Advancing Antibody Development: An Investigation into Improved Classification and Prediction of Antibody-Antigen Binding Affinities Using AI Approaches

By

Adeel Nisar

(Registration No: 00000400908)

Department of Sciences

School of interdisciplinary Engineering and Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Advancing Antibody Development: An Investigation into Improved Classification and Prediction of Antibody-Antigen Binding Affinities Using AI Approaches

By

Adeel Nisar

(Registration No: 00000400908)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Bioinformatics

Supervisor: Dr. Mehak Rafiq

School of interdisciplinary Engineering and Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms Adeel Nisar Registration No 00000400908 of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: _____

DR. MEHAK RAFIQ
Assistant Professor
SINES, National University
of Science & Technology
H-12 Islamabad

Name of Supervisor: Dr Mehak

Date: 16/09/24

For HOD Sci

Signature of HoD with stamp: _____

Dr. Mian Ilyas Ahmad
HoD Engineering
Professor
SINES - NUST, Sector H-12
Islamabad

Date: 16/09/2024

Countersign by

Signature (Dean/Principal): _____

Dr SYED IRTIZA ALI SHAH
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Date: 18 SEP 2024

AUTHOR'S DECLARATION

I Adeel Nisar hereby state that my MS thesis titled "Advancing Antibody Development: An Investigation into Improved Classification and Prediction of Antibody-Antigen Binding Affinities Using AI Approaches" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Adeel Nisar

Date: 18/09/2024

DEDICATION

I dedicate this thesis to my exceptional parents, siblings, friends, and teachers whose unconditional love, support, and guidance led me to this world of accomplishment.

ACKNOWLEDGEMENTS

All praise is for **Almighty Allah**, the ultimate source of all knowledge. By His grace, I have reached this stage of knowledge with the ability to contribute something beneficial to His creation. My deepest respects are to **the Holy Prophet Hazrat Muhammad (PBUH)**, the symbol of guidance and fountain of knowledge.

I earnestly thank my supervisor, **Dr. Mehak Rafiq** for her keen interest, invaluable guidance, encouragement, and continuous support throughout my research journey. I am extremely grateful for her thought-provoking discussions, sound advice, and valuable suggestions. Her mentorship has enabled me to tackle problems more meaningfully and provided me with the resources to pursue my objectives diligently and sincerely.

I am also thankful to my co-supervisor **Dr. Salma Sherbaz** and GEC committee members **Dr. Masood ur Rehman Kiyani** and **Dr. Zartasha Mustansar** who guided me throughout my project and offered valuable feedback and suggestions to refine my thesis. Additionally, I acknowledge all the faculty members of SINES for their kind assistance at various phases of this research.

My gratitude extends to my colleagues in the Data Analytics lab. Special thanks to **Ayesha Iman, Ariba Abbasi**, **Amman Safeer**, and **Talha Zuberi** for their continuous help and feedback at every stage of the research.

TABLE OF CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

AIF	Amino Acid Interface Fitness
CDR	Complementarity-Determining Region
DNN	Deep Neural Network
ELISA	Enzyme-Linked Immunosorbent Assay
Fc	Fragment Crystallizable
Fab	Fragment Antigen-Binding
ITC	Isothermal Titration Calorimetry
MSE	Mean Squared Error
PDB	Protein Data Bank
PyRosetta	Python-based Rosetta Software Suite
R^2	Coefficient of Determination
RBF	Radial Basis Function
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
SIN	Significant Interaction Network
SPR	Surface Plasmon Resonance
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	Extreme Gradient Boosting
dMaSIF	Differentiable Molecular Surface Interaction Fingerprinting

ABSTRACT

The effective use of antibodies in personalized medicine hinges on their ability to bind strongly and specifically to their target antigens. This binding capability is critical for designing precise and successful therapeutic treatments tailored to individual patients. Identifying factors that predict binding affinity is essential to developing therapeutic antibodies with enhanced efficacy. However, experimental methods to measure antibody-antigen interactions are often costly and time-consuming, posing a challenge due to the diverse structures of antibody regions that influence binding. This study employs advanced machine learning and deep learning techniques, including Neural Networks, XGBoost, Random Forest, and Support Vector Machines (SVM), to classify and predict antibody-antigen binding affinities based on structural antibody data. Using data from antibody-antigen complexes sourced from the Structural Antibody Database (SAbDab), key features are extracted, including structural, energetic, and interaction-based metrics, to differentiate antibodies with high and low binding affinities. The models are evaluated using classification metrics (accuracy, precision, recall, F1 score) and regression metrics (MSE, R^2 score), with Random Forest emerging as the top performer in binding affinity prediction. By integrating neural networks with traditional machine learning approaches, this research provides a more efficient alternative to experimental methods, enhancing our understanding of molecular-level interactions between antibodies and antigens. This approach contributes to the optimization of therapeutic antibody design, further advancing the field of personalized medicine.

CHAPTER 1: INTRODUCTION

Sales of antibodies are projected to exceed $11 billion by 2024, reflecting their increasing significance in treating diseases like cancer, autoimmune disorders, and infections. Therapeutic antibodies, particularly monoclonal antibodies (mAbs), have evolved into essential tools in modern biopharmaceuticals due to their precision and effectiveness.

This chapter focuses on the structural aspects of antibodies, particularly the complementarity-determining regions (CDRs) that drive antigen-binding specificity. Understanding how variations in these regions influence therapeutic efficacy is crucial for optimizing treatments. Additionally, the importance of binding affinity, a key factor in antibody-antigen interactions, is discussed along with the limitations of traditional methods for predicting these interactions. Finally, the chapter introduces the role of artificial intelligence (AI) in overcoming these challenges by improving the accuracy and efficiency of binding affinity predictions, offering a more streamlined approach to therapeutic antibody development.

1.1 Therapeutic antibodies

Antibodies, also known as immunoglobulins, are specialized proteins generated by the immune system to identify and neutralize harmful pathogens such as bacteria and viruses[1]. Therapeutic antibodies represent a crucial class of biopharmaceuticals, distinguished by their ability to precisely target and treat a wide range of diseases, including cancer, autoimmune disorders, and infectious diseases. Since the advent of monoclonal antibodies (mAbs) over thirty years ago, the field has undergone significant advancements, particularly in antibody engineering, resulting in a broad spectrum of therapeutic modalities[2]. These antibodies are meticulously engineered to bind with high specificity to antigens, typically proteins present on the surface of pathogens or aberrant cells. This antigen-antibody interaction can neutralize the target, recruit immune effector cells, or

deliver cytotoxic agents directly to diseased cells, thereby amplifying the therapeutic efficacy while minimizing adverse effects[3].

The global market for therapeutic monoclonal antibodies is expected to surge, reaching an estimated $479.0 billion by 2028, with a robust compound annual growth rate (CAGR) of 14.1% starting from 2023. By 2024, fully human or humanized monoclonal antibodies are predicted to dominate, with 21 of the top 50 best-selling pharmaceuticals falling into this category, underscoring their critical role in modern therapeutics[3]. Key drugs such as Keytruda (pembrolizumab), Humira (adalimumab), and Dupixent (dupilumab) are anticipated to maintain their leading positions, reflecting the ongoing dominance of monoclonal antibody therapies in the pharmaceutical market[4].

1.2 Structure of antibody

1.2.1 Basic Structure

Antibodies are Y-shaped molecules made up of four polypeptide chains: two identical heavy chains and two identical light chains, which are structurally similar. These chains are held together by disulfide bridges, creating a flexible and functional configuration. Each antibody consists of two heavy chains, each approximately 50 kDa, and two light chains, each around 25 kDa. The heavy chains are longer and more complex than the light chains. Both the heavy (H) and light (L) chains are composed of variable (V) and constant (C) domains[5]. The variable regions, located at the tips of the Y-shaped structure, are where antigens bind. In contrast, the constant regions, positioned near the base of the antibody, determine the antibody's class or isotype and its functions, such as complement activation or receptor binding. Functionally, antibodies can be divided into two fragments: Fab (fragment antigen-binding) and Fc (fragment crystallizable). The Fab region includes the variable domains of both heavy and light chains, responsible for antigen binding. The Fc region, made up of the constant domains of the heavy chains, interacts with cell receptors and immune molecules to trigger immune responses [6].

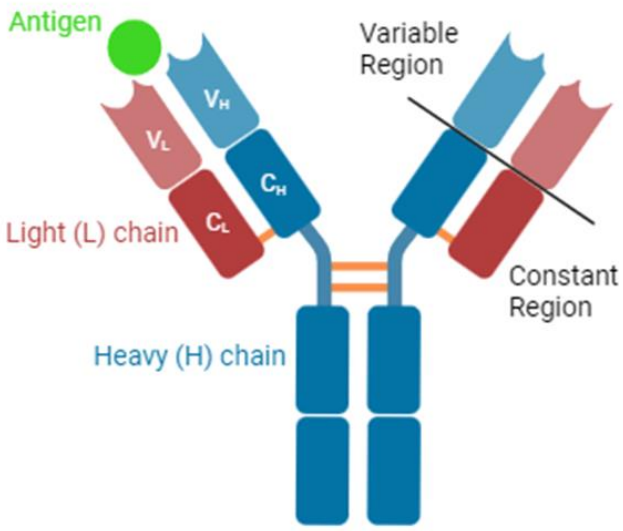

Figure 1.1: Structure of antibody

1.2.2 Complementarity-Determining Regions (CDRs)

The antigen-binding specificity of an antibody is largely dictated by the complementarity-determining regions (CDRs), which are highly variable loops within the variable regions of the heavy and light chains. Each variable region contains three CDRs, designated as CDR1, CDR2, and CDR3, with CDR3 being the most variable and critically important for determining the diversity of antigen recognition. These CDRs collectively form the paratope, the specific part of the antibody that binds directly to the epitope on the antigen[7]. The unique structure and amino acid sequence of the CDRs enable the antibody to recognize and bind to a broad range of antigens with remarkable specificity. The diversity of CDRs is generated through V(D)J recombination during B-cell development, a process that is further refined by somatic hypermutation, which introduces point mutations in the CDRs, enhancing antigen affinity and fine-tuning the immune response[8].

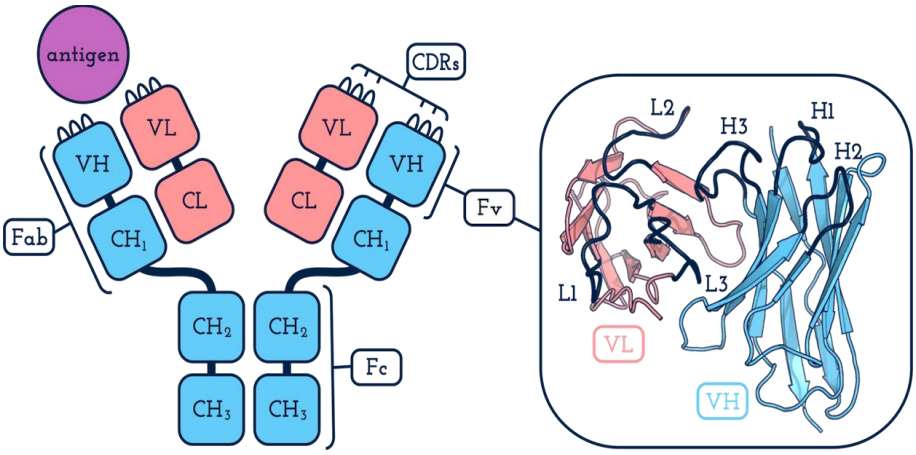

Figure 1.2: Structure of antibody, highlighting CDR region

1.2.2 Role of Structure in Therapeutic and Personalized Antibodies

The complex structure of antibodies, characterized by the antigen-binding specificity within the complementarity-determining regions (CDRs) and the functional versatility of the Fc region, is crucial to their use in therapeutic and personalized medicine. The precise configuration of the variable regions, especially the CDRs, enables the creation of monoclonal antibodies that can specifically target antigens linked to diseases such as cancer, autoimmune disorders, and infections. This structural precision not only boosts the effectiveness of antibody-based therapies by ensuring high-affinity binding to the target but also minimizes off-target effects, thereby reducing potential side effects[9]. Furthermore, engineering the Fc region to modulate immune responses—such as enhancing antibody-dependent cellular cytotoxicity (ADCC) or complement activation—expands the therapeutic capabilities of antibodies[2]. In personalized medicine, the structural adaptability of antibodies supports the development of customized treatments that align with an individual's unique genetic and antigenic profile, resulting in more effective and tailored therapeutic strategies. Consequently, the fundamental structure of antibodies is integral not only to their biological function but also to their expanding role in advancing precision medicine[10].

1.3 Binding Affinity and Its Critical Role

1.3.1 Binding Affinity Overview

Binding affinity refers to the degree of interaction between two molecules, particularly between an antibody and its corresponding antigen. This interaction is a crucial aspect of protein-protein interactions, where the antibody binds to a specific epitope on the antigen[11]. Binding affinity is typically quantified by the ratio of the association rate to the dissociation rate of the antibody-antigen complex in a solution. This is expressed through the dissociation constant (K_D), which is inversely proportional to the binding constant. A low K_D value indicates a high binding affinity, signifying that the antibody can effectively bind to the antigen even at low antigen concentrations[12]. The binding affinity is stabilized by various non-covalent forces, including hydrogen bonds, van der Waals forces, electrostatic forces, and hydrophobic interactions[11]. Affinity measures the strength of interaction between an epitope and an antibody's antigen binding site. It is defined by the same basic thermodynamic principles that govern any reversible biomolecular interaction:

$$K_A = \frac{[Ab\text{-}Ag]}{[Ab]\,[Ag]}$$

- $\mathbf{K_A}$ = affinity constant
- **[Ab]** = molar concentration of unoccupied binding sites on the antibody
- **[Ag]** = molar concentration of unoccupied binding sites on the antigen
- **[Ab-Ag]** = molar concentration of the antibody-antigen complex

In other words, K_A describes how much antibody-antigen complex exists at the point when equilibrium is reached.

1.3.2 Factors Influencing Binding Affinity

Several factors influence the strength of the interaction between an antibody and its antigen. The most critical factor is antigen fit, where the antibody's complementarity-determining regions (CDRs) must precisely align with the epitope on the antigen. This immunological specificity arises from the lock-and-key fit between the antibody's paratope and the antigen, meaning any alteration in the shape of either the antibody or the antigen can significantly impact binding affinity[13]. Additionally, environmental conditions such as pH, temperature, and ionic strength can affect the stability and strength of the antibody-antigen complex. Interference from other molecules present in the system can also impact binding, either by obstructing the interaction or by reducing the effective concentrations of the antibody or antigen in the solution, thereby altering the apparent binding affinity. Finally, post-synthesis modifications to the antibody, such as glycosylation, can further influence its binding profile and overall interaction strength[14].

1.3.3 Importance of Binding Affinity in Therapeutics

Therapeutic applications of antibody depend entirely on its ability to bind strongly to its antigen. The results in high binding capacity means that the antibody has the ability to strongly and specifically bind to its target antigen with few molecules thus improving the therapeutic index by reducing the amount of the drug that can cause side effects. For instance, whereby in cancer therapy, highly specific antibodies could selectively bind to tumor-associated antigens and hence enhance targeting and destruction of the cancer cell. In infectious disease treatments, the high affinity of drugs can enhance the blocking of pathogens' ability to infect the host or help to remove the pathogen from an individual's system[15]. Further, engineering antibodies for optimization of their binding affinities means that therapies can be created and designed that will work better and are specifically designed for the intended application. Hence, it is important and effective to understand and manage the binding affinity of antibodies for designing antibody-based therapeutics[3].

1.4 Challenges in Traditional Antibody Antigen Binding Affinity Prediction

1.4.1 Traditional Antibody-Antigen Binding Affinity Estimation Methods

Traditional techniques for determining the association constant between antibodies and antigens are crucial for understanding the nature and strength of these interactions. These methods are generally classified into experimental and computational approaches. Experimental techniques such as surface plasmon resonance (SPR), x-ray crystallography, isothermal titration calorimetry (ITC), and enzyme-linked immunosorbent assay (ELISA) provide direct measurements of the interactions under laboratory conditions, enabling the determination of binding constants and other kinetic parameters. On the other hand, computational methods utilize algorithms and simulations to model the interactions between antibodies and antigens, estimating binding affinities based on the structural and energetic properties of the molecules[16]. Each approach has its own advantages and limitations, making them complementary tools in the study of antibody-antigen interactions.

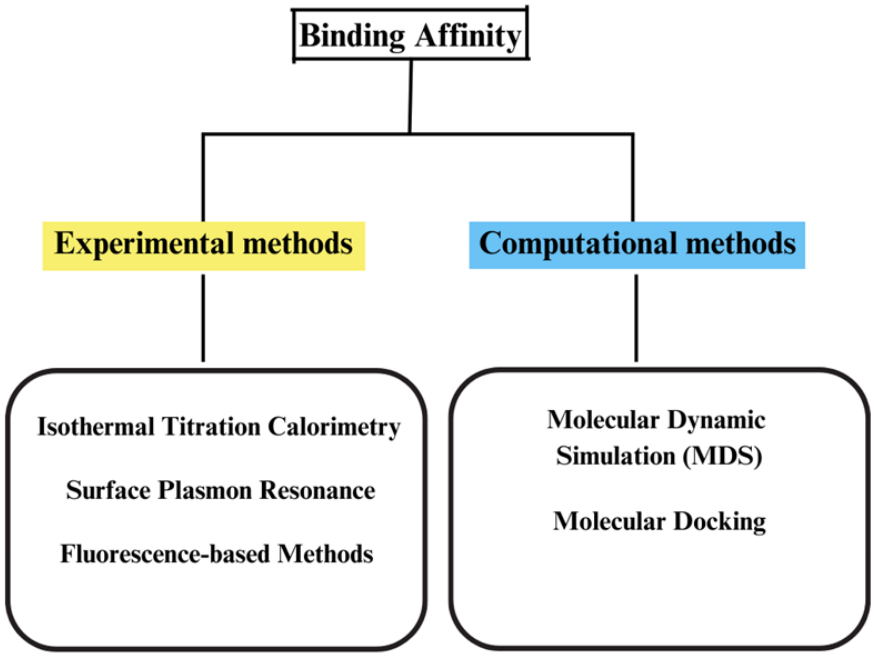

Figure 1.3: Binding affinity traditional methods

1.4.2 Experimental Challenges

A key challenge in experimental methods for estimating antibody-antigen binding affinity is the necessity for highly purified proteins and the need to maintain precisely controlled conditions. Variations in factors like temperature, pH, and ionic strength can significantly impact results, leading to discrepancies across different experiments. The dynamic nature of protein-protein interactions further complicates accurate binding measurements, especially with low-affinity interactions or rapid binding kinetics[17]. Techniques such as surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC), though powerful, demand sophisticated instrumentation and are often both time-consuming and resource intensive[18]. While enzyme-linked immunosorbent assay (ELISA) is popular for its simplicity and cost-effectiveness, it can encounter issues like nonspecific binding and artifacts from surface immobilization, potentially compromising the accuracy of binding affinity measurements. X-ray crystallography, another critical method, offers high-resolution structural insights but requires crystallization of the antibody-antigen complex, a process that is both challenging and time-consuming[19].

Additionally, the crystalline state may not fully capture the dynamic nature of interactions in solution, leading to potential discrepancies in observed binding affinities. Moreover, these methods often fail to accurately represent binding affinities in physiological contexts, as they typically assess interactions in isolated systems that do not account for the complexity of cellular environments[20].

1.4.3 Computational Challenges

Some challenges associated with molecular docking and molecular dynamics simulations in estimating binding affinity are also noteworthy. One significant difficulty lies in accurately reproducing the fine-grained and dynamic nature of protein-protein interactions. These methods often require simplifications and approximations, which can compromise the accuracy of predicted binding affinities. For instance, docking

algorithms may struggle with pose prediction when induced fit involves substantial conformational changes upon binding[21]. Additionally, while molecular dynamics simulations offer more detailed computational analysis, they come with higher computational costs and longer processing times, particularly when simulating large systems or extended timeframes. These methods are also highly sensitive to the choice of force fields and algorithms, which may not fully capture all aspects of antibody-antigen complex formation. Another major challenge is modelling solvent effects and ionic strength, which are crucial for accurate system predictions but are notoriously difficult to represent accurately[22].

These challenges underscore the complexities involved in estimating antibody-antigen binding affinity using both experimental and computational techniques, highlighting the ongoing need for advancements in technology and methodologies to achieve more accurate predictions.

1.5 Emergence of AI in Antibody Optimization

The integration of artificial intelligence (AI) into antibody optimization marks a significant advancement in biotechnology, revolutionizing the production of therapeutic antibodies in the healthcare sector. Recently, AI has been applied at various stages of antibody design and optimization, particularly during the early phases of drug discovery, where it aids in predicting binding affinity and structural conformation. These advancements are transforming the way antibodies are developed, enhancing their efficacy and specificity while minimizing side effects.

1.5.1 AI in Drug Discovery

AI has rapidly become an indispensable tool in drug discovery, particularly in the early stages of antibody development. By analyzing large datasets, AI algorithms can identify potential therapeutic targets and predict which antibodies might have the highest efficacy[23]. Machine learning models are used to sift through vast libraries of antibody sequences and structures, identifying those with the desired characteristics for further

development[24]. AI-driven platforms can also predict off-target effects and potential toxicity, helping to streamline the drug discovery process and reduce the time and costs associated with bringing new therapies to market. The ability of AI to process and analyze data at a scale and speed far beyond human capabilities is leading to more informed decision-making and the rapid identification of promising antibody candidates[25].

1.5.2 Advances in Structure Prediction using AI

Among the many impacts of AI on the antibody optimization process, structure prediction stands out as one of the most significantly transformed areas. A precise spatial understanding of an antibody is essential for grasping the architecture of related molecules. While traditional methods like X-ray crystallography and cryo-electron microscopy (cryo-EM) provide accurate results, they are both time-consuming and costly[26]. AI-driven approaches, such as DeepMind's AlphaFold, have demonstrated the potential to solve the structure prediction challenge with near-experimental accuracy[27]. These AI models utilize deep learning to predict the spatial arrangement of atoms in a protein based solely on its amino acid sequence, allowing scientists to anticipate the likely structure of antibodies before synthesis. This capability enhances the efficiency of the antibody design process by enabling rapid iteration and optimization based on predicted structures, ultimately leading to the development of more effective and precise antibody-based therapies[28].

1.6 Problem Statement and Solution

The rapid and cost-effective prediction of antibody-antigen binding affinity is essential for advancing the development of therapeutic antibodies. This section discusses the limitations of current methodologies and introduces an AI-driven approach as a solution to improve both accuracy and efficiency in predicting binding affinities.

1.6.1 Problem Statement

There is a need to improve the accuracy and efficiency of predicting antibody-antigen binding affinity. Traditional methods, whether experimental or computational, are often

slow and costly, which hampers the pace of developing and optimizing new therapeutic antibodies. Experimental techniques demand highly controlled conditions and substantial resources, while computational approaches, despite their power, can be both computationally expensive and may lack the precision required for complex interactions. The current landscape necessitates a faster, more cost-effective solution to predict binding affinities with high accuracy, which is crucial for advancing the design of therapeutic antibodies and the progression of personalized medicine.

1.6.2 Solution

The proposed solution involves performing a comparative analysis of various machine learning and deep learning algorithms to pinpoint the most accurate model for predicting antibody-antigen binding affinity. By systematically evaluating these AI models, the objective is to identify an optimal approach that improves prediction accuracy while minimizing the time and cost associated with traditional methods. This strategic application of AI will streamline the antibody design process, enabling the faster and more efficient development of therapeutic antibodies.

1.6.3 Objectives

Following are the objectives of this research:

- Analyze the structural features of antibody-antigen complexes to identify key determinants of binding affinity.
- Develop machine learning models to classify antibodies based on binding affinity and predict continuous affinity values.
- Validate the performance and robustness of the models using various evaluation metrics and data preprocessing techniques.

CHAPTER 2: LITERATURE REVIEW

2.1 Antibody-Antigen Interactions

Antibody-antigen interactions are fundamental to the immune response and are crucial in the development of therapeutic antibodies. These interactions hinge on the precise binding of an antibody to its specific antigen, a process primarily driven by the complementarity-determining regions (CDRs) of the antibody. The high specificity and affinity with which antibodies bind to antigens are key to their effectiveness in neutralizing pathogens or modulating immune responses[29]. Given the importance of these interactions, enhancing and understanding binding affinity is vital for developing effective therapeutic antibodies. High-affinity antibodies are particularly valuable in therapeutic applications because they can bind tightly to target antigens, ensuring a robust and precise immune response. This specificity is essential not only for therapeutic efficacy but also for minimizing off-target effects that could lead to adverse reactions[30].

As advancements in antibody design and optimization continue, structural considerations play an increasingly critical role. Understanding how antibodies interact with antigens at the molecular level directly impacts their binding affinity and specificity. The following section will delve into these structural aspects, highlighting their significance in designing more effective antibody therapeutics.

2.1.1 Structural Considerations in Antibody-Antigen Interactions

Antibody-antigen interactions are defined by several critical structural features that significantly influence binding efficacy. In 2022, Lee et al. analysed nearly 200 high-resolution antibody-peptide complex structures, highlighting the predominance of conformational epitopes as the primary antibody binding sites[31]. This finding underscores the importance of structural data in elucidating binding mechanisms. Typically, epitopes consist of 3-8 sequential patches of residues, with the longest patch

averaging 5-7 residues. The primary interactions at the interface are hydrogen bonds and hydrophobic interactions. Germline-encoded antibodies often utilize more hydrophobic interactions, whereas affinity-matured antibodies increasingly rely on polar contacts, such as hydrogen bonds, to achieve higher specificity. This idea is supported by extensive studies that examine the structural basis of antibody-antigen interactions, emphasizing the role of conformational flexibility in antigen recognition and the structural dynamics of the antibody affinity maturation process [32].

Notably, residues that form polar bonds and participate in hydrophobic clusters are often found at the boundaries of hydrophobic regions, which allows for specificity while balancing the reduced hydrophobic interface size during affinity maturation. Additionally, the framework regions and constant domains of antibodies can contribute to antigen binding through non-local and allosteric effects, as indicated by various studies analysing the properties of antibody-antigen interfaces[33].

To efficiently study these interactions, computational methods have become invaluable. In 2023, Madsen et al. employed graph convolutional neural networks to aggregate properties across local regions of proteins, learning predictive structural representations of binding interfaces that enhance the understanding of antibody-antigen interactions[34]. Attention layers in these models can explicitly encode the context of the antibody-antigen pair, capturing the specificity of their interactions. Furthermore, transfer learning from general protein-protein interaction data can provide a prior for the specific case of antibody-antigen interactions. These attention layers not only improve prediction performance but also offer biologically interpretable insights into the mechanisms of antibody-antigen interaction[33].

Overall, the structural basis of antibody-antigen recognition involves a complex interplay of conformational epitopes, hydrogen bonding, hydrophobic interactions, and contributions from both the variable and constant regions of antibodies, making it a fertile area for research and therapeutic development[32]. Given the complexity of these structural interactions, robust and comprehensive databases are essential for predicting antibody-antigen interactions, which we will explore in the next section.

2.2 Databases for Antibody-Antigen Interaction Prediction

Databases for antibody-antigen interaction prediction are essential for advancing therapeutic antibody development. SAbDab(Structural Antibody Database) offers the collection of antibody molecules from the PDB(Protein Data Bank). As highlighted by Schneider et al. (2022), SAbDab provides precise annotations and enhanced search capabilities to allow for the assessment of structures in terms of, for example, sequence similarity and antigen selectivity[35].

Another important source is the Antigen-Antibody Interaction Database, AgAbDb, that provides the determinants of Ag-Ab interactions as Kulkarni-Kale et al. (2014) have also pointed out. This database offers binding site residues and interacting pairs and as such is important for assessing the accuracy of algorithms for B-cell epitopes[36].

The PDBbind database is also important because it gathers the experimental binding free energy data (Kd, Ki, and IC50) for more than 23,000 protein–ligand complexes. PDBbind was founded in 2004 and it integrates energetic and structural information to support the investigations on molecular recognition. PDBbind includes protein-protein, protein-ligand, and protein-nucleic acid interactions, providing detailed structural and binding affinity data essential for molecular docking and scoring methods.[37].

Subsequent developments in the use of structural databases in combination with machine learning and deep learning models have improved the specificity of the antibody-antigen interactions. These databases are used by the researchers to derive extensive structural and binding data that is then used to train models that can effectively predict the binding affinities and to determine the key features that are likely to impact these interactions.

2.3 Machine Learning Approaches to Antibody-Antigen Interaction Modeling

About a decade ago, the initial efforts to improve the reliability of antibody-antigen interaction models began incorporating machine learning concepts that have since gained widespread attention, as noted by Tharakaraman et al. [38] During this period, researchers like Quinlan et al. (2017) featurized antibody-antigen interfaces by analysing various physicochemical and geometric properties of amino acids at these interfaces[39]. Using these features, several scoring functions were developed, and linear regression approaches were applied to differentiate between native antibody-antigen complex poses and decoy poses. This methodology enabled the prediction of a model for a broad-spectrum anti-Dengue virus antibody, leading to engineered mutations that increased its affinity to Dengue virus serotype 4 by more than 450-fold while slightly improving affinity to the other three serotypes, as demonstrated by Robinson et al[40]. This achievement stands as one of the earliest successful applications of computational methods based on simple machine learning principles, such as featurization and regression, resulting in a significant enhancement of antibody binding affinity[41].

Moreover, the physicochemical and geometric characteristics of these interfaces were also implicated in the molecular modelling of viral escape from neutralizing antibodies. This includes scenarios where escape occurs through multiple epistatic mutations, such as with the BA.1 Omicron variant of SARS-CoV-2, as described by Tharakaraman et al. and Miller et al[42].

Machine learning approaches to antibody-antigen interactions employ a diverse array of models to enhance the design and optimization of therapeutic antibodies, significantly advancing the field. Xu et al. (2021) introduced a method for featuring antibody-antigen interfaces using traditional machine learning techniques. Their approach involved representing the binding interface with a two-dimensional matrix, which was then analysed to identify complex interaction patterns that could distinguish antibody-antigen complexes from other protein-protein interactions[43]. This pioneering work laid a solid foundation for further research in this area.

Building on this, Wang et al. (2022) developed regression models incorporating structural features extracted from binding sites to estimate the binding energy between antibodies and antigens. Their work demonstrated the efficiency of supervised learning techniques, particularly linear regression and support vector regression, in measuring interaction strength[44]. This marked a significant advancement in designing antibodies with greater selectivity and affinity for their targets.

Recent developments underscore the critical role of feature selection in binding affinity prediction. In 2023, Zhang et al. compared the use of comprehensive feature sets—including various structural and physicochemical properties—with simpler models based solely on basic sequence information. Their findings revealed that while complex features can enhance predictive accuracy, simpler models like logistic regression and decision trees can also be highly effective, particularly when computational resources are limited[45]. This comparison highlighted the importance of balancing model complexity according to the specific needs of the study.

Additionally, the incorporation of network-based features has proven beneficial in improving binding affinity predictions. By modelling antibody-antigen interactions as networks, researchers can quantify the contributions of different residues to binding affinities[46]. For instance, in 2022, Makowski et al. proposed a model that combined linear discriminant analysis (LDA) with conventional machine learning approaches to simultaneously optimize site-specific and non-specific binding. This approach demonstrated the flexibility of machine learning architectures in interpreting complex biological interactions[45].

2.4 Deep Learning Approaches to Antibody-Antigen Interaction Modeling

Recent advancements in deep learning have significantly impacted the modeling of antibody-antigen interactions, particularly through the use of advanced algorithms for structure prediction and binding affinity prediction. In the domain of structure prediction, models like AlphaFold and RoseTTAFold have made remarkable strides. DeepMind’s AlphaFold, which utilizes a transformer-based system to predict protein structures from

amino acid sequences, has demonstrated exceptional performance in predicting 3D conformations. AlphaFold has generated a vast array of predicted structures, fueling further research across various biological fields[47]. Similarly, RoseTTAFold employs a three-track convolutional neural network (CNN) architecture that simultaneously processes sequence, distance, and 3D coordinate data, enabling fast and accurate prediction of protein structures, including those relevant to human health[48].

Deep learning has also been successfully applied to predicting antibody-antigen binding interactions. For instance, in 2021, Xu et al. developed a deep learning framework that utilized CNNs to analyze structural features of antibody-antigen interfaces, achieving high accuracy in distinguishing between complexes[49]. Building on this, Makowski et al. (2022) introduced AbAgIntPre, a deep learning model designed to predict antibody-antigen interactions from sequence data alone. This model achieved an impressive AUC of 0.82 on an independent test dataset, demonstrating the superior ability of deep learning to capture non-linear feature representations from sequence information[45].

Moreover, the integration of structural data with deep learning models has further enhanced predictive power. Besides these improvements, a more recent paper by Liu and his team put forward S3AI, which is a new antibody-antigen interaction prediction model, marked by interpretability and based on structural data together with the actual chemical rules encoded in the model. S3AI performed much better than the state-of-the-art methods and demonstrated good generalization when predicting the new antibody-antigen pairs making it applicable for large-scale antibody optimization and screening[50].

As a result, this paper has major improvements in predicting the antibody-antigen interactions while having major drawbacks as well. One major drawback is that it is heavily dependent on the quality of the training data; the model can be highly sensitive to the data that has been fed to it, and many potentially possible interactions may not be covered. S3AI has demonstrated OOD generalization, but it could be less performant when facing new combinations of antibodies and antigens not trained. Moreover, S3AI may not incorporate all the facets of the binding process because it has embedded structural knowledge and chemical priors. Training and applying such deep learning models also require heavy

computations that may also prove costly to researchers with meager resources. In summary, although S3AI has made a great improvement in the field of antibody-antigen interaction prediction, there are still many issues that need to be focused in the following research, such as the data quality, the model generalization, and the computational cost.

Despite the successes of deep learning in antibody design, there are notable limitations. However, there are several challenges associated with deep learning in antibody design. One major drawback is the need for high-quality training data, as the performance of deep learning applications is highly dependent on both the quantity and quality of the data used. Current models, such as AlphaFold and RoseTTAFold, may not perform optimally with certain protein families or specific conformational states, indicating that further refinement and calibration are necessary[51].

Despite these challenges, deep learning has significantly advanced the modelling of antibody-antigen interactions, including structure prediction and binding affinity estimation. These models employ sophisticated techniques to capture the complex interactions within biological systems, providing a strong foundation for more effective antibody design and screening. However, they also highlight areas that require continued improvement to enhance accuracy and applicability across a broader range of proteins and interactions.

2.5 Literature Gap

Despite the impressive advancements in modelling antibody-antigen interactions using machine learning and deep learning techniques in recent years, there are still areas in the literature that require further improvement. While early machine learning models have been instrumental in predicting binding affinities and optimizing therapeutic antibodies, they often rely on simplified features of the antibody-antigen interaction space and lack comprehensiveness. Although deep learning models have significantly enhanced structure prediction and demonstrated potential in improving binding affinity predictions, their performance is heavily dependent on the quality of the training data. These models often struggle with robustness when tested on unseen antibody-antigen pairs, particularly those

involving less commonly described protein families. Additionally, the computational demands of training and utilizing these models can be prohibitive, making them costly and difficult to deploy in larger-scale scientific studies.

Moreover, previous research has predominantly focused on accuracy enhancement, without adopting a holistic strategy that integrates various biological and structural data sources for more accurate and versatile predictions. The literature also highlights a critical need to improve the interpretability of deep learning models, as many of these approaches function as 'black boxes,' making it challenging to explain the rationale behind specific predictions. This lack of interpretability can hinder the refinement and validation of models, particularly in clinical settings where understanding the decision-making process is crucial.

In conclusion, while machine learning and deep learning have made significant strides in modelling antibody-antigen interactions, there is an urgent need for more accurate, explainable, and robust models that can be trained on diverse datasets and perform well across a range of biological scenarios. Addressing these gaps will be essential for advancing antibody development and effectively integrating AI into this field.

CHAPTER 3: METHODOLOGY

This chapter outlines the detailed methodology employed in this study, starting from data acquisition to the final evaluation of machine learning models. The chapter begins with the process of gathering data from the Structural Antibody Database (SAbDab) and continues with the thorough preprocessing steps applied to the dataset. It describes the removal of non-relevant entries, the structural optimization of antibody-antigen complexes, and feature extraction procedures that were crucial for building effective predictive models. Additionally, this chapter covers the feature selection techniques used to refine the dataset and enhance model performance. Finally, it introduces the machine learning algorithms implemented for classification and prediction tasks, followed by the evaluation metrics used to assess the accuracy and efficiency of these models.

3.1 Methodology Overview

This is the basic methodology flow used in this study. Data was first acquired from the Structural Antibody Database (SAbDab), where 470 antibody-antigen complexes were selected based on relevant criteria, such as the type of antigen and true affinity values. The dataset was then preprocessed to remove non-relevant entries like nanobodies and homologous antibodies to avoid potential biases in the model. Following this, the structural data in PDB format was processed, which involved the removal of water molecules and non-essential ligands, replacement of missing side chain atoms, and relaxation of complexes using PyRosetta to ensure optimal structural integrity.

After preprocessing, a subset of 16 high-importance features was extracted based on both complex computational methods and simpler structural properties. These features were then subjected to feature selection techniques, including RandomForestClassifier and correlation analysis, to improve the predictive power of the models. Finally, four machine learning models—XGBoost, Random Forest, Support Vector Machine (SVM), and Deep

Neural Networks (DNN)—were implemented to classify binding affinities and predict continuous binding affinity values. The models were evaluated using classification metrics (accuracy, precision, recall, F1 score) and regression metrics (MSE, R^2 score), ensuring a thorough evaluation of their performance in predicting antibody-antigen binding affinities.

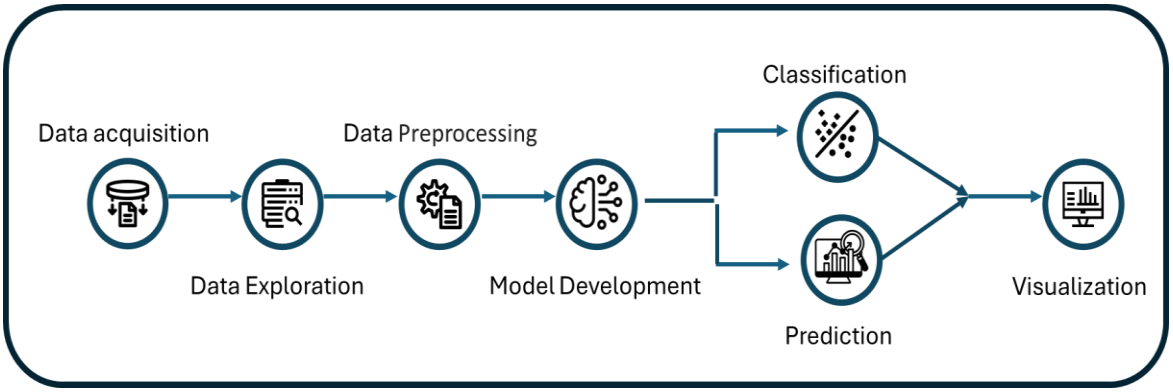

Figure 3.1: Methodology Overview

For this study, data of 470 antibody-antigen complexes was obtained from the Structural Antibody Database (SAbDab), a comprehensive resource maintained by the Oxford Protein Informatics Group for the exploration of antibody structures. The dataset was filtered to include only those complexes where the antigen type was a protein and the affinity value was marked as true, ensuring the selection of relevant antibody-protein interactions. The PDB files for these complexes were then downloaded in the Chothia numbering scheme, which reassigns residues in antibody variable regions based on structural landmarks, facilitating consistent analysis and comparison of antibody structures.

3.1.1 Data Characteristics

The dataset used for predicting antibody-antigen binding affinity contains data from 470 antibody-antigen complexes. Key characteristics include the `pdb_id` for each complex, the `affinity` value, and the `affinity_method` used to determine the binding affinity (e.g., SPR). The dataset also includes detailed information about the antigen chains

(`antigen_chain_1`, `antigen_chain_2`, etc.), the antigen's name, sequence, and associated heteroatoms. In addition to these biological details, the dataset features annotations based on the Chothia and North numbering schemes, which provide insight into the structure of the antibody's variable regions. Furthermore, the dataset includes categorical indicators for the presence of antigen chains and structural alignment, enabling a thorough understanding of the antibody-antigen interaction.

3.2 Data Preprocessing

This section discusses the data preprocessing pipeline, as depicted in the flowchart. It begins with the initial preprocessing of the affinity summary dataset, where non-relevant entries, such as nanobodies and homologous antibodies, are removed to avoid potential biases. The next step involves the preprocessing of PDB files, where water molecules and irrelevant ligands are systematically removed to maintain focus on the antibody-antigen interactions. Afterward, missing side chain atoms are replaced and repacked to ensure complete structural representation. Finally, the antibody-antigen complexes undergo a relaxation process, optimizing their structural integrity and energy landscape, making them suitable for further computational analysis, particularly in binding affinity predictions.

Figure 3.2: Affinity Data Preprocessing

3.2.1 Initial Preprocessing of Affinity Summary Dataset

In the initial preprocessing phase, the dataset was filtered to eliminate potential confounders and sources of data leakage, such as nanobodies and homologous antibodies. Nanobodies, which are single-domain antibodies derived from camelids, were excluded due to their consistently lower binding affinities compared to full-sized antibodies (median affinity of 10 nM vs. 3 nM). Removing nanobodies ensured that the model would not learn an artificial correlation between antibody format and affinity, which could skew performance results. Additionally, homologous antibodies—closely related antibodies with more than 95% sequence identity in the heavy chain—were removed to prevent data leakage during cross-validation, as these variants could lead to overfitting. This refinement resulted in a dataset of 356 unique, affinity-labeled antibodies.

3.2.2 Preprocessing of PDB files

The PDB files for each antibody-antigen complex, present in Chothia format, were downloaded from the Structural Antibody Database (SAbDab) and filtered according to the criteria established during the initial preprocessing, based on their pdb_id. This ensured that only relevant structures, aligned with the refined dataset of unique antibodies, were selected for further analysis. The structural data was then processed to extract key features of antibody-antigen interactions, following several essential steps:

3.2.2.1 Removal of Water Molecules and Ligands

To maintain a clear focus on the antibody-antigen binding interfaces, all water molecules and non-essential ligands were systematically removed from the PDB files. This step was crucial for eliminating any extraneous molecular elements that do not directly participate in the interaction. The Biopython package was employed for this task, ensuring that the processed structures were clean and suitable for further structural analysis, without interference from solvent or irrelevant ligands.

3.2.2.2 Replacement and Repacking of Missing Side Chain Atoms

To ensure complete structural representation of each antibody-antigen complex, missing side chain atoms were replaced and repacked using the PyRosetta package in Python. PyRosetta, a robust protein modeling and design toolkit, was employed to reconstruct missing atoms, especially in cases where side chain residues were incomplete or absent in the original PDB files. The software uses advanced algorithms to predict optimal side chain conformations, minimizing steric clashes and preserving the structural integrity of the protein. PyRosetta's side chain repacking functionality places the side chains in their most energetically favorable configurations, ensuring accurate modeling of the protein structure. This step was crucial for maintaining structural fidelity, as missing or inaccurately modeled atoms could distort the interpretation of antibody-antigen interactions, ultimately affecting the reliability of binding affinity predictions and other downstream analyses.

3.2.2.3 Pareto-optimal Relaxation of Complexes

Following the repacking of side chains, the entire antibody-antigen complexes were subjected to Pareto-optimal relaxation using PyRosetta. This was performed on an High Performance Computing cluster, where each complex took approximately 30 minutes on average to relax. The relaxation process utilized the FastRelax protocol in PyRosetta, which optimizes both the side chains and backbone of the complexes. The FastRelax protocol is crucial for ensuring that the structural models are energetically favorable while minimally perturbing the backbone (mean RMSD < 1 Å). A custom MoveMap was employed, allowing movements of both the backbone and side chains for thorough structural refinement. This balance between structural optimization and preservation of the native backbone is essential for maintaining biologically relevant conformations.

The importance of Pareto-optimal relaxation lies in its ability to correct any steric clashes or unfavorable interactions introduced during PDB processing (e.g., repacking of missing atoms), without introducing significant structural deviations. This refinement enhances the reliability of the structural models, making them more suitable for tasks such as binding affinity prediction, protein-protein docking, and antibody design. By optimizing the energy landscape while preserving native-like conformations, the relaxed complexes provide a high-quality foundation for further computational and experimental analyses.

3.3 Feature Extraction

In the study by Miller et al. (2023), a comprehensive set of features involved in antigen-antibody interactions was compared, examining over 200 distinct features spanning various structural, energetic, and interaction-based metrics. These features were derived from several advanced methodologies, including molecular surface analysis, interaction networks, and residue-level interactions, providing a deep insight into the factors contributing to binding affinity. Building upon their findings, the present work refines and focuses on a subset of these features that show the most promise for predicting binding affinity. Sixteen high-importance features were selected, and their performance

was compared across different machine learning models to optimize classification accuracy and binding affinity prediction.

Feature extraction is a critical step in analyzing antigen-antibody complexes, where two categories of features were utilized: complex feature-sets derived from advanced computational methods, and simple feature-sets computed directly from the antibody-antigen complexes. The aim was to identify characteristics relevant to binding affinity through both high-level molecular interactions and straightforward structural properties.

3.3.1 Complex Feature-Sets

The complex feature-sets are derived from established algorithms and computational workflows, providing detailed insights into molecular interactions at the antigen-antibody interface. These feature sets include:

3.3.1.1 Significant Interaction Network (SIN) Features

The SIN features quantify the strength of interactions between the antibody's paratope and the antigen's epitope. For example, SIN_L1_avg_epilope measures the average interaction strength of the L1 chain, while SIN_avg_paratope captures the overall strength of the paratope. These features help to pinpoint key interaction sites responsible for binding affinity.

3.3.1.2 PyRosetta-Based Features

PyRosetta provides energetics-based features crucial for evaluating molecular stability and compatibility. For instance, pyrosetta_Interaction_total_energy assesses the total interaction energy between the antigen and antibody, while pyrosetta_sc_total measures surface complementarity. These features offer insights into the structural and energetic favorability of the binding event.

3.3.1.3 dMaSIF-site Features

The dMaSIF-site (Differentiable Molecular Surface Interaction Fingerprinting) focuses on molecular surface interactions, particularly shape and electrostatic complementarity. dMaSIF_avg_paratope measures surface interactions across the paratope, and dMaSIF_L2_total captures the interaction score for the L2 chain, identifying crucial hotspots that drive binding affinity.

3.3.1.4 Amino Acid Interface Fitness (AIF) Features

AIF features evaluate the fitness of amino acids at the interface, providing scores like AIF_L1_avg_epilope, which measures L1 chain residues' fitness in interacting with the epitope, and AIF_avg_paratope, which scores the paratope's overall contribution to binding. These features help identify residues that stabilize the antigen-antibody complex.

3.3.2 Simple Feature-Sets

The simple feature-sets provide straightforward structural characteristics of the antibody-antigen complexes. These features, while computationally less intensive, contribute valuable information to understanding binding dynamics:

3.3.2.1 Amino Acid (aa) Counts Features

This feature set quantifies the types and frequencies of amino acids involved at the binding interface. Features such as aa_counts_IE measure interaction energy (IE) at the amino acid level, while aa_counts_RE capture residue-level contributions to binding.

3.3.2.2 Amino Acid Counts by CDR

These features analyze interactions within the antibody's complementarity determining regions (CDRs), key sites for antigen recognition. For instance, aa_counts_CDR_L3_charged_aromatic tracks charged and aromatic residue

interactions in the L3 chain, while aa_counts_CDR_L2_aromatic_aromatic captures aromatic interactions in the L2 chain.

3.3.2.3 Multivalent Interaction Features

Multivalent interactions significantly enhance binding strength by creating multiple points of contact between the antigen and antibody. Features such as multivalent_L1 and multivalent_L2 measure how many antigen sites interact simultaneously with the L1 and L2 chains of the antibody.

3.3.2.4 Ab Info Features

This set describes the structural characteristics of the antibody, such as CDR lengths. For instance, H3_length measures the length of the H3 region, known for its flexibility, while L3-9,10 identifies specific positions within the L3 chain that influence antigen binding.

3.4 Feature Selection

After the extraction of 16 initial features, feature selection was conducted to identify the most relevant features that contributed significantly to the model's predictive power. This step was essential to enhance the model's accuracy and reduce potential multicollinearity. Two primary methods were employed: RandomForestClassifier and Correlation Analysis.

3.4.1 RandomForestClassifier

The RandomForestClassifier is a tree-based ensemble learning algorithm that operates by constructing multiple decision trees during training. Each tree is built using a random subset of features and samples, and the final prediction is made based on the aggregated output of all the trees. One of the significant advantages of this method is its ability to provide an importance score for each feature, indicating how much each feature contributes to the overall predictive power of the model.

In this study, RandomForestClassifier was used to rank the importance of features. The importance score assigned by this classifier helps to identify which features are most influential in reducing prediction uncertainty and improving model accuracy. This method is particularly valuable for feature selection because it accounts for complex, non-linear relationships between features and the target variable. Features with high importance scores are prioritized for inclusion in the final model, while those with lower scores can be deprioritized or removed, optimizing the model's performance and simplifying its structure.

3.4.2 Correlation Analysis

Correlation analysis was employed to identify and address potential multicollinearity among the features. Multicollinearity occurs when two or more features are highly correlated, meaning they provide overlapping or redundant information. Including highly correlated features in a model can lead to overfitting, reduce interpretability, and increase model complexity without adding predictive value.

A correlation matrix was generated to quantify the relationships between features using Pearson correlation coefficients. The matrix displays the strength of the relationship between each pair of features, with coefficients ranging from -1 (strong negative correlation) to +1 (strong positive correlation). Features exhibiting high correlation (close to +1 or -1) were flagged for further investigation, and one feature from each correlated pair was removed to reduce redundancy. This step ensures that the remaining features contribute unique information, improving the generalizability and stability of the model.

3.5 Classification and Prediction Models for Binding Affinity

In this study, four machine learning models were implemented for both classification and regression tasks. The primary goal was to develop models capable of classifying binding affinities into low and high categories, as well as predicting continuous binding affinity values. For the **classification task**, a binding affinity threshold of **9** was set. Binding affinities above 9 were labeled as **high affinity**, while those below or equal to 9 were classified as **low affinity**. This binary classification framework was used to

distinguish between stronger and weaker molecular interactions, which is critical for assessing the efficacy of antigen-antibody binding.

For the **prediction task**, regression models were employed to estimate continuous binding affinity values, providing deeper insights into the strength of molecular interactions based on the selected features.

3.5.1 Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting algorithm that builds an ensemble of decision trees sequentially. Each tree in the model learns from the errors of the previous one, which helps minimize prediction errors using gradient descent. This model is designed to optimize performance by focusing on minimizing residual errors from previous iterations. XGBoost is highly effective for handling large, structured datasets and is robust against missing values and class imbalances.

In this study, XGBoost was implemented for both classification and regression tasks. Its ability to capture non-linear relationships between features, combined with boosting techniques, makes it highly suited for detecting subtle patterns in complex datasets. Regularization methods, such as shrinkage and subsampling, were applied to control overfitting, ensuring the model's generalization to new data.

3.5.2 Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using different random subsets of data and features. Each tree in the forest operates independently, and the final prediction is made by averaging the predictions (for regression) or by majority vote (for classification). The Random Forest model works by reducing variance and increasing prediction stability through its aggregation of multiple models, which mitigates overfitting.

In this study, Random Forest was applied to both classification and regression tasks. Its strength lies in its ability to handle high-dimensional data while automatically ranking

the importance of features. By averaging the predictions of multiple decision trees, Random Forest reduces the likelihood of overfitting, resulting in more stable and reliable predictions.

3.5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning model that finds the hyperplane that best separates data points into different classes (in classification) or fits a line (in regression). In SVM, the model works by maximizing the margin between the data points of different classes, ensuring that the model achieves the optimal separation of categories. For regression, Support Vector Regression (SVR) fits a hyperplane that captures the relationship between features and the target variable.

In this study, SVM was used to classify binding affinities as either high or low based on the threshold of 9. The Radial Basis Function (RBF) kernel was employed to handle non-linearity in the data, allowing the model to classify complex relationships between features and binding affinity categories. SVR was also applied to predict continuous binding affinity values, capturing non-linear trends in the data.

3.5.4 Deep Neural Network (DNN)

Deep Neural Networks (DNNs) are a type of artificial neural network composed of multiple layers of neurons that process information through non-linear transformations. The model works by learning from the input data in successive layers, where each layer captures increasingly complex patterns in the data. DNNs are particularly useful for modeling intricate relationships in data where simpler machine learning models may struggle.

In this study, a DNN was implemented using the PyTorch framework for both classification and regression tasks. The network architecture included multiple layers, consisting of an input layer, hidden layers, and an output layer. For the classification task, the DNN was trained to distinguish between high and low binding affinities. For the regression task, the network was trained to predict continuous binding affinity values. The

use of ReLU activation functions introduced non-linearity, allowing the model to capture complex interactions between molecular features.

3.6 Model Evaluation

After implementing the machine learning models for both classification and regression tasks, the next step involved evaluating the performance of each model. This evaluation process used different metrics depending on whether the task was classification or regression, ensuring a comprehensive assessment of model performance.

3.6.1 Classification Evaluation

For the classification tasks, where the goal was to classify binding affinities into low and high categories based on a threshold of 9, several metrics were employed to assess the effectiveness of the models:

- Confusion Matrix

A confusion matrix was generated for each model to examine the classification performance. The confusion matrix provides detailed insights into the number of true positives, true negatives, false positives, and false negatives. This helped in understanding where the models were performing well and where they were making errors, especially in misclassifying low and high binding affinities.

The confusion matrix offers a visual representation of the classification outcomes, allowing for the identification of areas where the model might need improvement, such as reducing false positives or false negatives.

- Accuracy

This metric measures the proportion of correct classifications out of the total number of predictions, providing an overall measure of the model's performance.

$$\text{Accuracy} = \frac{\text{No of correct prediction}}{\text{No of total prediction}}$$

- Precision

Precision represents the ratio of true positive classifications to the total number of positive predictions, reflecting how well the model avoids false positives.

$$\text{Precision} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Positive}}$$

- Recall

Recall (or sensitivity) measures the proportion of actual positive instances that the model successfully identifies, highlighting the model's capacity to detect high binding affinities.

$$\text{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Negative}}$$

- F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a balanced measure in cases of uneven class distribution.

$$F1_{score} = \mathbf{2} * \frac{\boldsymbol{Precision * Recall}}{\boldsymbol{Precision + Recall}}$$

3.6.2 Prediction Evaluation

For the regression tasks, where the objective was to predict continuous binding affinity values, the following metrics were used to evaluate the models:

- Mean Squared Error (MSE)

MSE measures the average of the squared differences between the predicted and actual values. A lower MSE indicates that the model's predictions are closer to the true binding affinity values.

- R^2 Score

The R^2 score represents the proportion of variance in the binding affinity that the model explains. A score closer to 1 indicates a better fit to the data.

- Scatter plot

Scatter plots were generated to compare predicted vs. actual binding affinity values. Points close to the diagonal line on the plot indicate that the model’s predictions are well-aligned with the actual data, offering a visual confirmation of the model's accuracy.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents the key findings from the study, focusing on the machine learning models developed for classifying and predicting antigen-antibody binding affinities. It begins by outlining the steps taken in feature engineering, where a selection of the most relevant features was made to enhance the models' predictive capabilities. Further engineering techniques were applied to these features to optimize model performance, ensuring robust and accurate results.

The analysis then delves into the evaluation of four machine learning models—XGBoost, Random Forest, Support Vector Machine (SVM), and Deep Neural Network (DNN)—which were implemented to classify binding affinities into low and high categories, as well as predict continuous binding affinity values. Each model's performance was rigorously assessed using evaluation metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and Mean Squared Error (MSE).

The discussion highlights how these models performed in distinguishing between low and high binding affinities, as well as in predicting affinity values. By comparing the results across different models, this chapter offers insights into the most effective approaches for modeling antigen-antibody interactions and predicting binding strengths, contributing to advancements in computational antibody design.

4.1 Data Overview

The dataset used in this research consisted of antigen-antibody complexes, each identified by unique PDB (Protein Data Bank) IDs. The primary data included structural information in the form of PDB files, which contained detailed representations of the antigen-antibody interactions. After the initial data collection, the PDB files were processed to conform to the Chothia numbering scheme, ensuring consistency in the representation of antibody variable regions. This step was crucial for accurately mapping the antibody's complementarity-determining regions (CDRs) and their interactions with the antigen.

Once the PDB files were filtered based on the preprocessing criteria, they underwent relaxation using PyRosetta, a robust computational toolkit designed for protein modeling. The relaxation process optimized the structural conformations of the antigen-antibody complexes, minimizing steric clashes and improving the overall energy profile of the models. The output from this preprocessing and relaxation pipeline provided high-quality, relaxed PDB files, which served as the foundation for subsequent feature extraction and model training.

4.1.1 Experimental Methods Used to Measure Antibody Affinity

In antibody research, several experimental techniques are employed to measure binding affinities between antibodies and antigens. The choice of method can influence the accuracy and precision of the binding measurements, which are essential for building reliable computational models. The dataset used in this study contains a variety of experimental methods, each contributing to the collection of affinity data.

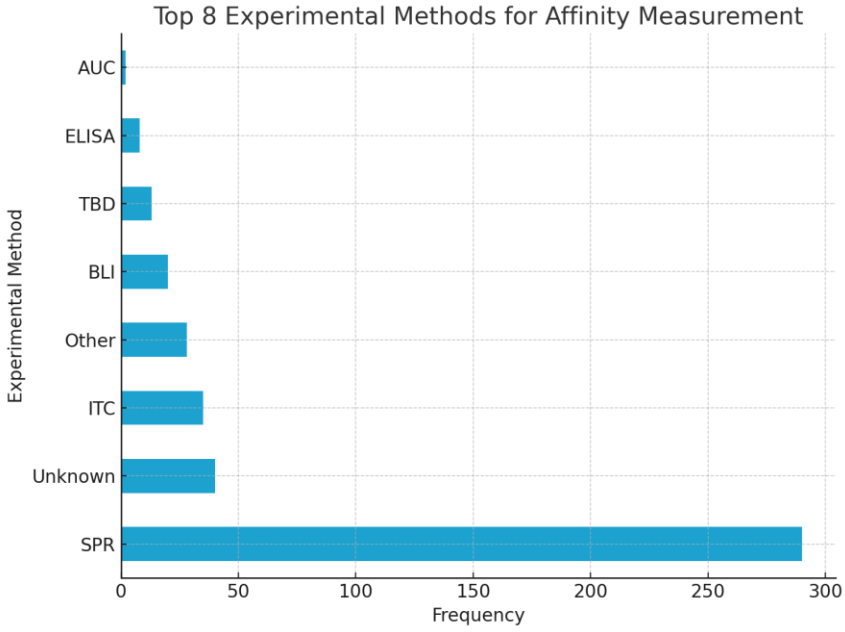

Figure 4.1: Bar chart of antibody affinity methods

The bar chart presented above illustrates the frequency of the **top 8 experimental methods** used to measure antibody affinity in the dataset. The most prominent method, **Surface Plasmon Resonance (SPR)**, is shown to be the dominant technique, significantly surpassing other methods such as **ITC (Isothermal Titration Calorimetry)** and **Biolayer Interferometry (BLI)**. Lesser-used methods, such as **AUC** and **ELISA**, are also represented but occur with much lower frequency.

These methods play a crucial role in evaluating the strength of interactions, which is essential for guiding therapeutic antibody design and enhancing the prediction of binding affinities in computational models.

4.1.2 Engineered vs. Non-Engineered Antibodies

The dataset used in this study includes both engineered antibodies, which have been modified to enhance properties such as binding affinity and stability, and non-engineered antibodies, which occur in their natural form. Differentiating between these two types is crucial as it provides insight into how molecular engineering influences antibody function and their potential therapeutic applications.

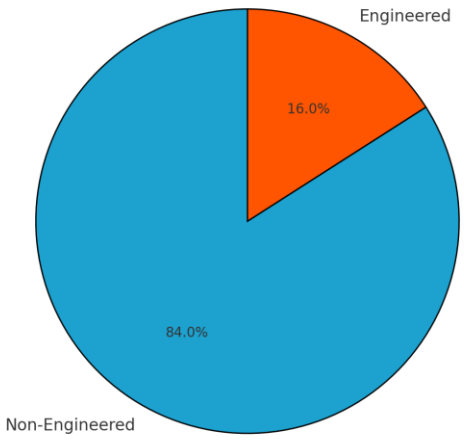

Figure 4.2: Pie chart of engineered and non-engineered samples

This pie chart illustrates the proportion of **engineered** and **non-engineered** samples in the dataset.

Engineered samples refer to antibodies that have been modified or optimized to enhance certain properties, such as binding affinity or stability, whereas **non-engineered samples** are natural, unmodified antibodies that occur in their original form without any alterations.

This emphasizes the distinction between engineered and non-engineered antibodies in the dataset, providing insights into how modifications may affect binding affinity, which is crucial for designing effective therapeutic antibodies and for understanding the role of naturally occurring antibodies in immune responses.

4.1.3 Affinity Distribution Across the Dataset and Non-PPI Subset

The strength of antigen-antibody binding is measured by binding affinity, and this study examines how these affinities are distributed across the dataset. The dataset includes both protein-protein interaction (PPI) and non-PPI interactions, which may display different binding characteristics due to the nature of the molecular interactions involved.

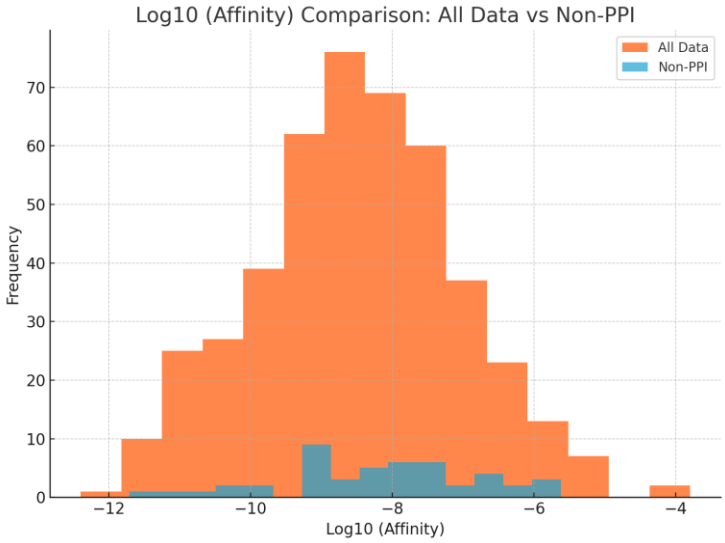

Figure 4.3: Histogram of affinity comparison

The above histogram compares the **Log10 (Affinity)** values between the entire dataset (in orange) and the **non-PPI subset** (in blue). Both distributions center around similar affinity ranges, but the non-PPI subset exhibits a slightly narrower distribution compared to the overall data.

The peak of the distribution is around a log10 affinity of -10 for both groups, indicating that most interactions in both the complete dataset and the non-PPI subset have moderate to strong binding affinities. The non-PPI data shows a more concentrated distribution, with fewer extreme values compared to the entire dataset.

This histogram is important because it highlights the distribution of binding affinities across the entire dataset compared to the non-PPI subset, providing insights into the variability and strength of antibody-antigen interactions. The comparison allows for a better understanding of how specific types of interactions (non-PPI) may differ in binding characteristics from the overall dataset, with key trends in affinity measurements being identified and analyzed.

4.1.4 Resolution Distribution of Structural Data

Structural resolution is a key factor in determining the quality and accuracy of the molecular models used in computational predictions. Higher-resolution structures provide more reliable details on molecular interactions, which are crucial for building accurate models of antibody-antigen binding. The dataset includes structures with varying resolution levels, which directly affect the precision of the binding affinity predictions.

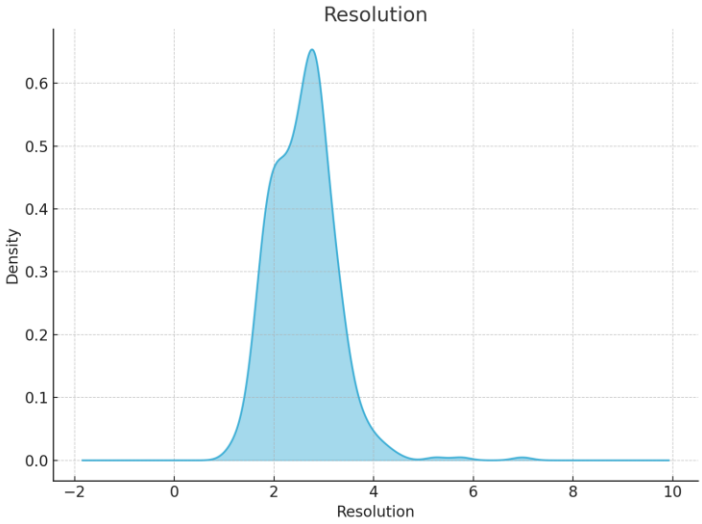

Figure 4.4: Density plot of distribution of resolution values

The refined density plot above represents the distribution of **resolution** values in the dataset. The x-axis shows the resolution, with lower values indicating higher-quality structural data, and the y-axis represents the density of occurrences for each resolution value.

Resolution refers to the quality of the structural data, typically measured in Ångströms. Lower resolution values indicate higher-quality structures. This plot shows a clear peak around **2 Ångströms**, signifying that most structural data in the dataset is of high quality. The density gradually decreases as resolution values increase, indicating fewer entries with lower-quality data.Understanding the resolution distribution is crucial for assessing the reliability and accuracy of the structural data used in antibody-antigen binding affinity predictions.

4.2 Feature Engineering

In their study, Miller et al. (2023) explored over 200 distinct features related to antigen-antibody interactions, encompassing a wide range of structural, energetic, and interaction-based metrics. These features were derived from advanced techniques such as molecular surface analysis, interaction networks, and residue-level interaction analysis, offering valuable insights into the factors influencing binding affinity. Their comprehensive approach highlighted the complexity of these interactions and the need for carefully selected features when predicting binding affinity[52].

Building on their work, this study focused on refining and narrowing down the most relevant features for the specific task of predicting binding affinity. A total of **16 high-importance features** were selected based on their potential to provide meaningful predictive power. These features were chosen to balance complexity and interpretability while ensuring they captured the key structural and energetic aspects of antigen-antibody binding.

Additionally, **further feature engineering techniques** were applied to these selected features to enhance their predictive capability. These techniques included normalization, scaling, and transformations to improve model performance and ensure that the features were appropriately aligned for the machine learning algorithms used. By preprocessing the features in this way, we ensured that the models could effectively capture the relationships between the features and the binding affinities, thus optimizing the predictive accuracy.

Table 4.1: Features involved in binding affinity with overview

Feature Set	Feature Name	Overview
SIN Features	SIN_L1_avg_epilope	Measures the average interaction strength between the L1 chain of the antibody and the epitope.
SIN Features	SIN_avg_paratope	Captures the average interaction strength across the entire paratope (the antigen-binding site on the antibody).
Amino Acid Count feature	aa_counts_IE	Represents interaction energy (IE) of amino acids at the antigen-antibody interface.
Amino Acid (aa) Counts Features	aa_counts_RE	Refers to residue-level counts, capturing how often certain residues appear at the interface.
Amino Acid Counts by CDR	aa_counts_CDR_L3_charged_aromatic	Tracks the presence of charged and aromatic residues in the L3 chain of the antibody.
Amino Acid Counts by CDR	aa_counts_CDR_L2_aromatic_aromatic	Measures aromatic residues in the L2 chain that interact with aromatic residues in the antigen.
PyRosetta-based Features	pyrosetta_Interaction_total_energy	Quantifies the total interaction energy between the antibody and antigen.
PyRosetta-based Features	pyrosetta_sc_total	Measures shape complementarity (sc_total) between the antigen and antibody.
dMaSIF-site Features	dMaSIF_avg_paratope	Measures the average surface interaction score across the entire paratope.
dMaSIF-site Features	dMaSIF_L2_total	Captures the total molecular surface interaction score for the L2 chain.
AIF Features	AIF_L1_avg_epilope	Represents the fitness of residues in the L1 chain that interact with the epitope.
AIF Features	AIF_avg_paratope	Reflects the overall fitness of the paratope residues in terms of their contribution to the antigen-antibody interface.

Multivalent Interaction Features	multivalent_L1	Measures the number of epitope sites with multivalent interactions involving the L1 chain.
Multivalent Interaction Features	multivalent_L2	Tracks multivalent interactions for the L2 chain.
Length and Positional Features	H3_length	Represents the length of the H3 region, which often plays a crucial role in antigen recognition and binding.
Length and Positional Features	L3-9,10-A	Refers to specific residues in the L3 chain (positions 9 and 10).

After extracting the initial 16 features, the next step involved selecting the most relevant features that contributed significantly to the model's predictive power. Feature selection was performed using **tree-based methods** and **correlation analysis** to identify the most influential features and reduce multicollinearity.

4.2.1 Tree-Based Feature Importance:

One of the primary methods used for feature selection was **RandomForestClassifier**, a tree-based ensemble model that naturally ranks features based on their importance in improving the model's prediction accuracy. The **RandomForestClassifier** assigns an importance score to each feature, which is proportional to its contribution to reducing uncertainty (impurity) during the decision-making process.

These features were identified as the most crucial predictors and were selected for further analysis. The bar plot (Figure 1) illustrates the importance of each feature, with **pyrosetta_sc_total** and **pyrosetta_Interaction_total_energy** having the highest importance scores, indicating their significant role in predicting binding affinity.

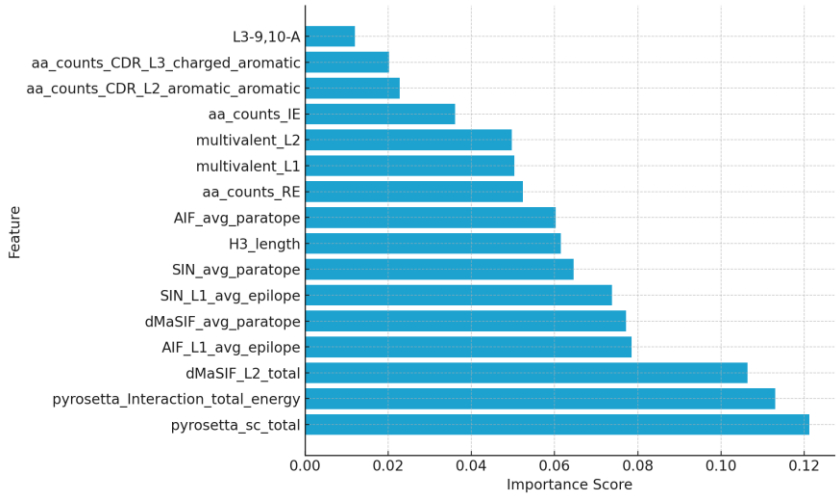

Figure 4.5: Feature importance bar plot

This bar plot visualizes the importance scores assigned to each feature by the **RandomForestClassifier**. Based on the bar plot, the features with the highest importance scores are **pyrosetta_sc_total** and **pyrosetta_Interaction_total_energy**, which reflect key aspects of molecular interactions, specifically energy-related metrics. These scores indicate their critical role in determining binding affinity and protein-ligand stability. Additionally, features related to surface interaction, such as **dMaSIF_L2_total** and **AIF_L1_avg_epilope**, are also highlighted for their contribution to how well a ligand fits and binds to a protein's surface, further affirming their importance in the model.

The **epitope** and **paratope** features, including **SIN_L1_avg_epilope** and **dMaSIF_avg_paratope**, are ranked highly as well, indicating that these regions, crucial for ligand-receptor interaction, play an essential role in binding properties. Furthermore, the **multivalent_L1** feature is shown to be more influential than **multivalent_L2**, which suggests that multivalency in this region significantly impacts binding. Features associated with charge and aromatic amino acid counts, such as **aa_counts_CDR_L3_charged_aromatic** and **aa_counts_CDR_L2_aromatic** also contribute by influencing molecular affinity through hydrogen bonding and hydrophobic interactions. The **H3_length** feature, despite being structural, shows importance, suggesting that the length of this region affects protein-ligand interactions and overall binding properties.

.

4.2.2 Correlation Analysis:

In addition to feature importance scores, a **correlation matrix** (Figure 2) was generated to examine the relationships between the features. The correlation matrix helps identify multicollinearity, which occurs when two or more features are highly correlated. In such cases, the redundant features can be removed or combined to simplify the model and improve performance.

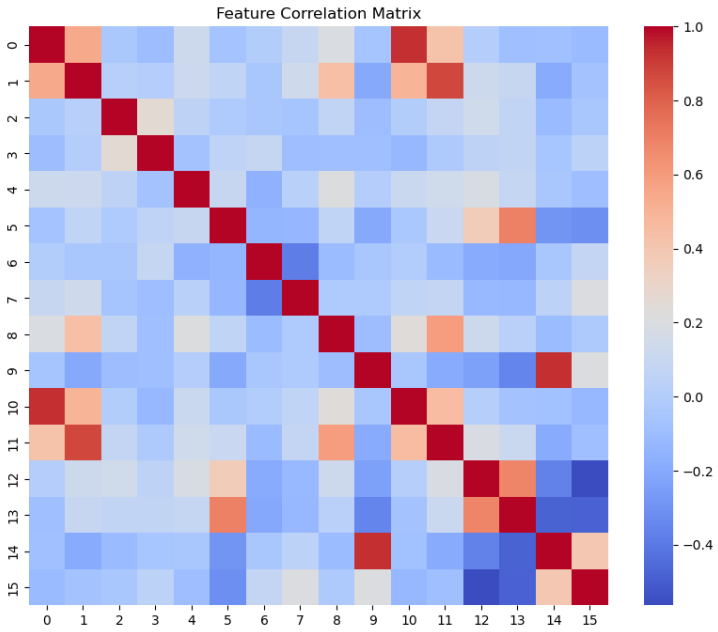

Figure 4.6: Feature correlation matrix

This heatmap represents the correlation between the features, with colors representing the strength of the correlation. By analyzing this matrix, we identified potential multicollinearity issues, and redundant features were removed to improve the model's robustness.

The correlation matrix further defends the selected features by showing generally low to moderate correlation among them, indicating that they contribute unique information and are not redundant. Although features like **pyrosetta_sc_total** and

pyrosetta_Interaction_total_energy exhibit some correlation due to their similar energy-based nature, they still provide distinct insights into different aspects of molecular scoring, which justifies including both. Features like **aa_counts_CDR_L2_aromatic_aromatic** and **multivalent_L1** show minimal correlation with others, demonstrating that they offer complementary information. The independence of key structural and functional metrics, such as **SIN_L1_avg_epilope** and **dMaSIF_L2_total**, underscores their unique contributions to the model.

By combining **RandomForestClassifier** feature importance and **correlation analysis**, we selected 10 features from the initial 16. These selected features were then used in the final model to ensure both high predictive power and minimal multicollinearity, ultimately improving the overall model performance.

- SIN_L1_avg_epilope
- aa_counts_IE
- aa_counts_CDR_L3_charged_aromatic
- aa_counts_CDR_L2_aromatic_aromatic
- pyrosetta_Interaction_total_energy
- pyrosetta_sc_total
- dMaSIF_L2_total
- AIF_L1_avg_epilope
- multivalent_L1
- H3_length

4.3 Data Preparation

Data preparation is a critical step in ensuring that the dataset is properly formatted and cleaned before any machine learning techniques are applied. This process includes handling missing values, normalizing feature scales, and addressing class imbalances to ensure more accurate and reliable results during model training and evaluation.

4.3.1 Handling Missing Data

In real-world datasets, missing values are a common challenge and can significantly impact the performance of machine learning models. To address this issue, missing values in the dataset were handled by replacing them with the median value of each respective feature. This imputation strategy helps to preserve the overall distribution of the data without introducing bias that could occur with methods like mean imputation. The median is particularly robust to outliers, making it an effective choice for datasets with skewed distributions.

Using the median ensures that the missing values are substituted in a manner that maintains the dataset's statistical integrity, allowing the model to operate on complete data and reducing the risk of data distortion.

4.3.2 Normalization of Features Using Scaling

Machine learning models often require features to be on a similar scale to ensure effective learning, especially in algorithms that rely on distance measurements or gradient-based optimization. In this dataset, feature scaling was applied using **StandardScaler**. StandardScaler transforms the features by subtracting the mean and scaling them to unit variance, resulting in a standard normal distribution with a mean of 0 and a standard deviation of 1.

This ensures that all features contribute equally to the model and prevents features with larger ranges from dominating those with smaller ranges. By applying feature scaling,

the model's learning process is optimized, particularly in cases where gradient-based algorithms or distance-based models are used.

4.3.3 Data Balancing Using SMOTE

Handling imbalanced data is a key concern in many machine learning tasks, as it can skew model performance and reduce its ability to generalize across different data distributions. In this context, the dataset under consideration showed an imbalance between the samples, with significantly more instances of low binding affinity compared to high binding affinity. This imbalance can lead to biased learning outcomes, where the model may favor the majority group, resulting in poor generalization for the under-represented group (Figure 1).

To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic data points for the under-represented class, helping to balance the dataset. This technique reduces the risk of overfitting to the majority class by ensuring that both groups are adequately represented during training.

Before applying SMOTE, the data distribution was as follows:

- Low Binding: 272 samples
- High Binding: 156 samples

After applying SMOTE, the dataset increased to 552 samples, achieving the following distribution:

- Low Binding: 276 samples
- High Binding: 276 samples

Figure 1 below shows the class distribution before and after applying SMOTE. The application of SMOTE ensured that the data was balanced, providing the model with a more evenly distributed dataset for learning purposes.

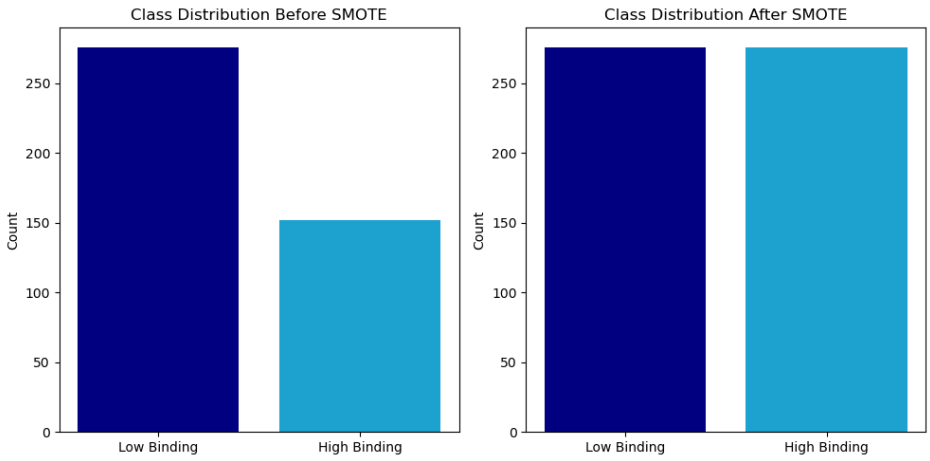

Figure 4.7: Data Distribution Before and After SMOTE

By balancing the dataset through SMOTE, the model becomes better equipped to handle variability across the dataset, improving its overall learning capabilities and minimizing bias that can occur due to the original imbalance.

4.4 Classification Model

Various machine learning algorithms were employed to learn patterns from the data and optimize performance metrics through model tuning and evaluation. The objective of this step was to develop models that generalize well to unseen data, ensuring robust and reliable performance in real-world applications. The focus was on building classification models to predict low and high binding affinities. The dataset was categorized into two classes based on a binding affinity cutoff value of 9, where values above the cutoff were labeled as high binding and those below as low binding. This binary classification allowed for the development of machine learning models that could differentiate between the two binding categories.

4.4.1 Model Selection

To achieve this goal, four different machine learning models were selected: XGBoost, Random Forest, Support Vector Machine (SVM), and a Neural Network trained

using PyTorch. Each model was chosen for its distinct strengths in handling complex datasets and optimizing performance.

XGBoost, a gradient-boosting algorithm, was chosen for its efficiency in building sequential decision trees and its ability to handle structured and imbalanced datasets. Random Forest, an ensemble learning method, builds multiple decision trees and combines their predictions, providing robustness and accuracy while reducing the risk of overfitting. SVM, known for its use of hyperplanes to separate data into classes, was selected due to its effectiveness in binary classification tasks where the margin between classes is clearly defined. Finally, a Neural Network implemented using PyTorch was chosen to capture complex non-linear relationships in the dataset, offering flexibility in feature interactions.

To optimize each model's performance, we employed GridSearchCV, a systematic hyperparameter tuning method. This approach also included k-fold cross-validation, which helps prevent overfitting and ensures that the models generalize well to new, unseen data.

4.4.2 Model results and evaluation

The performance of the selected machine learning models was evaluated using several key metrics: accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). These metrics provide a comprehensive view of how well each model performed on the dataset, allowing us to assess the trade-offs between true positives, false positives, and overall prediction quality. Below is a summary of the results achieved by each model:

Table 4.2: Classification evaluation report

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
XGBoost	0.973	1.000	0.942	0.970	0.996
RandomForest	0.937	0.941	0.923	0.932	0.991

Model	Accuracy	Precision	Recall	F1 Score	AUC
SVM	0.748	0.722	0.750	0.736	0.815
Neural Network	0.757	0.727	0.769	0.748	0.810

Each model's performance was optimized through a rigorous process of hyperparameter tuning using GridSearchCV, and the results presented here are reflective of their best configurations.

In the following sections, we will delve into the specific results for each model, discussing their confusion matrices, the best hyperparameters identified, and how they performed overall.

4.4.2.1 Extreme Gradient Boosting

The XGBoost model performed effectively in distinguishing between low and high binding affinities. The confusion matrix indicates the accuracy of the model in classifying the two categories. Specifically:

- **Low Binding (0):** The model correctly classified 59 out of 59 instances, achieving 100% accuracy for this class.
- **High Binding (1):** Out of 52 actual high binding instances, the model correctly predicted 49, resulting in a true positive rate of approximately 94.2%. Only 3 instances were misclassified as low binding, demonstrating strong performance in identifying high binding affinities.

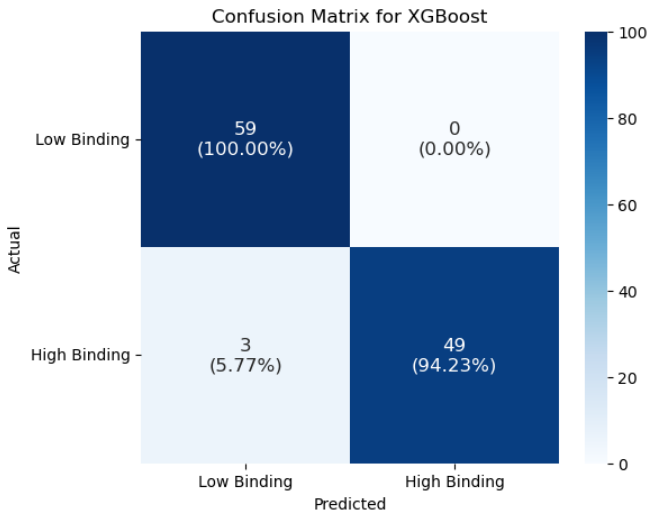

Figure 4.8: Confusion matrix for XGBoost

The best hyperparameters for the XGBoost model, as identified through GridSearchCV, are as follows:

- **colsample_bytree:** 0.6: This parameter controls the fraction of features (columns) to be randomly sampled for each tree. Using 60% of the features ensures diverse trees, preventing overfitting.

- **gamma:** 0.3: The gamma parameter adds regularization, ensuring a minimum loss reduction required to make a further split on a leaf node. This helps in controlling the complexity of the model by reducing overfitting.

- **learning_rate:** 0.1: The learning rate determines how much the model changes with each iteration. A value of 0.1 is moderate and allows the model to converge slowly, thus improving generalization.

- **max_depth:** 6: This restricts the maximum depth of each tree. A depth of 6 prevents the model from becoming too complex, reducing the risk of overfitting.

- **min_child_weight:** 1: This parameter controls the minimum sum of instance weight needed in a child. A lower value allows for more splits, improving the model's ability to capture subtle patterns.

- **n_estimators:** 100: The number of trees in the ensemble. A value of 100 ensures enough complexity while keeping the computation time reasonable.

- **subsample:** 0.8: This parameter controls the fraction of observations (rows) used to train each tree. Using 80% of the data helps in making the model robust by preventing it from overfitting to specific data samples.

The XGBoost model displayed strong predictive capability, particularly in identifying high binding affinities, with an accuracy of 97.3%. The model's hyperparameter tuning contributed to its balance between precision and recall, ensuring that it generalizes well to unseen data while minimizing overfitting.

4.4.2.2 Random Forest

The Random Forest model also exhibited strong performance in distinguishing between low and high binding affinities. The confusion matrix reflects the model's accuracy in classifying the two categories:

- **Low Binding (0):** The model correctly classified 56 out of 59 instances, resulting in a high accuracy of 94.92% for this class. Only 3 instances were misclassified as high binding.

- **High Binding (1):** Out of 52 actual high binding instances, the model correctly predicted 48, achieving a true positive rate of approximately 92.31%. Only 4 instances were misclassified as low binding.

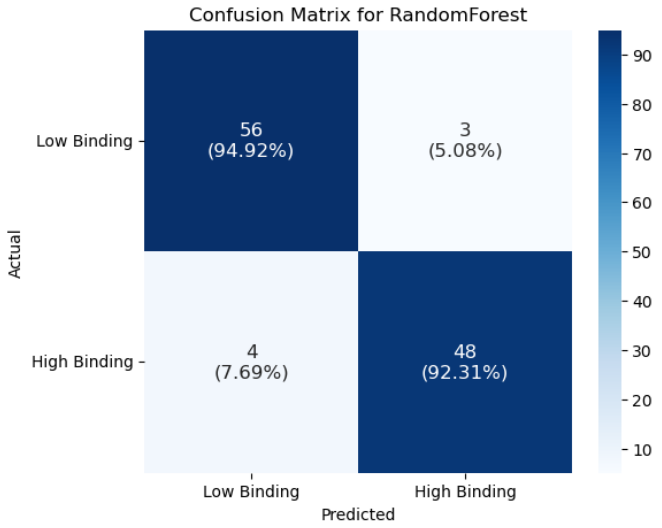

Figure 4.9: Confusion Matrix for random forest

The best hyperparameters for the Random Forest model, as identified through GridSearchCV, are as follows:

- **max_depth:** 10: This limits the depth of each tree, helping the model avoid becoming too complex and thus reducing the risk of overfitting.
- **min_samples_leaf:** 2: This parameter defines the minimum number of samples required to be at a leaf node. Setting it to 2 ensures that the trees do not become too finely tuned to the data, thus improving generalization.
- **min_samples_split:** 5: This controls the minimum number of samples required to split an internal node. A value of 5 encourages the model to only make splits where there is a meaningful division in the data, preventing overfitting.
- **n_estimators:** 150: This defines the number of trees in the forest. A higher number of trees, like 150, generally provides more robust and stable predictions without significantly increasing computation time.

The Random Forest model displayed excellent classification performance, particularly in identifying low binding affinities. With an overall accuracy of 93.7%, the model

effectively generalized to the dataset. Its hyperparameters balanced the complexity of the trees while ensuring robust prediction.

4.4.2.3 Support vector machine

The SVM (Support Vector Machine) model demonstrated moderate performance in distinguishing between low and high binding affinities, as reflected in the confusion matrix:

- **Low Binding (0):** The model correctly classified 44 out of 59 instances, achieving an accuracy of 74.58% for this class. However, 15 instances were misclassified as high binding.
- **High Binding (1):** Out of 52 actual high binding instances, the model correctly predicted 39, resulting in a true positive rate of 75%. However, 13 instances were misclassified as low binding, indicating room for improvement in capturing high binding affinities.

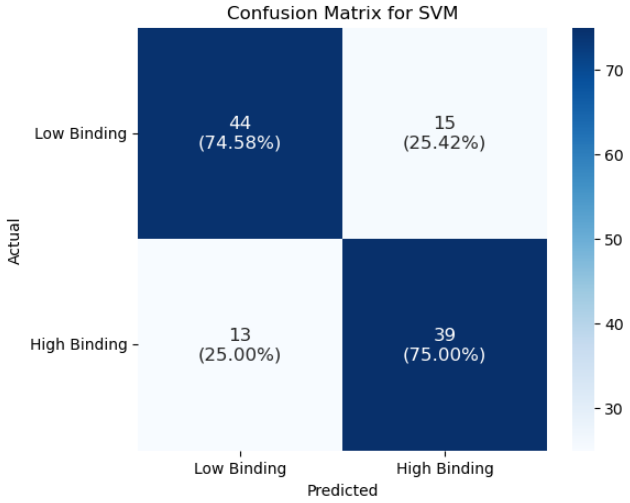

Figure 4.10: Confusion matrix for SVM

The best hyperparameters for the SVM model, as identified through GridSearchCV, are as follows:

- **C:** 1: This regularization parameter controls the trade-off between achieving a low training error and a low testing error. A value of 1 provides a good balance, preventing overfitting while still capturing patterns in the data.

- **gamma:** 'scale': This parameter defines how far the influence of a single training example reaches. Using 'scale' adjusts gamma based on the number of features, ensuring proper regularization and avoiding overfitting.

- **kernel:** 'rbf' (Radial Basis Function): This kernel is commonly used for non-linear data. It helps in mapping the input features into higher dimensions to find a linear separation in complex data.

The SVM model showed adequate classification performance with an accuracy of 74.8%. While it performed well in some cases, particularly in identifying low binding affinities, it misclassified a significant number of instances for both classes. Its hyperparameters provide a balance between complexity and generalization but might benefit from further tuning or additional features to improve performance.

4.4.2.4 Neural Network

The Deep Neural Network (DNN) model, as depicted in the confusion matrix, displayed moderate performance in distinguishing between low and high binding affinities. Below is a breakdown of its classification results:

- **Low Binding (0):** The model correctly classified 42 out of 59 instances, achieving an accuracy of 71.19% for this class. However, 17 instances were misclassified as high binding, indicating the model's struggle to perfectly capture the patterns for low binding affinities.

- **High Binding (1):** Out of 52 actual high binding instances, the model correctly predicted 39, resulting in a true positive rate of 75%. Nevertheless, 13 instances were incorrectly classified as low binding.

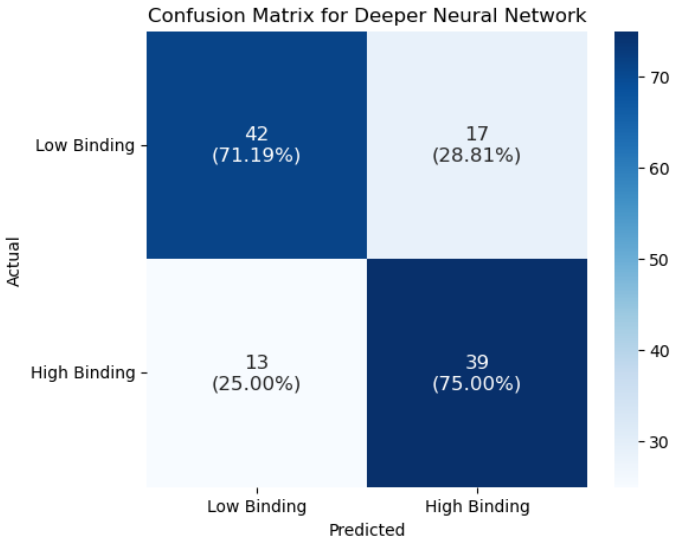

Figure 4.11: Confusion matrix for neural network

The architecture of this deeper neural network consists of three hidden layers with 256, 128, and 64 neurons respectively. It employs LeakyReLU activations and dropout to prevent overfitting, while a Sigmoid function in the final layer ensures that predictions are between 0 and 1, suitable for binary classification. The model is optimized using Adam optimizer and trained using Binary Cross-Entropy Loss.

The network was trained with the following tunable hyperparameters:

- **Learning Rate (lr) = 0.0001:** This controls how much to adjust the model with respect to the loss each time the model weights are updated. A lower learning rate was selected to ensure the model converges gradually without skipping over optimal weights.
- **Batch Size = 32:** Mini-batch gradient descent was used with batches of size 32 to balance computational efficiency and model convergence.
- **Epochs = 200:** The model was trained for 200 epochs, allowing it to learn from the data iteratively.

The Deeper Neural Network model showed moderate classification performance with an accuracy of 73.0%. While its recall for high binding affinity was satisfactory (75%), it

struggled to distinguish low binding instances, misclassifying 17 out of 59. Although the model captures non-linear relationships effectively, further tuning or additional training data may improve its performance. The AUC of 0.814 suggests a decent balance between precision and recall, but there's room for optimization.

4.4.3 Result review

Among the four models evaluated, XGBoost emerged as the top performer. Its ensemble of decision trees, gradient boosting mechanism, and effective regularization allowed it to capture complex patterns while preventing overfitting. The model's strong results across precision, recall, and AUC make it the most suitable option for distinguishing between low and high binding affinities. The hyperparameter tuning, particularly the control over learning rate, tree depth, and subsampling, ensured that XGBoost balanced complexity with performance, making it an excellent choice for this task.

RandomForest, while slightly behind XGBoost, still demonstrated robust performance without signs of overfitting. Its reliance on random sampling and averaging over multiple trees leads to a strong generalization ability. Although its metrics did not surpass XGBoost, the model's more stable behavior suggests that it may perform better on unseen data due to its nature of reducing variance. In contrast, SVM and the Deeper Neural Network struggled with tuning challenges and did not achieve the same level of generalization as the tree-based models. SVM's sensitivity to the correct kernel choice and the neural network's requirement for extensive training likely limited their potential in this context. Overall, RandomForest, with its simplicity and avoidance of overfitting, stands as a reliable contender for future applications.

4.5 Prediction Model

The primary goal of this phase is to predict the binding affinity of molecules based on the selected features from the dataset. Unlike classification tasks, where the outcome is categorical, this is a regression task where the target variable is continuous. The binding affinity, which reflects the strength of interaction between molecules, is a critical parameter in many scientific and pharmaceutical studies.

Through comprehensive training and tuning processes, these models are expected to generalize well and offer reliable predictions for future unseen data, allowing for robust performance in practical applications.

4.5.1 Model Selection

In the prediction task for binding affinity, we employed four machine learning models: XGBoost, Random Forest, Support Vector Machine (SVM), and a Neural Network. The objective of this approach was to predict binding affinity values based on selected features from the dataset. Each model was trained, evaluated, and fine-tuned using a systematic grid search approach, applying cross-validation to optimize the hyperparameters for better predictive performance.

The process began with the preparation of the dataset, where features were selected based on prior analysis. Following this, models were trained using a combination of grid search for hyperparameter tuning and k-fold cross-validation to ensure the models would generalize well to unseen data. Each model's performance was evaluated using Mean Squared Error (MSE) and the R^2 Score, two critical metrics for regression tasks. Additionally, scatter plots of actual vs predicted binding affinities were generated to visualize how well the models performed, and learning curves were plotted to assess how the models improved as more data was used during training. This comprehensive approach ensured that each model's strengths were fully leveraged, while also providing a detailed understanding of where each model performed best in the prediction of binding affinities.

4.5.2 Model Results and Evaluation

The results of the four machine learning models—XGBoost, Random Forest, SVM, and Neural Network—are evaluated based on two primary regression metrics: Mean Squared Error (MSE) and R^2 Score. These metrics help determine how well the models performed in predicting the binding affinity. Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values. A lower MSE indicates a better fit. R^2 Score represents the proportion of variance in the target variable

that is predictable from the features. A higher R^2 score, closer to 1, indicates a better fit of the model to the data.

The table below summarizes the model performance:

Table 4.3: Prediction evaluation report

Model	MSE	R^2 Score
XGBoost	0.527	0.704
RandomForest	0.319	0.814
SVM	0.877	0.640
Neural Network	1.767	0.526

The RandomForest model achieved the lowest MSE and the highest R^2 Score, indicating that it outperformed the other models in terms of predictive accuracy. The XGBoost model also demonstrated strong performance, while the SVM and Neural Network models showed lower predictive capabilities, as reflected in their higher MSE and lower R^2 Scores.

4.5.2.1 Extreme Gradient Boosting

The XGBoost model was applied to predict binding affinity, achieving an MSE of 0.527 and an R^2 score of 0.704. The relatively low MSE indicates that the model's predicted binding affinity values are close to the actual values, while the R^2 score of 0.704 shows that 70.4% of the variance in the binding affinity can be explained by the model. These metrics reflect the model’s strong performance in capturing the underlying patterns in the data.

The best hyperparameters for the **XGBoost** model are as follows:

- **colsample_bytree (0.8)**: This parameter controls the fraction of features used for each tree, promoting diversity and reducing overfitting.
- **gamma (0.3)**: Gamma adds regularization by controlling the minimum loss reduction required to make a split, helping to simplify the model and avoid overfitting.
- **learning_rate (0.1)**: The learning rate determines the step size for each update, with a moderate value allowing for slower, more stable convergence.
- **max_depth (8)**: This limits the depth of each tree, preventing excessive complexity while capturing relevant patterns.
- **n_estimators (200)**: The number of trees in the model, providing sufficient complexity to improve predictive accuracy without overfitting.

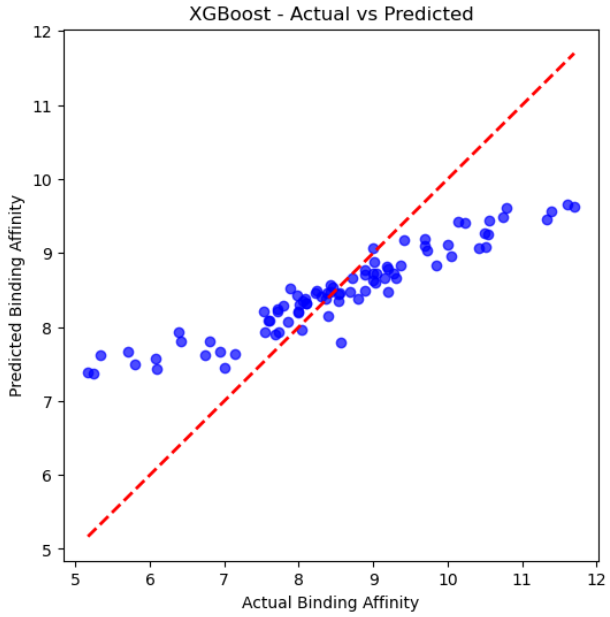

Figure 4.12: XGBoost model - Actual vs predicted binding affinity scatter plot

The scatter plot of Actual vs Predicted Binding Affinities (Figure X) visually illustrates the model's predictions. Most data points are concentrated along the red diagonal line,

indicating that the predicted values are well-aligned with the actual values. This further confirms the model's reliability in predicting binding affinity across different samples.

4.5.2.2 Random Forest

The Random Forest model was applied to predict binding affinity, yielding a Mean Squared Error (MSE) of 0.319 and an R^2 score of 0.814. The lower MSE indicates that the predicted binding affinity values closely match the actual values, while the R^2 score of 0.814 demonstrates that 81.4% of the variance in binding affinity is explained by the model. This highlights the model's strong predictive capability.

Here is a breakdown of the Random Forest hyperparameters:

- **max_depth: 10**: This limits the depth of each tree, preventing them from becoming overly complex and reducing the risk of overfitting.
- **n_estimators: 200**: This defines the number of trees in the forest. A higher number of trees improves the model's stability and accuracy by averaging over multiple predictions.
- **min_samples_leaf: 2**: This sets the minimum number of samples required to be at a leaf node, controlling how much a tree can be split.
- **min_samples_split: 5**: This specifies the minimum number of samples required to split an internal node, controlling the growth of the tree and preventing overfitting.

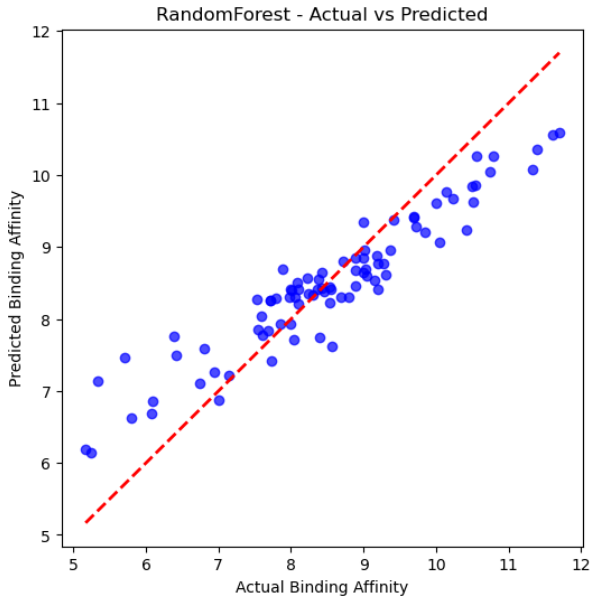

Figure 4.13: Random Forest model - Actual vs predicted binding affinity scatter plot

The scatter plot of Actual vs Predicted Binding Affinities (Figure X) provides a clear visualization of the model's performance. A large concentration of data points along the red diagonal line indicates that the predicted values closely follow the actual values. The alignment of these points reinforces the model's reliability in predicting binding affinity across different samples.

4.5.2.3 Support Vector Machine

The Support Vector Machine (SVM) model was applied to predict binding affinity, achieving an MSE of 0.877 and an R^2 score of 0.640. While the performance is not as strong as other models like Random Forest or XGBoost, the SVM still captures a substantial portion of the variance in the binding affinity. The R^2 score indicates that 64% of the variance in binding affinity can be explained by the model, although the MSE suggests that the predictions are less accurate than the leading models.

The best hyperparameters for the SVM model included a **C value of 1**, which controls the trade-off between achieving a low error on the training data and minimizing model complexity, ensuring the model does not overfit. The **gamma parameter set to 'scale'** adjusts the kernel width to automatically account for the feature variance. The **linear kernel** was chosen, which assumes a linear relationship between features and the target, and proved sufficient for the data structure.

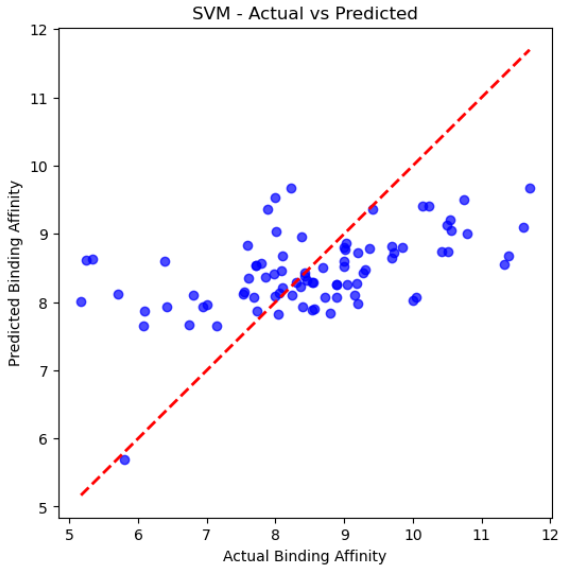

Figure 4.14: SVM model - Actual vs predicted binding affinity scatter plot

The scatter plot (Figure X) of actual vs. predicted binding affinities shows a wider spread of points around the diagonal red line compared to the other models, indicating slightly higher variability in predictions. Despite this, the SVM model demonstrates reasonable predictive performance with some room for improvement in future iterations.

4.5.2.4 Neural Network

For the neural network model, we implemented a a **Deep Neural Network (DNN)** using PyTorch, consisting of four fully connected layers. The model starts with an input layer that takes in 10 selected features from the dataset, passing them through the first fully connected layer with 64 neurons. This is followed by the first hidden layer, which contains 32 neurons and uses the ReLU activation function to

introduce non-linearity. The output from this hidden layer is passed to the second hidden layer, containing 16 neurons, where ReLU is again applied. Finally, the output layer consists of a single neuron responsible for predicting the binding affinity. Since this is a regression task, no activation function is applied in the output layer.

The neural network was trained using the **Mean Squared Error (MSE)** as the loss function and optimized using the **Adam optimizer** with a learning rate of 0.001. The training process was conducted over 100 epochs with a batch size of 32, ensuring that the model could iteratively update its weights to minimize the error between the predicted and actual binding affinity values.

After training, the neural network achieved an MSE of **1.767** and an R^2 score of **0.526**, reflecting moderate performance. While the model captured some patterns in the data, the relatively higher MSE and lower R^2 compared to other models suggest that the neural network struggled to generalize as effectively. In terms of hyperparameters, we trained the model for 200 epochs, allowing for sufficient iterations to learn from the data. A smaller learning rate of 0.0005 was chosen to ensure gradual convergence, while a batch size of 32 was used to optimize learning efficiency during training.

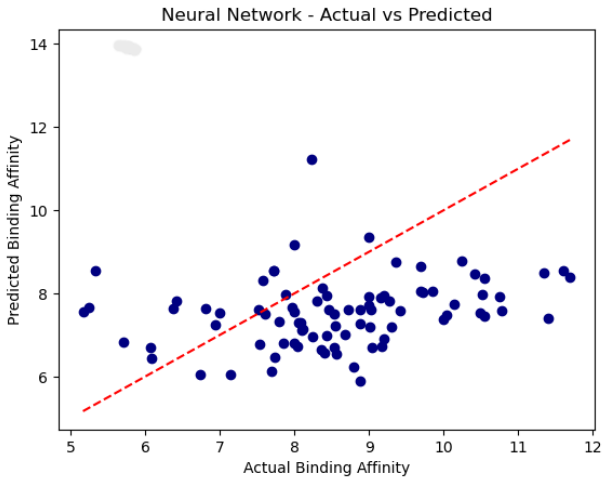

Figure 4.15: Neural network model - Actual vs predicted binding affinity scatter plot

The scatter plot of Actual vs. Predicted Binding Affinity (Figure X) demonstrates that the neural network predictions are more spread out compared to the XGBoost and Random Forest models. While some predicted values align with the actual values (close to the diagonal line), there is a noticeable dispersion, indicating that the neural network struggled to consistently capture the relationship between features and binding affinity.

4.5.3 Result review

In the evaluation of machine learning models for predicting binding affinity, the performance of the four models—XGBoost, Random Forest, SVM, and Neural Network—showed varying degrees of success. Among the models, **Random Forest** demonstrated the best overall performance with a low MSE of **0.319** and an R^2 score of **0.814**. This indicates that 81.4% of the variance in binding affinity was successfully captured by the model. The Random Forest model's success can be attributed to its ensemble learning approach, which builds multiple decision trees and averages the results. This method helps reduce overfitting and enhances the model's generalization ability. Furthermore, the hyperparameters—such as limiting the tree depth to 10 and using 200 estimators—ensured that the model remained flexible yet not overly complex. The balance between complexity and simplicity allowed Random Forest to consistently deliver accurate predictions across different samples, making it the best-performing model in this task.

On the other hand, the **Neural Network** showed the least promising results, with an MSE of **1.767** and an R^2 score of **0.526**, indicating that it struggled to capture the underlying patterns in the data. The challenge with the neural network model lies in the need for a larger dataset or more complex architectures to fully leverage its potential. Deep learning models typically excel when trained on vast amounts of data, and in this case, the dataset may not have been extensive enough for the network to generalize well. Additionally, while the neural network could have benefited from more fine-tuned hyperparameters or regularization techniques, the spread of predictions in the scatter plot indicates its difficulty in capturing the relationship between the features and the target variable.

While **XGBoost** and **SVM** also performed well, their results were not as strong as Random Forest. XGBoost achieved an MSE of **0.527** and an R^2 score of **0.704**, and its strong performance can be credited to its use of boosting techniques and effective handling of overfitting. **SVM**, with an MSE of **0.877** and an R^2 score of **0.640**, showed reasonable predictive capabilities but exhibited more variability in its predictions due to its linear kernel and sensitivity to noise in the data. However, both models showcased their strengths in prediction accuracy and model complexity, with XGBoost emerging as a competitive choice for predicting binding affinity.

In conclusion, Random Forest and XGBoost proved to be the most reliable models for predicting binding affinity, while SVM and the Neural Network faced challenges in terms of generalization and predictive accuracy.

CHAPTER 5: CONCLUSIONS AND FUTURE RECOMMENDATION

Antibody-antigen binding affinity plays a critical role in therapeutic antibody development, with strong and specific binding being essential for the effectiveness of treatments. Traditional methods for measuring binding affinities are both costly and time-consuming, necessitating the development of more efficient computational approaches. This study successfully integrates machine learning and deep learning techniques to classify and predict antibody-antigen binding affinities. Using a dataset of antibody-antigen complexes, we extracted key structural features and implemented models such as Random Forest, XGBoost, Support Vector Machines (SVM), and Neural Networks to

predict binding affinities. Among the models, Random Forest emerged as the most effective, providing high predictive accuracy for both classification and regression tasks. This study demonstrates the potential of combining structural data with advanced machine learning techniques to enhance our ability to predict binding affinities, offering a valuable tool for the future of therapeutic antibody design.

However, this research has some limitations. The relatively small size of the dataset may introduce biases and limit the generalizability of the models. Additionally, while the machine learning models showed strong predictive power, further validation with larger and more diverse datasets is essential to confirm the robustness and broader applicability of the models. Expanding the dataset and improving the feature selection process could further optimize the accuracy and reliability of the predictions.

Future work should aim to develop a comprehensive pipeline that takes antibody-antigen data as input and provides a full range of analysis plots and insights. This pipeline would automate tasks such as preprocessing antibody-antigen structural data, optimizing structures, and integrating machine learning models to classify and predict binding affinities. It would generate outputs like binding energy trends, RMSD plots, and predicted affinity scores, enabling researchers to easily interpret results and make informed decisions for therapeutic design. The flexible architecture would allow future enhancements, making it a valuable tool for antibody research and biopharmaceutical applications.

REFERENCES

[1] A. A. J. Vaillant, Z. Jamal, P. Patel, and K. Ramphul, "Immunoglobulin," *Encyclopedia of Respiratory Medicine: Volume 1-4*, vol. 1–4, pp. V2-314-V2-320, Aug. 2023, doi: 10.1016/B0-12-370879-6/00184-8.

[2] P. Sharma, R. V. Joshi, R. Pritchard, K. Xu, and M. A. Eicher, "Therapeutic Antibodies in Medicine," *Molecules*, vol. 28, no. 18, p. 6438, Sep. 2023, doi: 10.3390/MOLECULES28186438.

[3] R. M. Lu *et al.*, “Development of therapeutic antibodies for the treatment of diseases,” *Journal of Biomedical Science 2020 27:1*, vol. 27, no. 1, pp. 1–30, Jan. 2020, doi: 10.1186/S12929-019-0592-Z.

[4] “Therapeutic antibody potential in 2024 - Drug Discovery World (DDW).” Accessed: Aug. 21, 2024. [Online]. Available: https://www.ddw-online.com/therapeutic-antibody-potential-in-2024-29475-202405/

[5] M. L. Chiu, D. R. Goulet, A. Teplyakov, and G. L. Gilliland, “Antibody Structure and Function: The Basis for Engineering Therapeutics,” *Antibodies*, vol. 8, no. 4, Dec. 2019, doi: 10.3390/ANTIB8040055.

[6] “Janeway’s Immunobiology - Kenneth Murphy, Casey Weaver - Google Books.” Accessed: Aug. 21, 2024. [Online]. Available: https://books.google.com.pk/books?hl=en&lr=&id=GmPLCwAAQBAJ&oi=fnd&pg=PP1&dq=Murphy,+K.,+Weaver,+C.,+%26+Janeway,+C.+(2017).+Janeway%27s+Immunobiology+(9th+ed.).+Garland+Science.&ots=6crf36u1rg&sig=2_9QEsdxDfhLHJRDSoKQYU-2nfk&redir_esc=y#v=onepage&q&f=false

[7] L. A. Rabia, Y. Zhang, S. D. Ludwig, M. C. Julian, and P. M. Tessier, “Net charge of antibody complementarity-determining regions is a key predictor of specificity,” *Protein Engineering, Design and Selection*, vol. 31, no. 11, p. 409, Nov. 2018, doi: 10.1093/PROTEIN/GZZ002.

[8] L. Polonelli *et al.*, “Antibody Complementarity-Determining Regions (CDRs) Can Display Differential Antimicrobial, Antiviral and Antitumor Activities,” *PLoS One*, vol. 3, no. 6, Jun. 2008, doi: 10.1371/JOURNAL.PONE.0002371.

[9] M. L. Chiu, D. R. Goulet, A. Teplyakov, and G. L. Gilliland, “Antibody Structure and Function: The Basis for Engineering Therapeutics,” *Antibodies*, vol. 8, no. 4, Dec. 2019, doi: 10.3390/ANTIB8040055.

[10] W. R. Strohl, “Structure and function of therapeutic antibodies approved by the US FDA in 2023,” *Antib Ther*, vol. 7, no. 2, pp. 132–156, Mar. 2024, doi: 10.1093/ABT/TBAE007.

[11] K. Tsumoto and J. M. Caaveiro, “Antigen–Antibody Binding,” *Encyclopedia of Life Sciences*, pp. 1–8, Dec. 2016, doi: 10.1002/9780470015902.A0001117.PUB3.

[12] M. L. Chiu, D. R. Goulet, A. Teplyakov, and G. L. Gilliland, “Antibody Structure and Function: The Basis for Engineering Therapeutics,” *Antibodies*, vol. 8, no. 4, Dec. 2019, doi: 10.3390/ANTIB8040055.

[13] I. Sela-Culang, V. Kunik, and Y. Ofran, “The structural basis of antibody-antigen recognition,” *Front Immunol*, vol. 4, no. OCT, p. 64858, Oct. 2013, doi: 10.3389/FIMMU.2013.00302/BIBTEX.

[14] J. Charles A Janeway, P. Travers, M. Walport, and M. J. Shlomchik, “The interaction of the antibody molecule with specific antigen,” 2001, Accessed: Aug. 21, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK27160/

[15] J. Bostrom, C. V. Lee, L. Haber, and G. Fuh, “Improving antibody binding affinity and specificity for therapeutic development,” *Methods Mol Biol*, vol. 525, 2009, doi: 10.1007/978-1-59745-554-1_19.

[16] Z. Wang, G. Wang, H. Lu, H. Li, M. Tang, and A. Tong, “Development of therapeutic antibodies for the treatment of diseases,” *Molecular Biomedicine 2022 3:1*, vol. 3, no. 1, pp. 1–31, Nov. 2022, doi: 10.1186/S43556-022-00100-4.

[17] S. Conti, E. Y. Lau, and V. Ovchinnikov, “On the Rapid Calculation of Binding Affinities for Antigen and Antibody Design and Affinity Maturation Simulations,” *Antibodies*, vol. 11, no. 3, Sep. 2022, doi: 10.3390/ANTIB11030051/S1.

[18] J. J. Rhoden, G. L. Dyas, and V. J. Wroblewski, “A modeling and experimental investigation of the effects of antigen density, binding affinity, and antigen expression ratio on bispecific antibody binding to cell surface targets,” *Journal of*

Biological Chemistry, vol. 291, no. 21, pp. 11337–11347, May 2016, doi: 10.1074/jbc.M116.714287.

[19] I. Sela-Culang, V. Kunik, and Y. Ofran, "The structural basis of antibody-antigen recognition," *Front Immunol*, vol. 4, no. OCT, p. 64858, Oct. 2013, doi: 10.3389/FIMMU.2013.00302/BIBTEX.

[20] B. M. Sala, T. Le Marchand, G. Pintacuda, C. Camilloni, A. Natalello, and S. Ricagno, "Conformational Stability and Dynamics in Crystals Recapitulate Protein Behavior in Solution," *Biophys J*, vol. 119, no. 5, pp. 978–988, Sep. 2020, doi: 10.1016/J.BPJ.2020.07.015.

[21] S. Y. Huang and X. Zou, "Advances and Challenges in Protein-Ligand Docking," *Int J Mol Sci*, vol. 11, no. 8, p. 3016, Aug. 2010, doi: 10.3390/IJMS11083016.

[22] A. A. Adeniyi and M. E. S. Soliman, "Implementing QM in docking calculations: is it a waste of," *Drug Discov Today*, vol. 22, no. 8, pp. 1216–1223, Aug. 2017, doi: 10.1016/j.drudis.2017.06.012.

[23] R. Qureshi *et al.*, "AI in drug discovery and its clinical relevance," *Heliyon*, vol. 9, no. 7, Jul. 2023, doi: 10.1016/J.HELIYON.2023.E17575.

[24] A. Blanco-González *et al.*, "The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies," *Pharmaceuticals*, vol. 16, no. 6, p. 891, Jun. 2023, doi: 10.3390/PH16060891/S1.

[25] C. Arnold, "Inside the nascent industry of AI-designed drugs," *Nat Med*, vol. 29, no. 6, pp. 1292–1295, Jun. 2023, doi: 10.1038/S41591-023-02361-0.

[26] J. Kim, M. McFee, Q. Fang, O. Abdin, and P. M. Kim, "Computational and artificial intelligence-based methods for antibody development," *Trends Pharmacol Sci*, vol. 44, no. 3, pp. 175–189, Mar. 2023, doi: 10.1016/J.TIPS.2022.12.005.

[27] M. L. Fernández-Quintero *et al.*, "Challenges in antibody structure prediction," *MAbs*, vol. 15, no. 1, 2023, doi: 10.1080/19420862.2023.2175319.

[28] B. Abanades, W. K. Wong, F. Boyles, G. Georges, A. Bujotzek, and C. M. Deane, "ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins," *Communications Biology 2023 6:1*, vol. 6, no. 1, pp. 1–8, May 2023, doi: 10.1038/s42003-023-04927-7.

[29] J. Charles A Janeway, P. Travers, M. Walport, and M. J. Shlomchik, "Immunobiology," *Immunobiology*, no. 14102, pp. 1–10, 2001, Accessed: Aug. 23, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK10757/

[30] M. Kumar, A. Jalota, S. K. Sahu, and S. Haque, "Therapeutic antibodies for the prevention and treatment of cancer," *J Biomed Sci*, vol. 31, no. 1, pp. 1–12, Dec. 2024, doi: 10.1186/S12929-024-00996-W/FIGURES/5.

[31] J. H. Lee, R. Yin, G. Ofek, and B. G. Pierce, "Structural Features of Antibody-Peptide Recognition," *Front Immunol*, vol. 13, p. 910367, Jul. 2022, doi: 10.3389/FIMMU.2022.910367/BIBTEX.

[32] E. J. Sundberg, "Structural basis of antibody-antigen interactions," *Methods Mol Biol*, vol. 524, pp. 23–36, 2009, doi: 10.1007/978-1-59745-450-6_2.

[33] P. B. P. S. Reis, G. P. Barletta, L. Gagliardi, S. Fortuna, M. A. Soler, and W. Rocchia, "Antibody-Antigen Binding Interface Analysis in the Big Data Era," *Front Mol Biosci*, vol. 9, p. 945808, Jul. 2022, doi: 10.3389/FMOLB.2022.945808/BIBTEX.

[34] A. V. Madsen *et al.*, "Structural trends in antibody-antigen binding interfaces: a computational analysis of 1833 experimentally determined 3D structures," *Comput Struct Biotechnol J*, vol. 23, pp. 199–211, Dec. 2024, doi: 10.1016/J.CSBJ.2023.11.056.

[35] C. Schneider, M. I. J. Raybould, and C. M. Deane, "SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker," *Nucleic Acids Res*, vol. 50, no. D1, pp. D1368–D1372, Jan. 2022, doi: 10.1093/NAR/GKAB1050.

[36] M. I. J. Raybould *et al.*, “Thera-SAbDab: the Therapeutic Structural Antibody Database,” *Nucleic Acids Res*, vol. 48, no. D1, pp. D383–D388, Jan. 2020, doi: 10.1093/NAR/GKZ827.

[37] C. Ye, W. Hu, and B. Gaeta, “Prediction of Antibody-Antigen Binding via Machine Learning: Development of Data Sets and Evaluation of Methods,” *JMIR Bioinform Biotech 2022;3(1):e29404 https://bioinform.jmir.org/2022/1/e29404*, vol. 3, no. 1, p. e29404, Oct. 2022, doi: 10.2196/29404.

[38] K. Tharakaraman *et al.*, “Redesign of a cross-reactive antibody to dengue virus with broad-spectrum activity and increased in vivo potency,” *Proc Natl Acad Sci U S A*, vol. 110, no. 17, pp. E1555–E1564, Apr. 2013, doi: 10.1073/PNAS.1303645110/SUPPL_FILE/SAPP.PDF.

[39] D. S. Quinlan, R. Raman, K. Tharakaraman, V. Subramanian, G. Del Hierro, and R. Sasisekharan, “An inter-residue network model to identify mutational-constrained regions on the Ebola coat glycoprotein,” *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–8, Apr. 2017, doi: 10.1038/srep45886.

[40] L. N. Robinson *et al.*, “Structure-Guided Design of an Anti-dengue Antibody Directed to a Non-immunodominant Epitope,” *Cell*, vol. 162, no. 3, pp. 493–504, Nov. 2015, doi: 10.1016/j.cell.2015.06.057.

[41] Y. H. Wong *et al.*, “Molecular basis for dengue virus broad cross-neutralization by humanized monoclonal antibody 513,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–17, May 2018, doi: 10.1038/s41598-018-26800-y.

[42] N. L. Miller, R. Raman, T. Clark, and R. Sasisekharan, “Complexity of Viral Epitope Surfaces as Evasive Targets for Vaccines and Therapeutic Antibodies,” *Front Immunol*, vol. 13, p. 904609, Jun. 2022, doi: 10.3389/FIMMU.2022.904609/BIBTEX.

[43] C. Ye, W. Hu, and B. Gaeta, “Prediction of Antibody-Antigen Binding via Machine Learning: Development of Data Sets and Evaluation of Methods,” *JMIR Bioinform*

Biotech 2022;3(1):e29404 https://bioinform.jmir.org/2022/1/e29404, vol. 3, no. 1, p. e29404, Oct. 2022, doi: 10.2196/29404.

[44] G. Zhang, Z. Su, T. Zhang, and Y. Wu, "Machine-learning-based Structural Analysis of Interactions between Antibodies and Antigens," *bioRxiv*, Dec. 2023, doi: 10.1101/2023.12.06.570397.

[45] Y. Huang, Z. Zhang, and Y. Zhou, "AbAgIntPre: A deep learning method for predicting antibody-antigen interactions based on sequence information," *Front Immunol*, vol. 13, Dec. 2022, doi: 10.3389/FIMMU.2022.1053617.

[46] G. Zhang, Z. Su, T. Zhang, and Y. Wu, "Machine-learning-based Structural Analysis of Interactions between Antibodies and Antigens," *bioRxiv*, Dec. 2023, doi: 10.1101/2023.12.06.570397.

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

[48] M. Baek *et al.*, "Accurate prediction of protein structures and interactions using a 3-track neural network," *Science*, vol. 373, no. 6557, p. 871, Aug. 2021, doi: 10.1126/SCIENCE.ABJ8754.

[49] J. Jänes and P. Beltrao, "Deep learning for protein structure prediction and design—progress and applications," *Mol Syst Biol*, vol. 20, no. 3, p. 162, Mar. 2024, doi: 10.1038/S44320-024-00016-X.

[50] Y. Liu *et al.*, "Interpretable antibody-antigen interaction prediction by introducing route and priors guidance," *bioRxiv*, p. 2024.03.09.584264, Apr. 2024, doi: 10.1101/2024.03.09.584264.

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

[52] N. L. Miller, T. Clark, R. Raman, and R. Sasisekharan, “Learned features of antibody-antigen binding affinity,” *Front Mol Biosci*, vol. 10, p. 1112738, Feb. 2023, doi: 10.3389/FMOLB.2023.1112738/BIBTEX.