building more targeted therapeutic treatments against COVID-19 and other viral pathogens.

Figure 17: Structure details of the SARS-CoV-2 Spike protein.

A. Sequence of primary structure of SARS-CoV-2 showing the S1 subunit, S2 subunit, and cleavage points.

B. The different colors (blue, red, and green) represent each monomer that creates the trimer structure of SARS-CoV-2 Spike protein.

C. One monomer that is part of the Spike protein vital domains on the Spike protein are marked by colors. The RBD and NTD are on the outside of the protein to allow binding to host while the S2 subunit is located closer to the SARS-CoV-2 membrane.

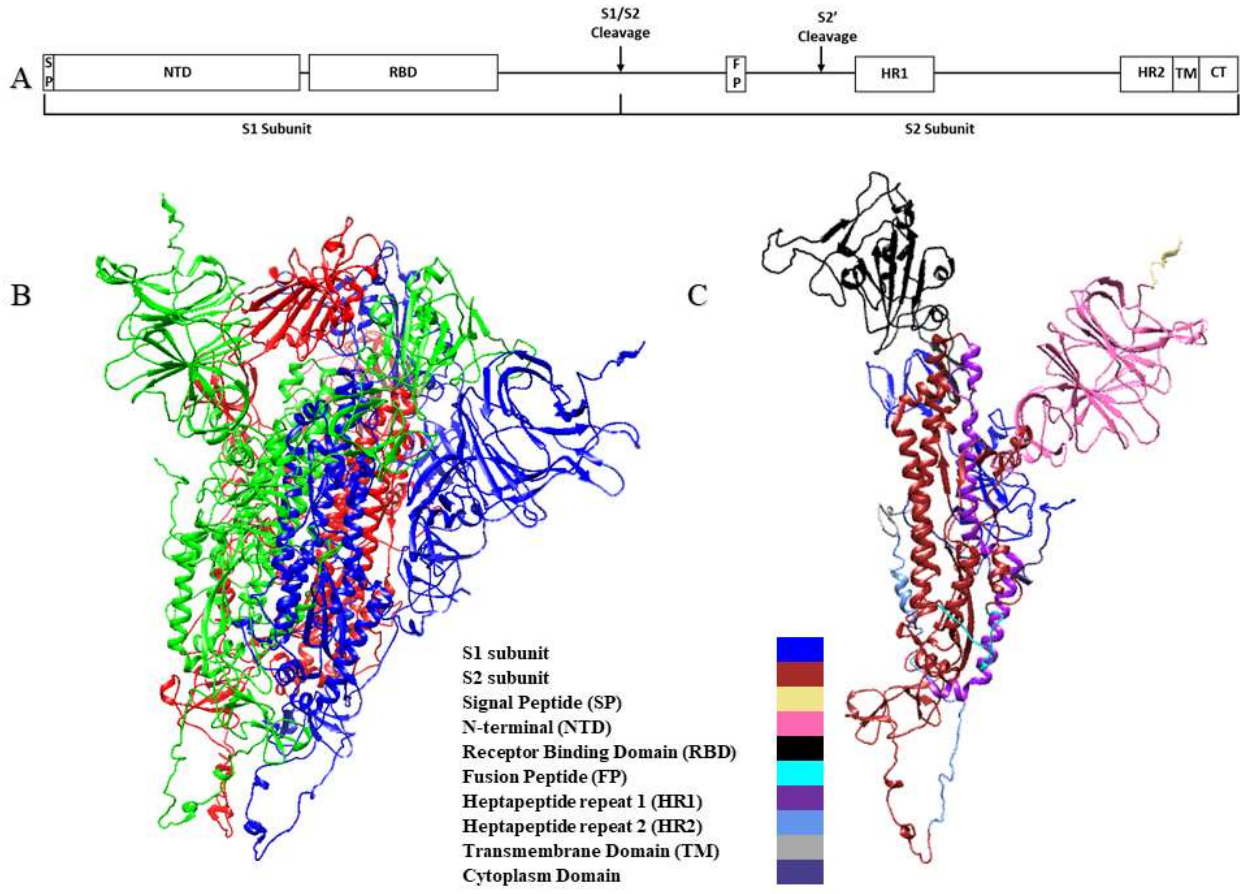


Table 29: The distribution of the alleles for the ethnicities. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. DRB1*03:01, DRB1*07:01, and DRB1*15:01 appears in most ethnicities. Using the allele information for each ethnicity, specific alleles that can greatly affect each ethnicity can be determined.

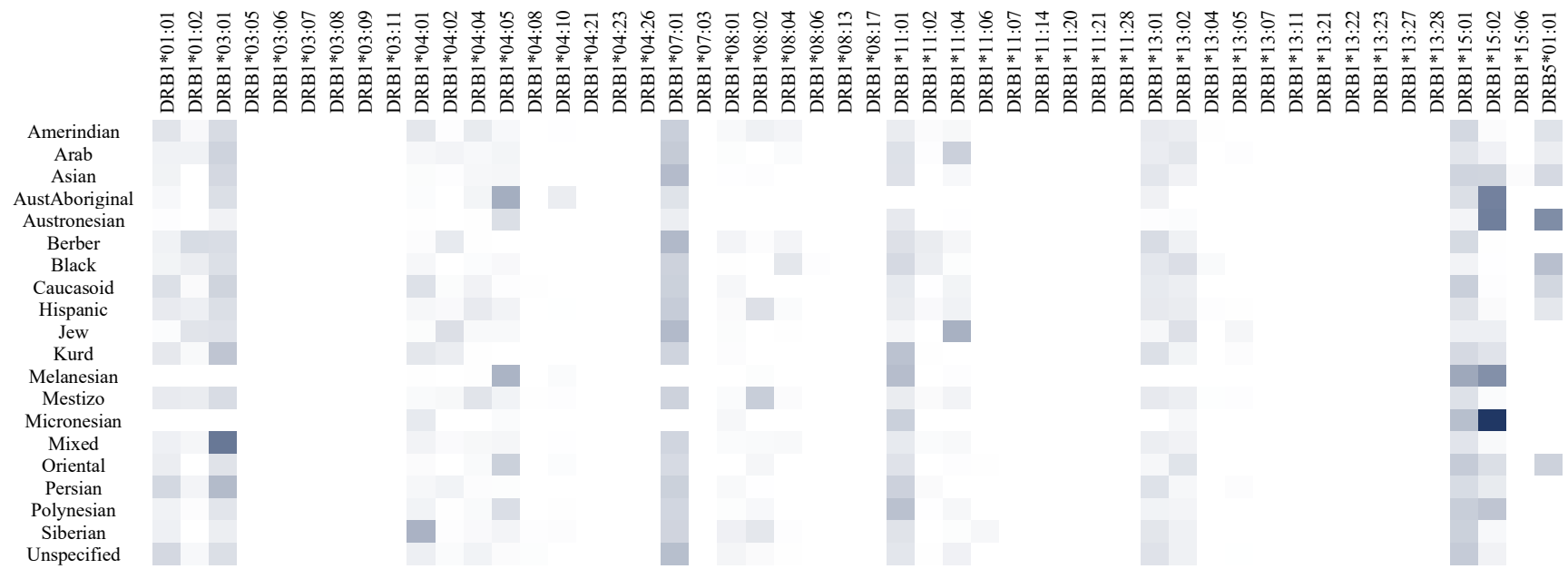
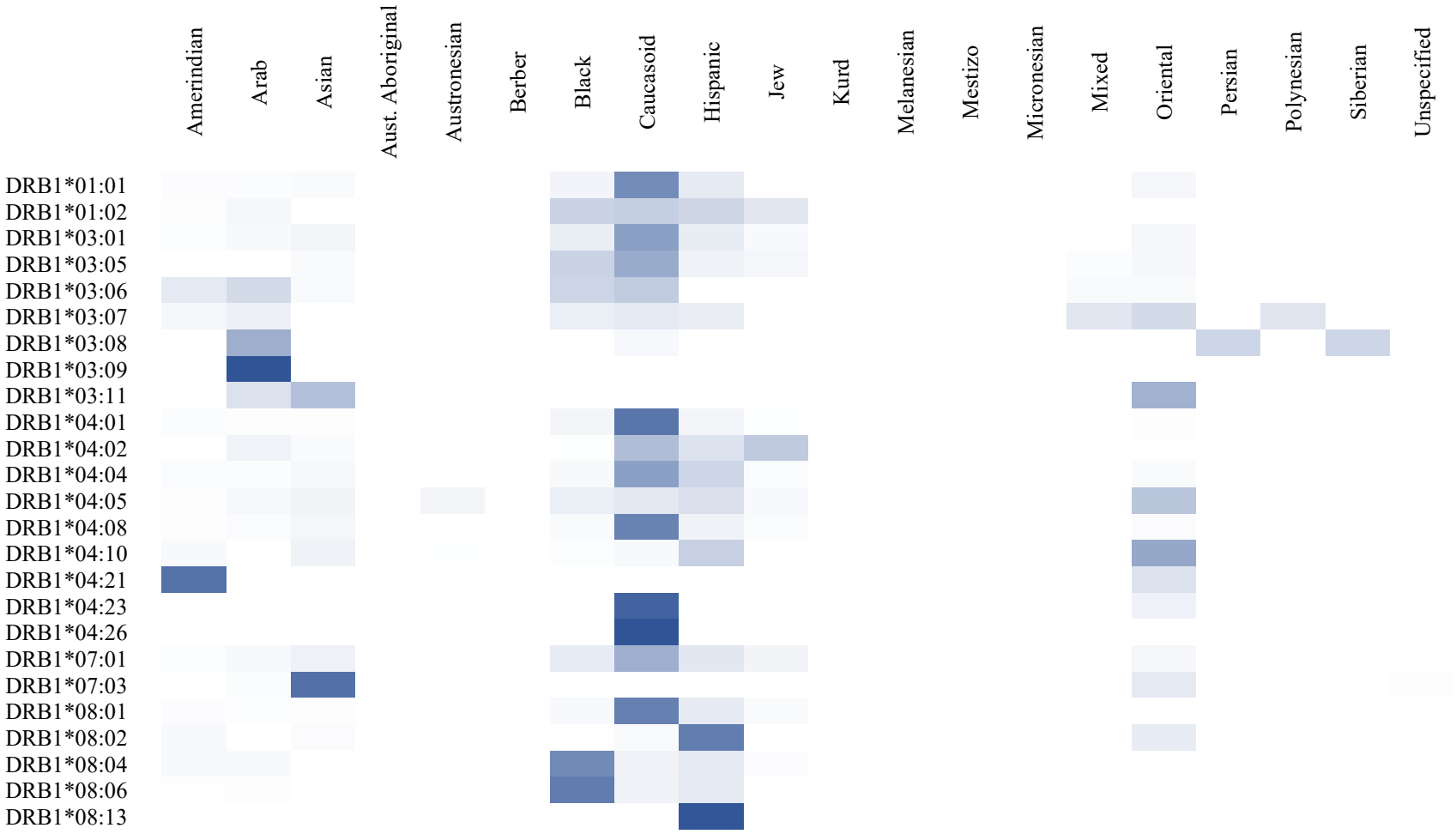


Table 30: The distribution of the ethnicities for the alleles. The pigmentation of the cell indicates no interactions to a high number of interactions from white to navy blue. Most of the alleles are observed in the Caucasoid group. A large set of the alleles are observed in the Asian, Black, Caucasoid, Hispanic, and Oriental groups.



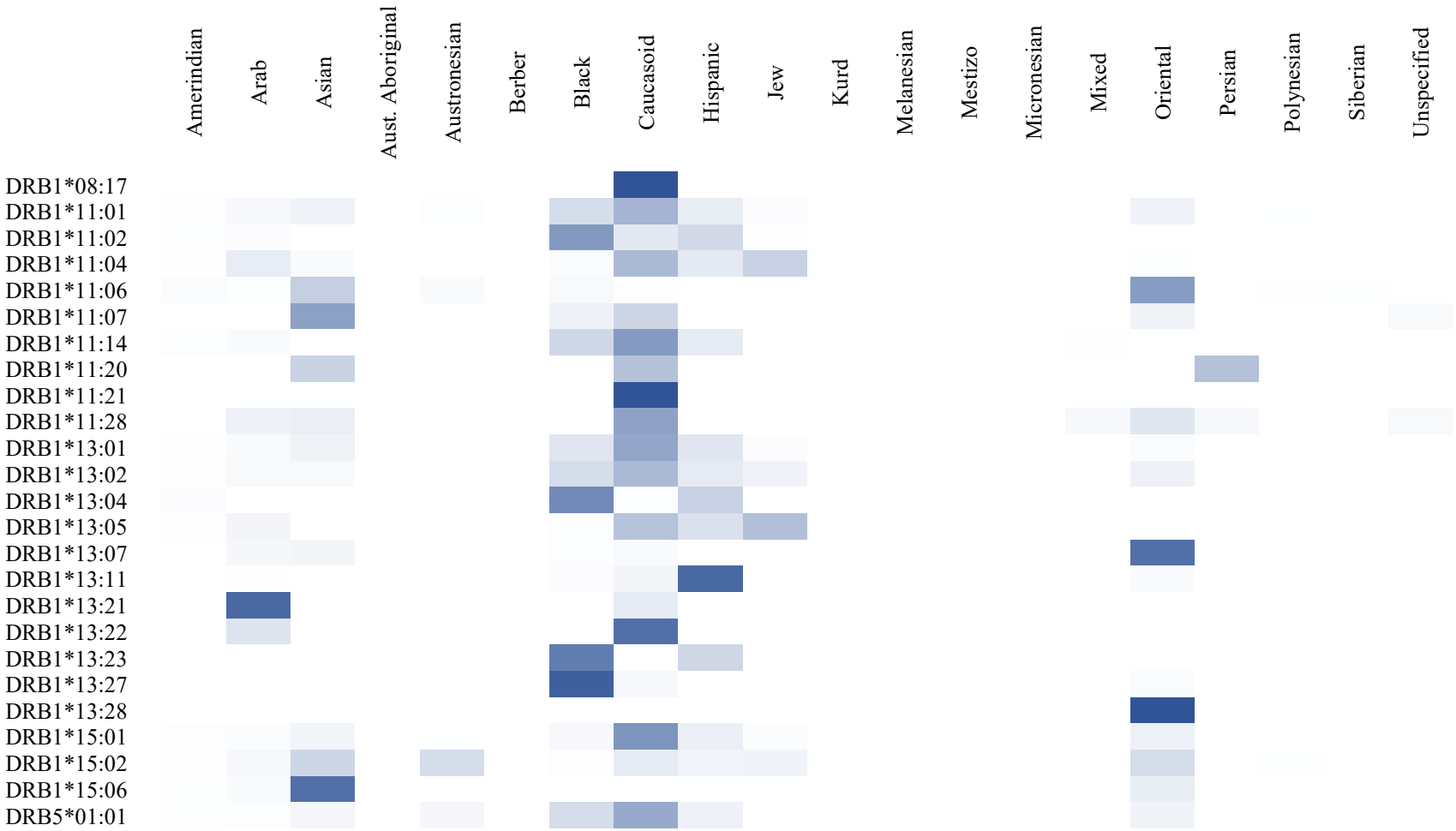

Amerindian
Arab
Asian
Aust. Aboriginal
Austronesian
Berber
Black
Caucasoid
Hispanic
Jew
Kurd
Melanesian
Mestizo
Micronesian
Mixed
Oriental
Persian
Polynesian
Siberian
Unspecified
DRB1*08:17
DRB1*11:01
DRB1*11:02
DRB1*11:04
DRB1*11:06
DRB1*11:07
DRB1*11:14
DRB1*11:20
DRB1*11:21
DRB1*11:28
DRB1*13:01
DRB1*13:02
DRB1*13:04
DRB1*13:05
DRB1*13:07
DRB1*13:11
DRB1*13:21
DRB1*13:22
DRB1*13:23
DRB1*13:27
DRB1*13:28
DRB1*15:01
DRB1*15:02
DRB1*15:06
DRB5*01:01

Table 31: Top (50%) alleles for ethnicities. DRB1*07:01 and DRB1*15:01 appear in the top percentage of alleles for most ethnicity. There are a couple of alleles that are only in one ethnicity including DRB1*01:02, DRB1*04:02, and DRB1*13:02. These are just a few examples of this occurring.

Ethnicity	1st allele	2nd allele	3rd allele	4th allele	5th allele	6th allele
Amerindian	DRB1*07:01	DRB1*15:01	DRB1*03:01	DRB5*01:01	DRB1*01:01	
Arab	DRB1*07:01	DRB1*11:04	DRB1*03:01	DRB1*11:01	DRB1*15:01	
Asian	DRB1*07:01	DRB1*15:01	DRB1*15:02	DRB1*03:01		
Aust. Aboriginal	DRB1*15:02	DRB5*01:01				
Austronesian	DRB1*15:02	DRB5*01:01				
Berber	DRB1*07:01	DRB1*15:01	DRB1*01:02	DRB1*13:01		
Black	DRB5*01:01	DRB1*07:01	DRB1*11:01	DRB1*13:02		
Caucasoid	DRB1*15:01	DRB1*07:01	DRB1*03:01	DRB5*01:01		
Hispanic	DRB1*07:01	DRB1*03:01	DRB1*08:02	DRB1*15:01	DRB5*01:01	DRB1*13:01
Jew	DRB1*11:04	DRB1*07:01	DRB1*04:02			
Kurd	DRB1*11:01	DRB1*03:01	DRB1*07:01	DRB1*15:01		
Melanesian	DRB1*15:02	DRB1*15:01				
Mestizo	DRB1*08:02	DRB1*07:01	DRB1*03:01	DRB1*15:01	DRB1*04:04	
Micronesian	DRB1*15:02					
Mixed	DRB1*03:01	DRB1*07:01	DRB1*15:01			
Oriental	DRB1*15:01	DRB1*04:05	DRB5*01:01	DRB1*07:01		
Persian	DRB1*03:01	DRB1*07:01	DRB1*11:01	DRB1*01:01		
Polynesian	DRB1*11:01	DRB1*15:02	DRB1*15:01	DRB1*07:01		
Siberian	DRB1*04:01	DRB1*15:01	DRB1*07:01	DRB1*11:01		
Unspecified	DRB1*07:01	DRB1*15:01	DRB1*01:01	DRB1*03:01		

Table 32: Effects of the Alpha Variant Mutations on Various Ethnic Groups. The mutations S494P and D1118H have detrimental effects on alleles that are more abundant in a large set of ethnic groups. Many of the mutations are located at the RBD and S2 subunits.

Allele	Mutation	Location	Change in Binding	Amerindian	Arab	Asian	Aust. Aboriginal	Austronesian	Berber	Black	Caucasoid	Hispanic	Jew	Kurd	Melanesian	Mestizo	Micronesian	Mixed	Oriental	Persian	Polynesian	Siberian	Unspecified
DRB1*01:01	S494P	RBD	6.67	■																■			■
	N501Y	RBD	6.67	■																■			■
	D614G	S1	25.00	■																■			■
	S982A	HR1	5.00	■																■			■
DRB1*01:02	E484K	RBD	13.00						■														
	S494P	RBD	15.00						■														
	D614G	S1	12.33						■														
	T716I	S2	12.33						■														
	S982A	HR1	5.00						■														
DRB1*03:01	S982A	HR1	-0.74	■	■	■					■	■		■		■		■		■			■
	D1118H	S2	-49.47	■	■	■					■	■		■		■		■		■			■
DRB1*04:01	N501Y	RBD	-20.93																			■	
	T716I	S2	25.35																			■	
	D1118H	S2	-3.49																			■	
DRB1*04:02	N501Y	RBD	20.83										■										
	T716I	S2	23.96										■										
	D1118H	S2	44.79										■										
DRB1*04:04	N501Y	RBD	-25.00													■							
DRB1*04:05	S494P	RBD	-29.79																■				
	N501Y	RBD	-9.57																■				
	D1118H	S2	21.28																■				
DRB1*07:01	S494P	RBD	-40.52	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	D614G	S1	44.83	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	T716I	S2	39.66	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
DRB1*08:02	S494P	RBD	-25.00									■				■							
DRB1*11:01	S494P	RBD	-18.08		■					■				■						■	■	■	
	S982A	HR1	15.06		■					■				■						■	■	■	
DRB1*11:04	S982A	HR1	27.11		■								■										
DRB1*13:01	D1118H	S2	45.46						■			■											■
DRB1*13:02	S494P	RBD	-29.55							■													
	N501Y	RBD	-3.41							■													
	D1118H	S2	34.09							■													
DRB1*15:01	E484K	RBD	33.67	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	S494P	RBD	-37.76	■	■	■			■		■	■		■	■	■		■	■		■	■	■

DRB1*15:02	S494P	RBD	-8.16																					
	A570D	S1	-39.80																					
DRB5*01:01	S494P	RBD	7.14																					
	D614G	S1	26.53																					

Table 33: Effects of the Beta Variant Mutations on Various Ethnic Groups. The mutations D215G, N501Y, A701V, and A570D are predicted to have detrimental changes. Most of the mutations are found on the S1 subunit. One detrimental change on S2 occurred for the allele DRB1*04:02, most commonly in the Jewish ethnicity.

Allele	Mutation	Location	Change in Binding	Amerindian	Arab	Asian	Aust. Aboriginal	Austronesian	Berber	Black	Caucasoid	Hispanic	Jew	Kurd	Melanesian	Mestizo	Micronesian	Mixed	Oriental	Persian	Polynesian	Siberian	Unspecified
DRB1*01:01	D215G	NTD	11.67	■																■			■
	N501Y	RBD	6.67	■																■			■
	D614G	S1	25.00	■																■			■
DRB1*01:02	D215G	NTD	28.33						■														
	E484K	RBD	13.00						■														
	D614G	S1	12.33						■														
DRB1*03:01	D215G	NTD	15.79	■	■	■					■	■		■		■		■		■			■
DRB1*04:01	D215G	NTD	-25.58																			■	
	N501Y	RBD	-20.93																			■	
	A701V	S2	10.23																			■	
DRB1*04:02	D215G	NTD	22.92										■										
	N501Y	RBD	20.83										■										
	A701V	S2	-23.96										■										
DRB1*04:04	D215G	NTD	-22.73													■							
	N501Y	RBD	-25.00													■							
DRB1*04:05	N501Y	RBD	-9.57																■				
DRB1*07:01	D614G	S1	44.83	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	A701V	S2	51.72	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
DRB1*08:02	D215G	NTD	25.00									■				■							
DRB1*11:01	D215G	NTD	60.24		■					■				■						■	■	■	
DRB1*11:04	D215G	NTD	60.24		■								■										
DRB1*13:01																							
DRB1*13:02	N501Y	RBD	-3.41							■													
DRB1*15:01	E484K	RBD	33.67	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	A701V	S2	33.67	■	■	■			■		■	■		■	■	■		■	■		■	■	■
DRB1*15:02	A570D	S1	-39.80			■	■	■							■		■				■		
DRB5*01:01	D215G	NTD	53.06	■			■	■		■	■	■							■				
	D614G	S1	26.53	■			■	■		■	■	■							■				

Table 34: Effects of the Gamma Variant Mutations on Various Ethnic Groups. The affected alleles are DRB1*01:01, DRB1*04:01, DRB1*04:04, and DRB1*15:01, more particularly affecting the Mestizo and Siberian ethnicities. Two alleles not affected by any mutations are DRB1*11:04 and DRB1*13:01.

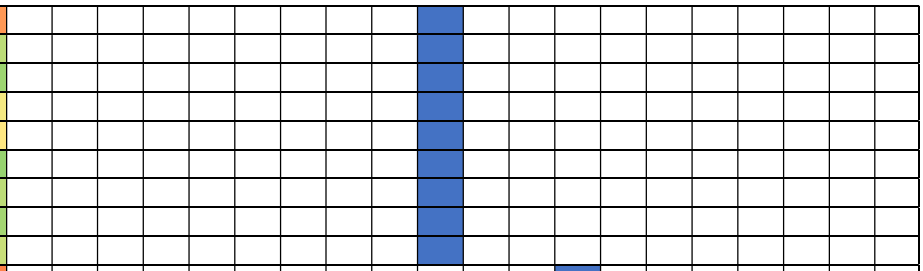
Allele	Mutation	Location	Change in Binding	Amerindian	Arab	Asian	Aust. Aboriginal	Austronesian	Berber	Black	Caucasoid	Hispanic	Jew	Kurd	Melanesian	Mestizo	Micronesian	Mixed	Oriental	Persian	Polynesian	Siberian	Unspecified
DRB1*01:01	L18F	NTD	-15.00	■																■			■
	D138Y	NTD	9.50	■																■			■
	N501Y	RBD	6.67	■																■			■
	D614G	S1	25.00	■																■			■
	T1027I	S2	11.67	■																■			■
DRB1*01:02	L18F	NTD	9.90						■														
	E484K	RBD	13.00						■														
	D614G	S1	12.33						■														
	T1027I	S2	28.33						■														
DRB1*03:01	L18F	NTD	9.90	■	■	■					■	■		■		■		■		■			■
	T1027I	S2	40.00	■	■	■					■	■		■		■		■		■			■
DRB1*04:01	N501Y	RBD	-20.93																			■	
DRB1*04:02	N501Y	RBD	20.83										■										
DRB1*04:04	N501Y	RBD	-25.00													■							
DRB1*04:05	D138Y	NTD	35.64																■				
	N501Y	RBD	-9.57																■				
DRB1*07:01	L18F	NTD	-12.93	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	D138Y	NTD	36.21	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	D614G	S1	44.83	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
DRB1*08:02	P26S	NTD	17.50									■				■							
	D138Y	NTD	14.75									■				■							
DRB1*11:01	D138Y	NTD	27.47		■					■				■						■	■	■	
DRB1*11:04																							
DRB1*13:01																							
DRB1*13:02	N501Y	RBD	-3.41							■													
DRB1*15:01	D138Y	NTD	39.80	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	R190S	NTD	-35.71	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	E484K	RBD	33.67	■	■	■			■		■	■		■	■	■		■	■		■	■	■
DRB1*15:02	D138Y	NTD	50.00			■	■	■							■		■				■		
DRB5*01:01	D614G	S1	26.53	■			■	■		■	■	■							■				

Table 35: Effects of the Delta Variant Mutations on Various Ethnic Groups. Mutations on the Spike protein for Delta variant are located mostly on the NTD of the S1 subunit. The mutations show a decrease in immunogenicity for this allele DRB1*07:01, affecting almost all ethnicities. The Jewish and Arab groups are affected by the detrimental changes in allele DRB1*11:04.

Allele	Mutation	Location	Change in Binding	Amerindian	Arab	Asian	Aust. Aboriginal	Austronesian	Berber	Black	Caucasoid	Hispanic	Jew	Kurd	Melanesian	Mestizo	Micronesian	Mixed	Oriental	Persian	Polynesian	Siberian	Unspecified
DRB1*01:01	D614G	S1	25.00	■																■			■
DRB1*01:02	V70F	NTD	-16.33						■														
	D614G	S1	12.33						■														
	D950N	HR1	12.50						■														
DRB1*03:01	T95I	NTD	-1.05	■	■	■					■	■		■		■		■		■			■
DRB1*04:01	V70F	NTD	3.49																			■	
	T95I	NTD	-23.26																			■	
DRB1*04:02	V70F	NTD	19.79										■										
	T95I	NTD	-32.29										■										
	L452R	RBD	12.50										■										
	D950N	HR1	35.42										■										
	T478K	RBD	-1.04										■										
DRB1*04:04	V70F	NTD	3.41													■							
	T95I	NTD	-40.91													■							
	T478K	RBD	-26.14													■							
	D950N	HR1	37.50													■							
DRB1*04:05	V70F	NTD	-8.51																■				
	T95I	NTD	-27.66																■				
	R158G	NTD	-18.09																■				
	A222V	NTD	5.32																■				
	D950N	HR1	48.94																■				
DRB1*07:01	T19R	NTD	-46.55	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	T95I	NTD	-37.07	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
	D614G	S1	44.83	■	■	■			■	■	■	■	■	■		■		■	■	■	■	■	■
DRB1*08:02	T19R	NTD	20.00									■				■							
	R158G	NTD	-21.25									■				■							
	L452R	RBD	35.00									■				■							
DRB1*11:01	A222V	NTD	6.02		■					■				■						■	■	■	
	L452R	RBD	-20.48		■					■				■						■	■	■	
DRB1*11:04	L452R	RBD	-63.86		■								■										
DRB1*13:01	T95I	NTD	-29.55						■			■											
	L452R	RBD	32.95						■			■											
DRB1*13:02	R158G	NTD	-27.27							■													
	L452R	RBD	11.36							■													
DRB1*15:01	T19R	NTD	37.76	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	V70F	NTD	-7.14	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	R158G	NTD	-36.74	■	■	■			■		■	■		■	■	■		■	■		■	■	■
DRB1*15:02	R158G	NTD	-46.94			■	■	■							■		■				■		
	L452R	RBD	-3.06			■	■	■							■		■				■		
DRB5*01:01	V70F	NTD	27.55	■			■	■		■	■	■							■				
	L452R	RBD	-40.82	■			■	■		■	■	■							■				
	D614G	S1	26.53	■			■	■		■	■	■							■				

Table 36: Effects of the Omicron Variant Mutations on Various Ethnic Groups. The mutations are distributed across both the S1 and S2 subunits. The Black ethnicity is highly affected by the changes in alleles DRB1*07:01, DRB1*11:01, DRB1*13:02, and DRB5*01:01 alleles. The mutations Y505H, located on the RBD, has severe negative effects in most of the mentioned alleles.

Allele	Mutation	Location	Change in Binding	Amerindian	Arab	Asian	Aust. Aboriginal	Austronesian	Berber	Black	Caucasoid	Hispanic	Jew	Kurd	Melanesian	Mestizo	Micronesian	Mixed	Oriental	Persian	Polynesian	Siberian	Unspecified
DRB1*01:01	A67V	NTD	5.00	■																■			■
	S371L	RBD	13.33	■																■			■
	S373P	RBD	4.50	■																■			■
	S375F	RBD	16.66	■																■			■
	N501Y	RBD	6.67	■																■			■
	Y505H	RBD	-8.33	■																■			■
	D614G	S1	25.00	■																■			■
	N856K	S2	5.00	■																■			■
	Q954H	HR1	-13.33	■																■			■
	L981F	HR1	16.67	■																■			■
DRB1*01:02	A67V	NTD	16.33						■														
	S371L	RBD	30.00						■														
	S373P	RBD	4.50						■														
	S375F	RBD	26.33						■														
	G496S	RBD	-10.00						■														
	Q498R	RBD	6.67						■														
	D614G	S1	12.33						■														
	Q954H	HR1	-13.33						■														
	N969K	HR1	-11.67						■														
DRB1*03:01	T95I	NTD	-1.05	■	■	■					■	■		■		■		■		■			■
	S373P	RBD	6.63	■	■	■					■	■		■		■		■		■			■
	S375F	RBD	-2.11	■	■	■					■	■		■		■		■		■			■
	Q498R	RBD	16.67	■	■	■					■	■		■		■		■		■			■
	L981F	HR1	-42.42	■	■	■					■	■		■		■		■		■			■
DRB1*04:01	A67V	NTD	18.61																			■	
	T95I	NTD	-23.26																			■	
	S371L	RBD	10.47																			■	
	S373P	RBD	-38.37																			■	
	S375F	RBD	-25.58																			■	
	G496S	RBD	-9.30																			■	
	Q498R	RBD	-45.35																			■	
	N501Y	RBD	-20.93																			■	
	N679K	S1	-24.42																			■	
	N856K	S2	3.49																			■	
	Q954H	HR1	19.77																			■	
	N969K	HR1	-17.44																			■	

DRB1*04:02	T95I	NTD	-32.29	
	L212I	NTD	20.83	
	S371L	RBD	30.21	
	S477N	RBD	3.13	
	T478K	RBD	-1.04	
	Q498R	RBD	32.29	
	N501Y	RBD	20.83	
	N764K	S2	28.13	
	Q954H	HR1	16.67	
DRB1*04:04	T95I	NTD	-40.91	
	L212I	NTD	5.68	
	S371L	RBD	56.82	
	S373P	RBD	-56.82	
	S375F	RBD	-26.14	
	S477N	RBD	3.41	
	T478K	RBD	-26.14	
	G496S	RBD	-25.00	
	Q498R	RBD	-25.00	
	N501Y	RBD	-25.00	
	N764K	S2	27.27	
	N969K	HR1	-31.82	
DRB1*04:05	A67V	NTD	5.32	
	T95I	NTD	-27.66	
	L212I	NTD	24.47	
	S371L	RBD	41.49	
	S373P	RBD	-17.02	
	G496S	RBD	-8.51	
	Q498R	RBD	-37.23	
	N501Y	RBD	-9.57	
	N679K	S1	-35.11	
	N764K	S2	-17.02	
	N856K	S2	3.19	
	N969K	HR1	-32.98	
DRB1*07:01	A67V	NTD	37.07	
	T95I	NTD	-37.07	
	S371L	RBD	-43.10	
	S373P	RBD	-43.10	
	S375F	RBD	78.45	
	Q498R	RBD	8.62	
	Y505H	RBD	-92.24	
	D614G	S1	44.83	
	Q954H	HR1	13.79	
DRB1*08:02	S375F	RBD	23.75	
	Q493K	RBD	17.50	
	Q498R	RBD	30.00	
	Y505H	RBD	-35.00	
	N764K	S2	13.75	
	N856K	S2	3.75	
	N969K	HR1	15.00	
DRB1*11:01	L212I	NTD	15.66	
	S371L	RBD	34.94	
	S373P	RBD	17.47	
	N764K	S2	14.46	
	N856K	S2	3.61	
	L981F	HR1	23.49	
DRB1*11:04	L212I	NTD	27.71	
	S373P	RBD	29.52	
	N764K	S2	26.51	
	N856K	S2	3.62	

DRB1*13:01	T95I	NTD	-29.55						■			■													
	Q954H	HR1	10.23						■			■													
DRB1*13:02	G496S	RBD	-9.09							■															
	Q498R	RBD	-4.55							■															
	N501Y	RBD	-3.41							■															
	Y505H	RBD	-33.64							■															
	N679K	S1	29.55							■															
	N764K	S2	34.77							■															
	N856K	S2	3.41							■															
	Q954H	HR1	30.23							■															
DRB1*15:01	S371L	RBD	78.16	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	Y505H	RBD	-35.71	■	■	■			■		■	■		■	■	■		■	■		■	■	■
	Q954H	HR1	39.80	■	■	■			■		■	■		■	■	■		■	■		■	■	■
DRB1*15:02	S375F	RBD	42.86			■	■	■							■		■				■		
	Y505H	RBD	-45.92			■	■	■							■		■				■		
DRB5*01:01	A67V	NTD	-3.06	■			■	■		■	■	■								■				
	G142D	NTD	-25.51	■			■	■		■	■	■								■				
	Q498R	RBD	10.20	■			■	■		■	■	■								■				
	Y505H	RBD	-25.51	■			■	■		■	■	■								■				
	D614G	S1	26.53	■			■	■		■	■	■								■				
	N764K	S2	27.55	■			■	■		■	■	■								■				
	Q954H	HR1	-29.59	■			■	■		■	■	■								■				

Table 37: Mutations Predicted to have Extremely Adverse Effects on T-Cell Epitopes. The mutations in the RBD (Positions 454 and 455) are of interest, since mutations in the RBD effect both infectivity and immune evasion and are more likely to be naturally selected in the evolutionary process.

F2D	V6A	L8D	S12G	R454A
F2P	V6C	L8P	S12K	R454C
F2S	V6T	L10D	V16D	R454D
F4D	V6W	L10E	L18D	R454E
F4E	V6D	L10H	V42D	R454P
F4G	V6E	L10K	V42E	R454S
F4H	V6G	L10Q	L48D	R454T
F4P	V6H	L10A	L48E	R454W
F4S	V6P	L10C	V120D	R454G
F4T	L7D	L10G	V120P	R454H
L5D	L7E	L10N	V120S	R454N
L5P	L7K	L10P	V120T	R454Q
L5T	L7G	L10R	I203D	L455A
L5A	L7H	L10S	I203E	L455C
L5C	L7P	L10T	A243D	L455D
L5E	L7Q	S12D	V320D	L455E
L5S	L7W	S12E	V320E	L455Q
L5G	L7S			L455T

Table 38: Predicted Mutation Effects for Position 505 and for Position 455 on the Spike protein. Y505 was a hotspot residue in binding to the ACE-2 receptor. The Omicron variant has the mutation Y505H, which is predicted to detrimental effects across the alleles, specifically affecting DRB1*07:01 and DRB1*07:03.

Y505																				
		Possible Mutations																		
		A	V	L	I	F	W	M	P	G	S	T	C	Y	N	Q	K	R	H	D	E
Alleles	DRB1*07:01	-92.2	-0.9	11.2	11.2	8.6	-2.6	-0.9	-92.2	-92.2	-92.2	-55.2	-92.2		-92.2	-92.2	-92.2	-92.2	-92.2	-92.2	-92.2
	DRB1*07:03	-92.2	-0.9	11.2	11.2	8.6	-2.6	-0.9	-92.2	-92.2	-92.2	-55.2	-92.2		-92.2	-92.2	-92.2	-92.2	-92.2	-92.2	-92.2
	DRB1*08:01	-44.2	-11.6	-11.6	-11.6			-11.6	-44.2	-44.2	-44.2	-44.2	-44.2		-44.2	-44.2	-44.2	-44.2	-44.2	-44.2	-44.2
	DRB1*08:13	-43.7	-11.5	-11.5	-11.5			-11.5	-43.7	-43.7	-43.7	-43.7	-43.7		-43.7	-43.7	-43.7	-43.7	-43.7	-43.7	-43.7
	DRB1*08:17	-42.6	-9.9	-9.9	-9.9			-9.9	-42.6	-42.6	-42.6	-42.6	-42.6		-42.6	-42.6	-42.6	-42.6	-42.6	-42.6	-42.6
	DRB1*15:02	-45.9	-10.2	-10.2	-10.2			10.2	-45.9	-45.9	-45.9	-45.9	-45.9		-45.9	-45.9	-45.9	-45.9	-45.9	-45.9	-45.9

L455																				
		Possible Mutations																		
		A	V	L	I	F	W	M	P	G	S	T	C	Y	N	Q	K	R	H	D	E
Alleles	DRB1*08:01	-32.6	-53.5		-41.9	-4.7	-36.0	12.8	-55.8	-36.0	-57.0	-59.3	-32.6	15.1	-53.5	-59.3	-16.3	-8.1	-57.0	-90.7	-90.7
	DRB1*08:17	-102.0	-42.6		-2.0	-29.7	-64.4	35.6	-72.3	-73.3	-73.3	-102.0	-102.0	11.9	-70.3	-102.0	-59.4	-25.7	-73.3	-102.0	-102.0
	DRB1*11:01	-43.4	-44.6		-15.7	-38.6	-48.2	-10.8	-40.4	-45.8	-56.6	-59.0	-43.4	-38.6	-49.4	-56.6	-45.8	-39.8	-43.4	-95.2	-78.3
	DRB1*11:04	-63.9	-63.9		-8.4	-63.9	-63.9	20.5	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9
	DRB1*11:06	-63.9	-63.9		-8.4	-63.9	-63.9	20.5	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9
	DRB1*11:28	-28.7	-29.9		-2.3	-24.1	-33.3	2.3	-25.9	-31.0	-41.4	-43.7	-28.7	-24.1	-34.5	-41.4	-31.0	-25.9	-28.7	-106.4	-73.1
	DRB1*13:04	-42.2	-45.6		-5.6	2.2	-34.4	8.9	-47.8	-48.9	-20.0	-77.8	-42.2	2.2	-8.9	-35.6	-33.3	-5.6	3.3	-77.8	-77.8
	DRB1*13:05	-28.7	-29.9		-2.3	-24.1	-33.3	2.3	-25.9	-31.0	-41.4	-43.7	-28.7	-24.1	-34.5	-41.4	-31.0	-25.3	-28.7	-106.4	-73.1
	DRB1*13:11	-63.9	-63.9		-8.4	-63.9	-63.9	20.5	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9	-63.9
	DRB1*13:21	-62.9	-60.7		-4.5	-31.5	-67.4	1.1	-60.1	-65.2	-91.0	-93.3	-62.9	-58.4	-87.6	-93.3	-65.2	-59.5	-62.9	-122.5	-122.5
	DRB1*15:02	-51.0	-45.9		1.0	-7.1	-51.0	2.0	-48.0	-49.0	-49.0	-51.0	-51.0	-42.9	-45.9	-51.0	-51.0	-43.9	-49.0	-64.3	-63.3

Figure 18: Distribution of the Mutations across the Spike protein in the different variants. The Spike protein of the SARS-CoV-2 shown with mutations found in the Alpha, Delta, and Omicron variants.

Alpha: 69del, 70del, 144del, E484K, S494P, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H, K1191N

Delta: T19R, V70F, T95I, G142D, 156del, 157del, R158G, A222V, W258L, K417N, L452R, T478K, D614G, P681R, D950N

Omicron: A67V, 69del,70del, T95I, 142-144del, Y145D, 211del, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F

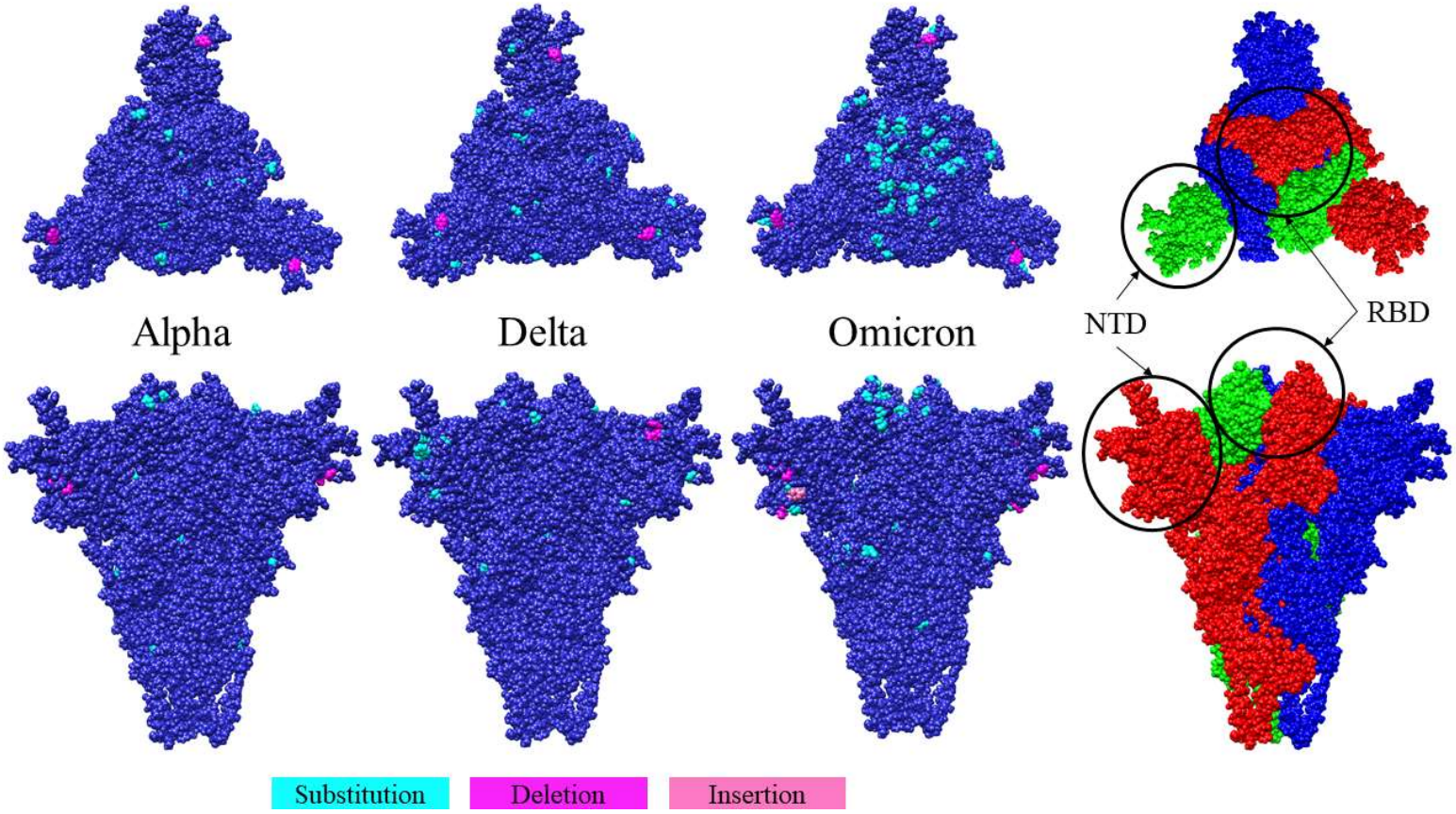


Figure 19: Distribution of the Surface Exposed Residues. This is image was created based on the numerical values of the Surface Accessible Surface Area (SASA) of each residue. It can be observed that only a few residues (0.46%) have very high exposure to the solvent, some residues (12.66%) have medium exposure, while a larger set of residues (70.13%) have a lower exposure of between 0 to 50 squared A°, and 16.71% of the residues have zero exposure.

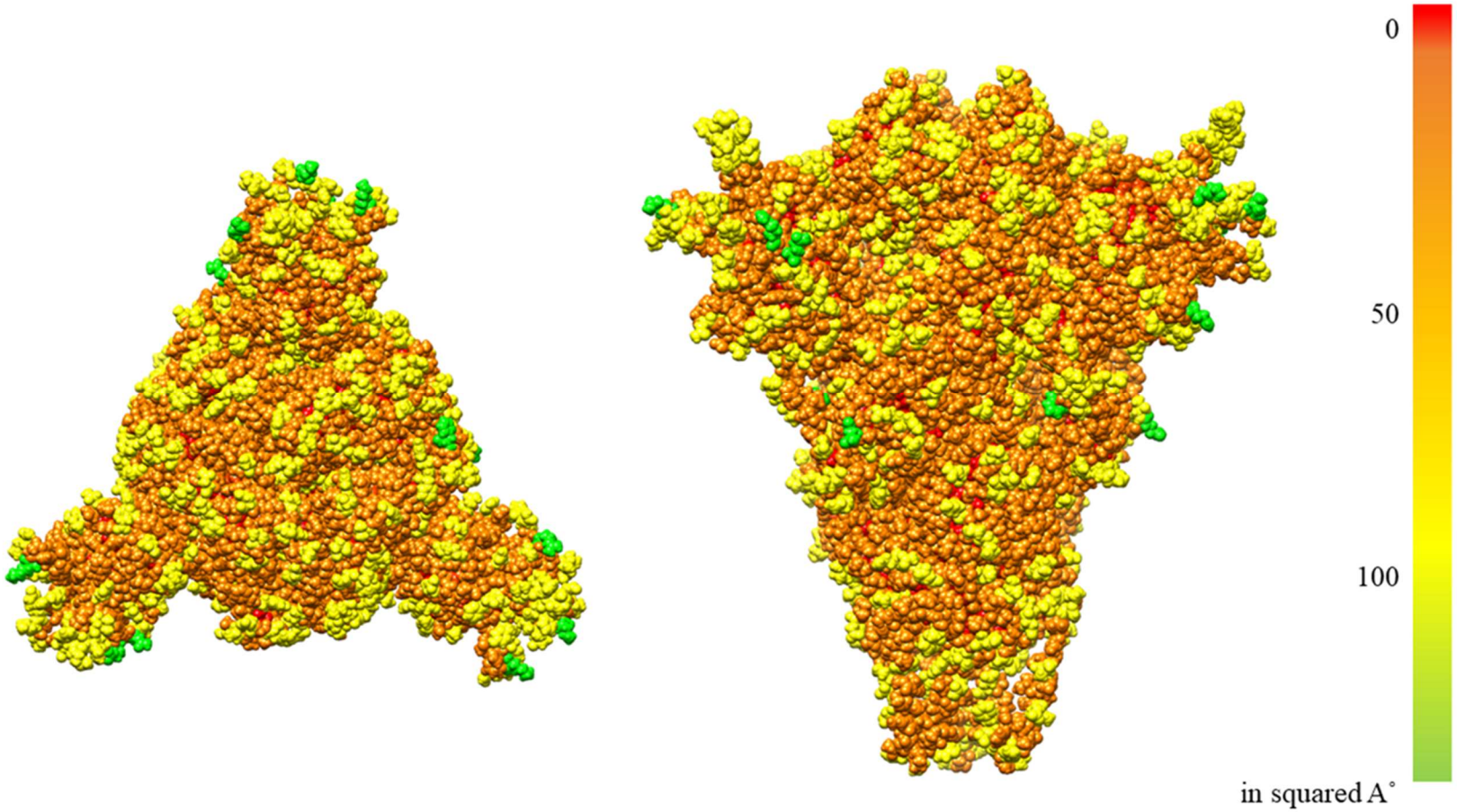
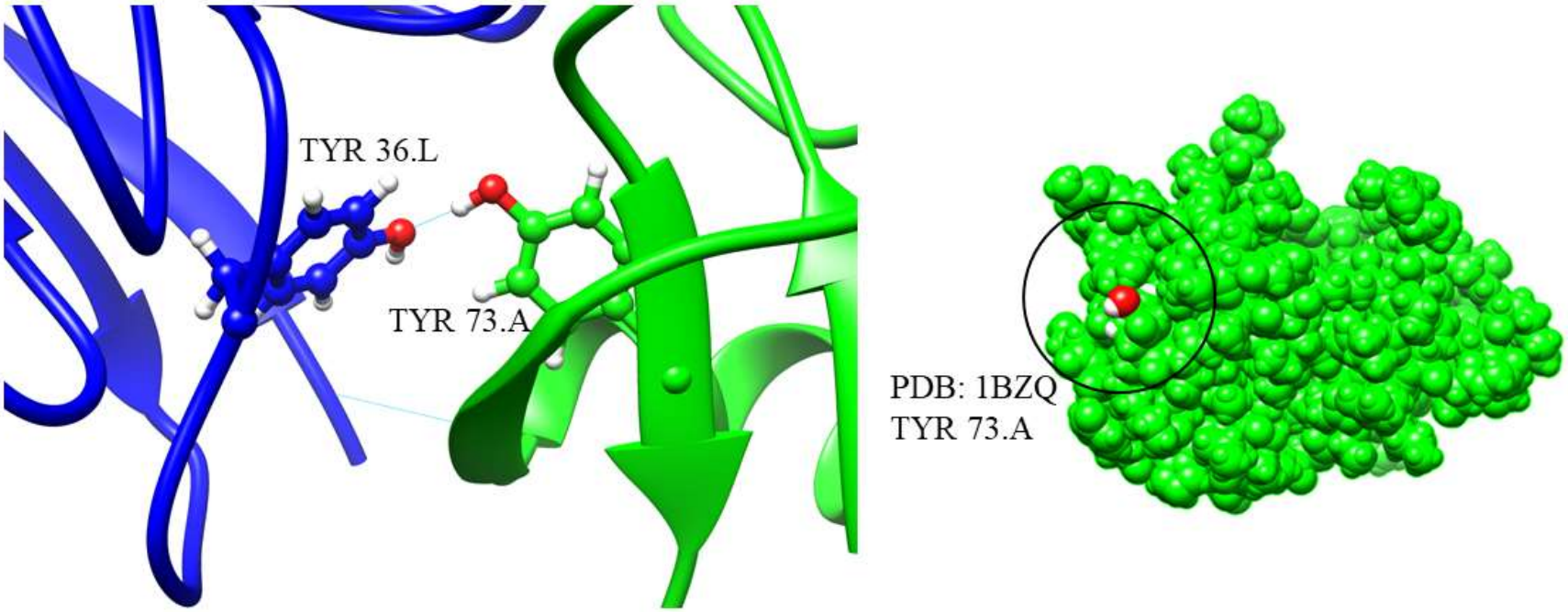


Figure 20: Residues with small solvent exposed surface area can contribute to binding. The Tyrosine residue in PDB 1BZQ has only the -OH group exposed, meaning a small SASA for the residue, yet Y73.A makes a strong hydrogen bond with Y36.L in the antibody.



References

[1] S. J. Gaudino and P. Kumar, “Cross-Talk Between Antigen Presenting Cells and T Cells Impacts Intestinal Homeostasis, Bacterial Infections, and Tumorigenesis”, *Frontiers in Immunology,* vol. 10. p. 360, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2019.00360

[2] D. R. Flower, “Vaccines: how they work,” *Bioinformatics for Vaccinology*, pp. 73–112, 2008.

[3] G. Li *et al.*, “Coronavirus infections and immune responses,” *Journal of Medical Virology*, vol. 92, no. 4. John Wiley and Sons Inc., pp. 424–432, Apr. 01, 2020. doi: 10.1002/jmv.25685.

[4] D. Wrapp *et al.*, “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation,” 2019.

[5] Y. Huang, C. Yang, X. feng Xu, W. Xu, and S. wen Liu, “Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19,” *Acta Pharmacologica Sinica*, vol. 41, no. 9. Springer Nature, pp. 1141–1149, Sep. 01, 2020. doi: 10.1038/s41401-020-0485-4.

[6] X. Xia, “Domains and functions of spike protein in sars-cov-2 in the context of vaccine design,” *Viruses*, vol. 13, no. 1. MDPI AG, Jan. 01, 2021. doi: 10.3390/v13010109.

[7] M. Hoffmann *et al.*, “SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor,” *Cell*, vol. 181, no. 2, pp. 271-280.e8, Apr. 2020, doi: 10.1016/j.cell.2020.02.052.

[8] L. Liu *et al.*, "Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike," *Nature*, vol. 584, no. 7821, pp. 450–456, 2020, doi: 10.1038/s41586-020-2571-7.

[9] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, no. 1, p. 40, 2008, doi: 10.1186/1471-2105-9-40.

[10] E. F. Pettersen *et al.*, "UCSF Chimera—A Visualization System for Exploratory Research and Analysis," *J Comput Chem*, vol. 25, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[11] H.-H. Bui *et al.*, "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, 2005, doi: 10.1007/s00251-005-0798-y.

[12] M. Nielsen, C. Lundegaard, and O. Lund, "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method," *BMC Bioinformatics*, vol. 8, no. 1, p. 238, 2007, doi: 10.1186/1471-2105-8-238.

[13] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites ," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, Dec. 2001, doi: 10.1093/bioinformatics/17.12.1236.

[14] P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower, "MHCPred: bringing a quantitative dimension to the online prediction of MHC binding," *Applied bioinformatics*, vol. 2, no. 1, pp. 63–66, 2003.

[15] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, and T. Li, "SVRMHC prediction server for MHC-binding peptides," *BMC Bioinformatics*, vol. 7, no. 1, p. 463, 2006, doi: 10.1186/1471-2105-7-463.

[16] P. Wang, J. Sidney, C. Dow, B. Mothé, A. Sette, and B. Peters, "A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach," *PLOS Computational Biology*, vol. 4, no. 4, p. e1000048, Apr. 2008.

[17] T. Sturniolo *et al.*, "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999, doi: 10.1038/9858.

[18] I. Dimitrov, P. Garnev, D. R. Flower, and I. Doytchinova, "MHC Class II Binding Prediction—A Little Help from a Friend," *Journal of Biomedicine and Biotechnology*, vol. 2010, p. 705821, 2010, doi: 10.1155/2010/705821.

[19] F. F. Gonzalez-Galarza *et al.*, "Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools," *Nucleic Acids Research*, vol. 48, no. D1, pp. D783–D788, Jan. 2020, doi: 10.1093/nar/gkz1029.

[20] A. Rambaut *et al.*, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature Microbiology*, vol. 5, no. 11, pp. 1403–1407, Nov. 2020, doi: 10.1038/s41564-020-0770-5.

[21] Á. O'Toole *et al.*, "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2," *Wellcome Open Research*, vol. 6, 2021, doi: 10.12688/wellcomeopenres.16661.1.

[22] Á. O'Toole *et al.*, "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool," *Virus Evolution*, vol. 7, no. 2, 2021, doi: 10.1093/ve/veab064.

[23] B. Lee and F. M. Richards, "The interpretation of protein structures: Estimation of static accessibility," *Journal of Molecular Biology*, vol. 55, no. 3, pp. 379-IN4, 1971, doi: https://doi.org/10.1016/0022-2836(71)90324-X.

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

[25] J. Shastri *et al.*, "Severe SARS-CoV-2 Breakthrough Reinfection With Delta Variant After Recovery From Breakthrough Infection by Alpha Variant in a Fully Vaccinated Health Worker ," *Frontiers in Medicine* , vol. 8. 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmed.2021.737007

[26] D. Planas *et al.*, "Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization," *Nature*, vol. 596, no. 7871, pp. 276–280, 2021, doi: 10.1038/s41586-021-03777-9.

[27] F. Amanat *et al.*, "SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD, and S2," *Cell*, vol. 184, no. 15, pp. 3936-3948.e10, 2021, doi: https://doi.org/10.1016/j.cell.2021.06.005.

[28] C. Xiangyang *et al.*, "A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2," *Science*, vol. 369, no. 6504, pp. 650–655, Aug. 2020, doi: 10.1126/science.abc6952.

[29] J. Zhang, T. Xiao, Y. Cai, and B. Chen, "Structure of SARS-CoV-2 spike protein," *Current Opinion in Virology*, vol. 50. Elsevier B.V., pp. 173–182, Oct. 01, 2021. doi: 10.1016/j.coviro.2021.08.010.

[30] S. Fatihi *et al.*, "A rigorous framework for detecting SARS-CoV-2 spike protein mutational ensemble from genomic and structural features," *Current Research in Structural Biology*, vol. 3, pp. 290–300, 2021, doi: https://doi.org/10.1016/j.crstbi.2021.11.002.

[31] M.-Y. Wang, R. Zhao, L.-J. Gao, X.-F. Gao, D.-P. Wang, and J.-M. Cao, "SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development", *Frontiers in Cellular and Infection Microbiology,* vol. 10. p. 724, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcimb.2020.587269

[32] C. Watanabe, Y. Okiyama, S. Tanaka, K. Fukuzawa, and T. Honma, "Molecular recognition of SARS-CoV-2 spike glycoprotein: quantum chemical hot spot and epitope analyses," *Chemical Science*, vol. 12, no. 13, pp. 4722–4739, 2021, doi: 10.1039/D0SC06528E.

[33] H. Lim *et al.*, "Hot spot profiles of SARS-CoV-2 and human ACE2 receptor protein protein interaction obtained by density functional tight binding fragment molecular orbital method," *Scientific Reports*, vol. 10, no. 1, p. 16862, 2020, doi: 10.1038/s41598-020-73820-8.

[34] T. N. Starr *et al.*, "Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding," *Cell*, vol. 182, no. 5, pp. 1295-1310.e20, 2020, doi: https://doi.org/10.1016/j.cell.2020.08.012.

[35] J. Verma and N. Subbarao, “Insilico study on the effect of SARS-CoV-2 RBD hotspot mutants’ interaction with ACE2 to understand the binding affinity and stability,” *Virology*, vol. 561, pp. 107–116, Sep. 2021, doi: 10.1016/j.virol.2021.06.009.

[36] J. A. Plante *et al.*, “Spike mutation D614G alters SARS-CoV-2 fitness,” *Nature*, vol. 592, no. 7852, pp. 116–121, Apr. 2021, doi: 10.1038/s41586-020-2895-3.

[37] S. Nersisyan, A. Zhiyanov, M. Shkurnikov, and A. Tonevitsky, “T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations,” *Nucleic Acids Research*, Aug. 2021, doi: 10.1093/nar/gkab701.

[38] A. Grifoni *et al.*, “Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals,” *Cell*, vol. 181, no. 7, pp. 1489-1501.e15, 2020, doi: https://doi.org/10.1016/j.cell.2020.05.015.

Chapter 6 Conclusion

The use of proteins in therapeutic developments is well-established. They are a clinically and commercially important class of treatment. Protein engineering and protein therapeutics have come a long way since the 1980s. There are several groups of therapeutic treatments that require proteins, e.g., hormones, blood factors, growth factors, antibody-based drugs, anticoagulants, enzymes, bone morphogenetic proteins, interferons, interleukins, and thrombolytics. Protein engineering provides the tools to customize existing proteins or to create novel proteins for all specific therapeutic needs. Both rational design and molecular evolution have provided scientists with knowledge about the structure-function relationships of proteins and the means to generate novel proteins with enhanced new properties.

The rapid growth of antibody therapeutics has created opportunities for antibody modification and antigen targeting against existing and novel diseases. Antibodies have high specificity and high affinity, making them highly efficient when designed for any targeted antigen. There are limitations to what is known about antibody binding and understanding antibody interfaces are crucial to designing new antibodies or modifying existing ones. Studying the relationship between the structures and binding mechanisms of PPIs is crucial to understanding how mutations can lead to novel pathogens that can evade the immune systems. This dissertation was built with projects directed toward generating data and knowledge that may be used in therapeutic design.

To be able to build tools that can better design protein interfaces, it was important to study their features and understand the possible paths to create change. The first project was based on studying the antibody-protein interfaces. A non-redundant database of 384 known antibody-protein structures was built and the interface properties of the structures were studied collectively. The analysis revealed that, on average, the seven most important residues contribute

almost 70% of the total binding energy, where the binding energy contributions follow an exponential decay. Charged and acidic amino acids were overrepresented in the significant residues. This dataset and the information about the interfaces were utilized to build the consecutive projects on this dissertation to better understand PPIs.

Protein structures have been studied extensively, with both current data (e.g., how particular sequences are shaped) and evolutionary data (e.g., how tolerated evolutionary amino acid mutations are in protein sequences). PPIs have not been studied to the same extent. It is also true that the relative importance of amino acids for stabilizing PPIs is likely different than for stabilizing protein structures. Therefore, the second project on this dissertation was based on understanding the effects of interface mutations on the binding function of antibodies. Protein structural mutation data is represented by similarity matrices, like PAM and BLOSUM, that was developed based on the rates of amino acid mutations in homologous protein sequences. The purpose of this project was to build similarity matrices for protein interactions using a systemic mutation analysis of the important amino acids to all the alternatives to calculate the corresponding changes in predicted binding energies. The goal was to gather knowledge about the mutational effects on antibody binding in a format that could be easily accessible for other studies regarding PPIs.

Early in the pandemic, a study led to an observation that SARS-CoV-2 evaded neutralizing SARS-CoV antibodies solely through the loss of favorable interactions. Theoretically, it was reasonable to hypothesize that loss of binding could occur due to either the loss of favorable interactions or the introduction of detrimental interactions. This observation motivated the third project on this dissertation, to examine how antigen mutations affect antibody binding in general. This project was also using a systemic mutational analysis, but it was more

specific to the binding interaction than the binding affinity, the residues selected for the mutational analysis were all the contact residues instead of just the important residues, and the mutational analysis was conducted for the antigenic residues only. Overall, the results show that antigen mutations can both introduce detrimental interactions and eliminate beneficial interactions, suggesting that the SARS-CoV-2 results of only eliminating beneficial interactions may be happenstance. Mutations in antigens resulting in new variants with reduced immunogenicity impact the efficacy of vaccines. Understanding how mutations in antigens affect antibody-antigen binding on specific interactions is an important path to creating vaccines and other therapeutic treatments which can provide enough protection against emerging variants.

As the pandemic progressed and new variants of the virus with different severity of infections in different countries were observed, the question of the efficacy of the vaccines developed against the new variants and whether people with different ethnic backgrounds was more vulnerable to the new variants arose. The final project of this dissertation was built on answering these questions. Immune responses to SARS-CoV-2 are mediated by the MHC Class II HLA genes. These genes vary from person to person and occur with different frequencies among ethnic groups. Of the four SARS-CoV-2 VOCs, the Beta and Gamma variants have not spread to the same extent as the Alpha, Delta, and Omicron variants. The calculations predicted they had fewer immune-evading mutations than Alpha, Delta, and Omicron. The analysis also shows a change in the distribution of the mutations in the variants, e.g., the Delta variant has more negative mutations than the Alpha variant and the majority are in the “tip” of the Spike protein (i.e., NTD and RBD), while the Omicron has more mutations in the RBD and NTD and in the S2 subunit than the Delta variant. This may explain why the Delta variant caused more severe infectivity than Alpha, and why, presumably, Omicron will cause increased infectivity but

less severity than Delta. There is not enough data about the Omicron variant in literature yet, but this is a reasonable conclusion to be made. The analysis also revealed mutations in the existing variants that are predicted to increase the immune response, they may possibly occur because they provide other evolutionary benefits. At the same time, the analysis was able to identify 1) mutation sites that can be a cause of concern for scientists regarding future variants of the virus and 2) ethnic groups that may be more vulnerable to current variants and future variants, requiring more targeted therapeutic treatments.

The high effectiveness of proteins in biological systems is a result of their mutational histories. Understanding the details of how mutations impact protein folding and interactions, and thus their functions in a biological system can provide insights towards improving the engineering of proteins. Mutational analyses are extensively used to study protein structures and functions for different purposes. For therapeutic purposes, such analyses can lead to more efficient engineering of proteins: from identifying beneficial interactions for improving protein binding to identifying how point mutations in pathogens can impact immune responses in humans. Each of the projects described here focuses on the effect of mutations on protein functions from a unique perspective, and each can contribute to the study of engineering proteins to meet different therapeutic needs. Identifying the features of the protein interfaces provided different perspectives of the interfaces that could be further studied. Quantifying the effects of point mutations on binding affinity and presenting that knowledge in a comprehensible format can contribute to the development of tools for improving the engineering of proteins. Studying the effects of mutations on the interface interactions provided an in-depth understanding of the antibody binding mechanisms to pathogens. Studying the effects of the mutations of a fast-mutating virus on people of different ethnic backgrounds has added novel directions to possible

therapeutic developments. This final study may also support possible directions to combating future pandemics more effectively.

The 2019 pandemic has proven that advances in therapeutic research are imminent to combat novel viruses. Mutant variants have caused unimaginable harm to humans and proved that one novel virus is not the only concern. The future holds scope for extensively engineering protein drugs to improve performance from a protein engineering perspective. The key pursuit for protein therapeutics is the development of better next-generation drugs, like enhanced antibodies and other proteins, with greater efficiency, more safety, or improved delivery. Tailoring antibodies for therapeutic applications is more established [1], given their high success rates [2]. Some antibody strengths and their limitations are broadly applicable to other protein therapeutics. One of the greatest strengths of antibodies in therapeutics is their ease in generating high affinity and high specificity to the desired target [3], yet their relatively large sizes limit their potential to penetrate tissue and localize to their targets.

Computational protein engineering tools have advanced to a point where they can guide protein therapeutic designs for optimizing favorable properties of proteins and creating novel activities. Understanding the strengths and limitations of protein therapeutics can pave how they may be improved. Structure-based, computational protein design can be potentially applied to antibody affinity maturation, stability improvement modification of PPIs, and minimization of protein aggregation [4]. Computational techniques have the advantage of their ability to work with more variables than experimental techniques. There is enough clinical and commercial success in protein therapeutics to motivate the use of computational techniques in their continued development into better drugs.

References

[1] D. S. Dimitrov, “Therapeutic proteins,” *Methods Mol Biol*, vol. 899, pp. 1–26, 2012, doi: 10.1007/978-1-61779-921-1_1.

[2] J. M. Reichert, “Monoclonal antibodies as innovative therapeutics,” *Curr Pharm Biotechnol*, vol. 9, no. 6, pp. 423–430, 2008.

[3] P. J. Carter, “Introduction to current and future protein therapeutics: A protein engineering perspective,” *Experimental Cell Research*, vol. 317, no. 9, pp. 1261–1269, 2011, doi: https://doi.org/10.1016/j.yexcr.2011.02.013.

[4] I. Hwang and S. Park, “Computational design of protein therapeutics,” *Drug Discovery Today: Technologies*, vol. 5, no. 2, pp. e43–e48, 2008, doi: https://doi.org/10.1016/j.ddtec.2008.11.004.