Generative AI-driven Antibody Design and Optimization: A Transformers Paradigm

by

Amman Safeer

(Registration No: 00000361289)

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

in

Bioinformatics

Supervisor: Dr. Salma Sherbaz

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

2025

03/2/24

Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr./Ms. Amman Safeer Registration No. 361289 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: Salma

Assistant Professor
SINES - NUST, Sector H-12
Islamabad

Name of Supervisor: Dr. Salma Sherbaz

Date: 03/02/2025

Signature of HoD with stamp:

Date: 03/02/2025

DR. MUHAMMAD TARIQ SAEED
Associate Professor
School of Interdisciplinary
Engineering & Sciences (SINES)
NUST, Sector H-12 Islamabad.

Countersign by

Signature (Dean/Principal):

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

Date: 04 FEB 2025

AUTHOR'S DECLARATION

I Amman Safeer hereby state that my MS thesis titled "Generative AI-driven Antibody Design and Optimization: A Transformers Paradigm" 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: Amman Safeer

Date: January 31th, 2025

DEDICATION

This research is dedicated to the greatest teacher for all mankind, Prophet Muhammad (PBUH). Secondly, I dedicate this research to my parents and siblings , who supported me in my ups and downs. And lastly, I dedicate this research to the Father of Algorithms, Muhammad Ibn Musa Al-Khwarizmi

ACKNOWLEDGMENTS

First, I am deeply grateful to **Allah Almighty**, whose countless blessings and guidance have brought me to this point in my educational journey. None of this would have been possible without His grace and mercy. I would like to express my deepest gratitude to my **Parents**, whose unwavering support and endless sacrifices have been my pillars of strength. Their encouragement and belief in me have always been the driving force behind my achievements.

I want to express my sincere gratitude to my former supervisor **Dr. Mehak Rafiq** and my current supervisor **Dr. Salma Sherbaz**. Your guidance, patience, and invaluable guidance have been of great help in completing this thesis. Your encouragement and belief in my abilities have helped me through the challenges of this research. I am deeply thankful to my GEC members, **Dr. M. Tariq Saeed** and **Dr. Zartasha Mustansar**, for their insightful guidance and constructive feedback. Your expertise and advice have helped shape this thesis and guided me through the research process. I am extremely grateful to my uncle **Muhammad Shakeel Anjum** for his everlasting support, encouragement and prayers.

I would like to extend my gratitude to my friends **Muneeb Nasir, Ghulam Murtaza** and **Saifullah Butt**, for their continuous support throughout this journey. I am also thankful to my colleagues from the Data Analytics Lab, particularly **Adeel Nasir**, **Ayesha Iman**, **Talha Zuberi**, **Ariba Abbasi** and **Inam Rafique** for their valuable collaboration and encouragement during the course of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ML	Machine Learning
MLM	Masked Language Modeling
ILM	Infilling Language Modeling
ALM	Autoregressive Language Modeling
GPT	Generative Pretrained Transformer
GPT-S	Generative Pretrained Transformer-Small
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robust BERT
ΔG	Gibbs Free Energy
K_d	Dissociation Constant
OAS	Observed Antibody Space
VoroMQA	Voronoi tessellation-based Model Quality Assessment
IgLM	Immunoglobulin Language Model

ABSTRACT

The human immune system generates high-affinity antibodies against pathogens and diseases. These antibodies serve as key therapeutic and diagnostic tools for disease classification and treatment. Traditional approaches like display technologies can generate potential antibody leads, but they come with challenges such as expressibility, viscosity, immunogenicity, and pharmacokinetics.. The recent advancements in AI have led to the foundation of the generative AI, which has also impacted the field of bioinformatics. The field of generative AI like transformer-based NLP models have significantly improved the development of better computational tools in protein and antibody design. These models can leverage protein and antibody sequence information to reduce the need for resource-intensive display technology experiments. However, existing models are often trained on multi-species datasets, which can introduce species-specific biases and limit their ability to generate diverse human antibodies. Here, this study proposed; AbSynth, a class of transformer-based antibody language models exclusively trained on 1 million human antibodies sequences dataset to improve generalizability in human antibody design. AbSynth models were tested on the natural antibody 1E6J to improve binding affinity. Of the 400 generated antibodies, 10 showed significant improvement in binding affinity.. Furthermore, AbSynth-generated sequences exhibited 96% humanness in heavy chain sequences under low sampling parameters, demonstrating its potential as an effective tool for designing and isolating diverse, humanized antibody candidates.

Keywords: generative AI, antibody, protein design, binding affinity, transformers.

CHAPTER 1: INTRODUCTION

The first chapter introduces the human immune system, exploring its various facets including innate and adaptive immunity. It highlights the development of antibodies through diverse methodologies while also shedding light on recent breakthroughs in Deep Learning and the emergence of Generative Artificial Intelligence, particularly in the realm of Protein Language Modeling.

1.1 Human Immune System

The human immune system is a complex network of cells, tissues, and molecules that plays a vital role in protecting the body against pathogens, such as viruses, bacteria, fungi, and parasites [1]. It consists of various components and functions to detect, identify, and neutralize invading pathogens while distinguishing them from the body's cells and tissues. The principal role of the immune system is to protect the host against invasion by infectious agents and other foreign substances. The immune system achieves this by recognizing and responding to a wide range of antigens, including pathogens and abnormal cells. The immune system is composed of two main defenses: innate immunity and adaptive immunity. Innate immunity is the first line of defense against pathogens and is antigen independent [2]. It is an immediate defense mechanism that is activated immediately or within hours of encountering an antigen. Innate immunity is characterized by its rapid response and lack of immunologic memory, meaning it does not "memorize" specific pathogens for future encounters. These innate immune responses include physical barriers like the skin and mucous membranes, as well as immune cells such as neutrophils, macrophages, and monocytes. Furthermore, soluble factors like cytokines and complement proteins also contribute to the innate immune response [3]. In contrast, adaptive immunity is a more specialized and specific immune response that develops over time. This type of immunity is antigen-dependent and has memory, meaning it can recognize and remember specific pathogens for future encounters. Adaptive immunity involves the activation of lymphocytes, specifically B cells and T cells, which produce antibodies and mediate cellular immune responses, respectively [2]. Innate and adaptive immune responses work together to provide a comprehensive defense against pathogens [1]. A healthy immune system can distinguish between the body's self-components and those of foreign origin [3]. It can effectively detect and neutralize invading pathogens, while also distinguishing them from the body's cells and tissues.

1.1.1 Innate Immunity

The innate immune system serves as the first line of defence against pathogens and is characterized by its rapid response and lack of immunological memory [4]. Upon encountering an antigen, innate immunity is activated to provide an immediate defence mechanism. This response includes physical barriers such as the skin and mucous membranes, which prevent the entry of pathogens into the body. In addition to physical barriers, innate immunity also involves various immune cells, including neutrophils, macrophages, and monocytes [5]. These cells play crucial roles in recognizing and eliminating pathogens. They can recognize conserved patterns on the surface of pathogens, known as pathogen-associated molecular patterns, through pattern recognition receptors. Through the activation of pattern recognition receptors, innate immune cells can quickly initiate inflammatory responses and release cytokines to recruit other immune cells to the site of infection.

1.1.2 Adaptive Immunity

The adaptive immune response is a more specialized and specific immune response that develops over time [4]. It is antigen-dependent and characterized by the activation of lymphocytes, including B cells and T cells. B cells produce antibodies, which are proteins that can specifically recognize and bind to antigens [6]. These antibodies can neutralize pathogens, mark them for destruction by other immune cells, or activate the complement system to eliminate the pathogen [4]. On the other hand, T cells are involved in cellular immune responses. They can recognize antigens presented by infected cells and directly kill the infected cells or release cytokines to recruit other immune cells to the site of infection. Overall, the adaptive immune response provides a highly targeted and specific defense against pathogens

1.2 Antibody

Antibody plural: Antibodies, also known as immunoglobulins, are a crucial component of the immune system. They are Y-shaped proteins produced by B lymphocytes in response to the presence of foreign substances, known as antigens, in the body [7]. Antibodies play a vital role in humoral immunity, which is one of the two types of specific immune responses mediated by the adaptive immune system. The cardinal features of antibodies and the immune system include specificity, diversity, memory, self-limitation, and discrimination of self from non-self. Antibodies exhibit specificity, which refers to their ability to recognize and bind to specific regions on antigens known as epitopes. Antibodies exhibit the ability to bind directly to these epitopic proteins and ultimately neutralize the pathogens.

1.2.1 Structural Features and Properties

The structure of antibodies is composed of two identical heavy chains and two identical light chains, interconnected by disulfide bonds. These chains are made up of constant regions and variable regions [8]. The constant regions of antibodies determine their class or isotype, such as IgG, IgM, IgA, IgD, and IgE. The variable regions of antibodies are responsible for their antigen-binding specificity and diversity. The constant region is termed as ‘Fc region’ and variable region is termed as ‘Fab region’ respectively. The constant region is responsible for mediating the biological activity or the technique involved to combat the binding antigen based on the Antibody’s isotype while the Fab region is responsible for the antibody’s specificity to bind the antigen (epitope) [9].

The variable regions are the crucial regions in the antibody’s characteristic specificity in antigen binding. The Fab regions comprise of the two identical heavy (H) and two identical light chains (L) that contain two variable regions (VH and VL) and small portions of constant regions at both heavy (CH) and light chains (CL). The variable regions are termed as ‘Fv’ regions. A single Fv is termed as single chain variable fragment (scFv), the scFv contains all the antigen-binding domains of the antibody. Each scFv region is constituted by four framework regions (FR) and three complementary determining regions (CDR). The FRs and CDRs have both been reported in the heavy and light chains Figure 1.1. These have a specific occurring pattern that is same for both heavy and light chain, each scFv chain (light and heavy) begins with FR1 followed by CDR1 and so on [10].

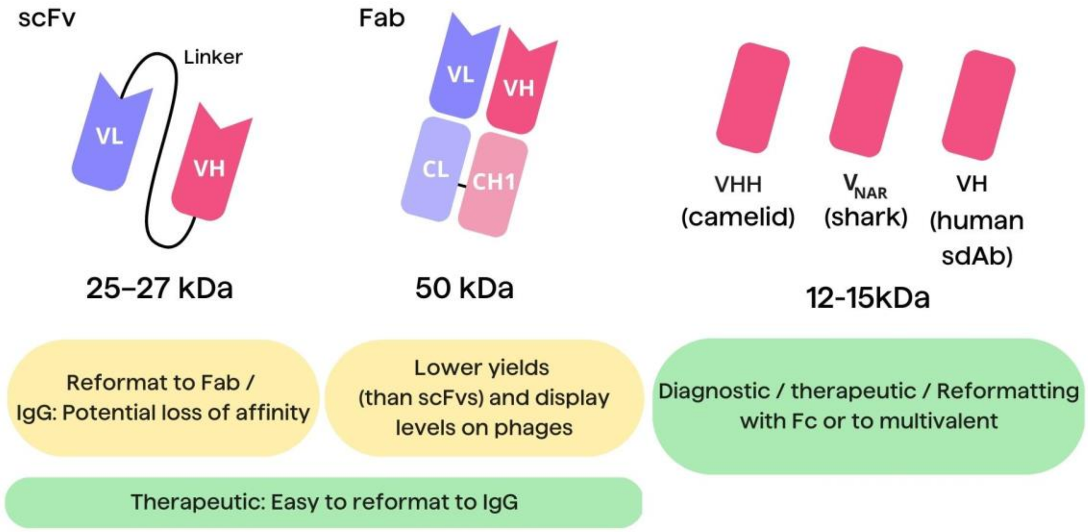


Figure 1.1: Antibody’s Physical Features.

The FRs and CDRs, both play important role in the binding of antigen to the antibody. The antigen-binding activity is shown maximum by the CDRs mostly observed at the CDRH3 position, the parts or regions of the antibody that contribute to the epitope (antigen) binding are termed as the paratopes. However, the FRs have been mostly shown to structurally support the binding activity with the antigen. The overall antigen binding and each region’s contribution can be ranked in such order [11]:

$$CDR3H;\ 29\% > CDR2H;\ 23\% > CDR3L;\ 21\% > CDR1H;\ 10\% > CDR1L;\ 9\% \\ > CDR2L; 4\%, and\ FR\ residues; 4\%$$

1.2.2 Applications

Antibodies higher specificity, strong binding to the target and effective neutralizing of the pathogens makes them suitable for treatments against infections, cancers, and other disorders. Furthermore, antibodies also bind to specific markers found on cells or organs making them for appropriate tool for diagnostics and laboratory experiments. Antibodies combat antigen in various ways that include neutralizing the microbes and toxins, microbes' phagocytosis and opsonization, antibody dependant cellular cytotoxicity (ADCC), or in other cases complement activation. Thus, understanding the structure-function relationships of antibodies is indispensable to develop a platform for protein engineering to generate functionally therapeutic antibodies [12]. Laboratory-made antibodies have been a better solution for therapeutic drug development, one such are monoclonal antibodies that are required in huge volumes of identical antibodies specific to a single epitope. Monoclonal antibodies can be utilized for diagnosis or treatment. They can neutralize pathogens or cancer cells directly, prevent the growth or aid the immune system in killing them. Interestingly, monoclonal antibodies have a high specificity for a specific epitopic target and limited cross-reactivity. Other antibodies, such as polyclonal antibodies, show heterogeneity in nature with higher overall affinity against the antigen due to the recognition of multiple epitopic targets and relatively more economical [13].

1.2.3 Antibody Isolation and Discovery Methods

The conventional antibody isolation methods include *in vitro* display technologies that have been successfully employed to isolate antigen-specific antibodies with therapeutic properties. The important feature that enables the selection of potential antibody with the desired properties is that the protein variant of the antibody is coupled with its genetic information and is referred to genotype phenotype coupling. There are several display technology platforms that are used till date, comprising mammalian surface display, ribosomal display, phage display and yeast surface display technology. The phage display being the state-of-the-art and widely used [14]. Furthermore, there is also a major development in the unconventional methods for antibody design and modelling. Antibody design workflows and pipelines have been evolving with time. These pipelines shorten the duration of antibody design and engineering processes due to the *in-silico* nature of their implementation.

1.2.3.1 Display Technologies

The display technologies were first introduced in 1985 and were first use in 1990s since their conception. The first display technology to be implemented was the phage display technology, that allowed the display of the antibodies on the surface of the bacteriophages. After that, several other technologies have been introduced that include yeast display, mammalian display, ribosome display and bacterial display (Table 1.1) [14]. These technologies vary in their media of antibody expression. The most used are phage display and yeast display.

Table 1.1: Antibody Display Technologies with respect to their generation methods used in antibody discovery.

Display Technology	**Description**	**Library Size**
Phage Display	• Antibodies are displayed on the surface of bacteriophages. • Allows rapid selection and optimization of antibodies.	**10^{10} – 10^{12}**
Mammalian Display	• Mammalian Cells are used for the expression of antibodies. • Allows full-length display with correct post translational modifications (PTM) and fold.	**10^{7} – 10^{9}**
Yeast Display	• Antibodies are displayed in yeast cells. • Full-length antibodies with high-throughput discovery.	**10^{9}**
Ribosome Display	• Cell-free display that relies on protein synthesis reaction of ribosomes; no host-cell transformation. • A full-length display is possible.	**10^{12} – 10^{15}**
Bacterial Display	• Antibodies are displayed on the outer or inner membrane of bacteria. • Fast growth but Gram-positive bacteria generate non-human PTMs.	**10^{11}**

1.2.3.2 In silico Design Methods

Antibodies are one of the most important class of proteins in terms of their therapeutic properties against different diseases like viral and cancers. Thus, the need for unconventional types of design and development techniques is crucial. Therefore, the *in silico* design methods have been getting more popularity and even reduce the need for time and complexity in the design and development processes. The design processes are assisted with the use of bioinformatics techniques like antibody numbering (annotation) through framework and CDR delimiting via numbering schemes, CDR modelling and structure prediction and refinement. Moreover, molecular docking can provide insights of the residues interaction that involves epitope-paratope interaction forming antigen-antibody complex which can provide binding affinity of the antibody [15]. Recently, the computational methods are gaining more popularity because of their ease of usage, speed and improvement in reliability [16]. The improvements in dry and wet lab technologies have been parallel which resulted in expansion of data generated through these methods. Recent expansion of data volume has created ideal case in the favour of machine learning and artificial intelligence applications in antibody design and engineering [17].

1.3 Generative Artificial Intelligence (Generative AI)

Recent developments in computing technologies and availability of vast developer communities have positively impacted the progress of AI methods exponentially. Therefore, forming a new field of deep learning and AI called the generative AI. Generative AI enables deep learning models not only to predict insights from data, but it is being employed to generate novel data as well. This has significantly shifted trend towards the use of generative AI in problem-solving [18]. The development of sophisticated deep learning models has impacted on antibody design methods as well. The use of deep learning in computer vision (CV) and natural language processing (NLP) models has increased the scope of generative AI and deep learning to be used on much more complex problems such as image generation and text generation with better accuracy and making them life-like. Therefore, these models have been deployed in proteins/antibody design to achieve novel and life-like proteins with greater pace and less complexity. One such example is the case of AlphaFold, which is an Evoformer-based deep learning model that can predict protein 3D structures fast and precise [19]. Another model IgFold [20], that combines the contextual representation of antibodies using AntiBERTy [21] with the geometric deep learning to predict 3D structures of the antibodies from the H and L chain sequence.

CHAPTER 2: LITERATURE REVIEW

In this chapter, the review of literature with different factors involved in the process of antibody design and development has been provided. In the first section, the antibody design in terms of display technologies is discussed with its strengths and its limitations. The section describes display technologies for their use of antibody hits discovery and affinity maturation along with the complexities and resource demanding factors to achieve mature antibodies and the significance of antibodies in therapeutics. This is followed by the section of unconventional methods like Protein Language Modeling (PLM) that has been utilized in terms of antibody sequence design by using Large Language Models (LLMs) to achieve diverse, human-like and affinity matured antibodies. The overall goal has been to summarize the conclusions and findings of the research that has been undertaken.

2.1 Display Technologies and their limitations

The process of progressing from 'target-to-hit' is a cornerstone in antibody drug discovery, typically identifying several potential lead candidates through methods such as hybridoma screening or phage and yeast display. Despite this, optimizing these leads demands substantial time and expense, forming most of the preclinical development phase. This challenge arises from the need to simultaneously address multiple factors such as expression levels, viscosity, pharmacokinetics, solubility, and immunogenicity [22]. Once a lead is identified, further engineering becomes necessary. High-throughput methods like phage and yeast display enable the screening of extensive mutagenesis libraries ($>1\times10^9$), primarily improving affinity or specificity to the target antigen. However, therapeutic antibodies must ultimately be expressed as full-length IgG in mammalian cells, necessitating subsequent optimization steps in this context [23]. Mammalian cells lack the ability to stably replicate plasmids, resulting in a low-throughput final development stage. Complex procedures for cloning, transfection, and purification limit screening to small libraries (~1×10^3), constraining changes to minor ones like point mutations. As a result, solving one development issue often exacerbates another or reduces antigen binding entirely, complicating multi-parameter optimization [24].

This limited scope frequently yields antibodies with suboptimal biophysical properties for clinical use. Such deficiencies can lead to side effects or even drug failure. For instance, self-administered subcutaneous antibody injections are increasingly favored for patients needing frequent doses. Yet, identifying highly soluble, low-viscosity antibodies that maintain high biological activity remains a significant challenge [25]. A notable example is Pfizer's withdrawal of Bococizumab, an anti-PCSK9 antibody, from clinical trials due to its immunogenicity undermining long-term treatment efficacy. In contrast, Sanofi and Regeneron's approved antibody, alirocumab, targets PCSK9 but exhibits negligible immunogenic effects [26] . Machine learning applied to biological sequence data presents a powerful tool for enhancing protein engineering by predicting genotype–phenotype relationships [27].

This capability arises from models that map complex relationships between sequence and function. However, collecting high-quality training data remains a critical obstacle in developing accurate machine-learning models. Directed-evolution systems address this issue by linking biological sequence data, such as DNA, RNA, or protein, to phenotypic outcomes [28]. Indeed, it has long been proposed that machine-learning models trained on mutagenesis library data could guide protein engineering effectively.

2.2 Natural Language Processing

Natural Language Processing (NLP) is a field of AI which enables machine to comprehend and process human language whether written or spoken, that enables computers to analyse, interpret and generate textual information of a given query to create human-like responses. NLP is a challenging task due to ambiguity, complexity and diversity in human language. Therefore, the initial NLP models like recurrent neural networks (RNNs) struggle with life-like language generation tasks in many cases. The length of the text and synonymous meanings of the words often creates problems for RNNs that are caused by vanishing and exploding gradient during training [29]. Therefore, pressing the need for better models for NLP task can be addressed through a **transformer architecture** shown in Figure 2.1 which was introduced in a paper, “Attention is All You Need” [30]. Transformers are called large language models (LLMs) that have an attention mechanism, which makes them superior in understanding context and process longer length of the textual data and better handling at variable length sequences. LLMs have sophisticated text processing architecture and are classified into three classes:

a. **Encoder**, that encodes the input sequence into contextual representations.
b. **Decoder**, that decodes the encoded representation into an output sequence.
c. **Encoder-Decoder**, both encoder and decoder work coherently and generate the output.

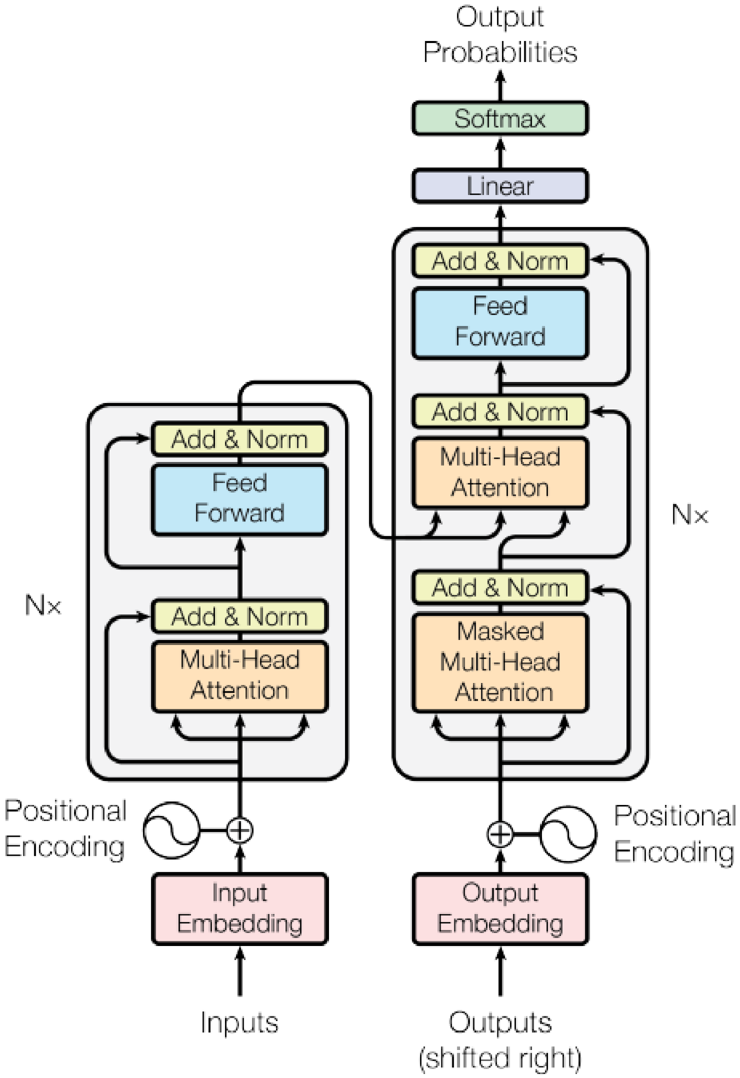


Figure 2.1: Basic Transformer Architecture

The transformers have also been utilized in protein sequence space to address various problems. One such example is evolutionary scale model for protein folding (ESMFold), which is trained on 250 million protein sequences for the purpose of protein 3D structure prediction. Transformers have also been deployed as an effective tool for *de novo* protein design and in many cases as a mutation tool to improve protein functions and properties by treating protein sequence as a text language of amino acids. [31].

2.3 Protein Language Modeling (PLM)

The recent progress in the implementation of natural language processing techniques has been revolving around unsupervised pre-training through utilization of large databases that contain raw protein sequences that particularly are used in situations where there is no availability of structural data. Different studies in this field have delved into a variety of pre-training tasks and the subsequent applications of models. A few noticeable instances include the **ESM** family of models, **AntiBERTa**, **IgLM** models family and **ProtGPT2**. The ESM performs masked language modeling during the training and exhibited effectiveness in protein representation learning, predicting effects of variants, and protein structure modeling [31]. The AntiBERTa model is also pre-trained on masked language modeling which is further fine-tuned on token classification model to predict paratopes from the input prompt sequence. The latter models have a pre-training strategy which was based on autoregressive language modeling. These models have been successfully utilized in protein variant designs such as IgLM focuses on Antibody Sequence Infilling [32]. ProtGPT2, that can design de novo protein sequences using greedy search strategy, then selects the residues that have the highest probabilities to occur in the sequence. Autoregressive language models have shown the ability to generate diverse protein sequences that can adopt natural folds, even in cases where residue composition shows significant divergence [33]. Some of these generated sequences have been noted to retain enzymatic activity which has been reported to be comparable to the naturally occurring proteins. Autoregressive language models have also demonstrated their effectiveness as powerful zero-shot predictors of protein fitness, with performance occasionally improving as the scale of the model increases.

Language models have also been pre-trained in the antibody modeling tasks specifically using the masked language modeling, these models were trained on the sequences in Observed Antibody Space (OAS) database. One such model is AntiBERTa, which is a single masked language model that was trained on a corpus data of 67M sequences (with 86 that included both heavy and light chain sequences [34]. AntiBERTa representations were used for the paratope prediction task from a provided prompt sequence. **BioPhi** developed by Merck, a platform based on **Sapiens** which is a pair of masked language models trained distinctly on each chain (each with 569K parameters) light and heavy. The heavy chain model was trained on 20M sequences of heavy chains, and the light chain model was trained on 19M sequences of light chains respectively, model also showed its effectiveness in humanization of the antibody as well [35]. Autoregressive models have been trained on nanobody sequences and used for library design; these models are capable of antibody sequence generation as well [16]. The study [36] conducted experimental validation on a set of nanobody sequences generated with autoregressive generative models, including generated CDR3 loops. The results demonstrated significant enhancements in viability and binding affinity of the sequences when compared with conventional methods, even though the generated sequences were over 1000-fold smaller. But the model was unidirectional, it could not directly perform well to redesign the CDR3 loop within and required needed to be oversampled to generate sequences that matched the residues that followed the loop. [32] developed IgLM models family, two models were trained on autoregressive language modelling, one being IgLM and second IgLM-S which is smaller in size compared to IgLM (IgLM having 12M parameters and IgLM-S having 1.4M parameters), the IgLM family used a bidirectional encoding method and was trained on 558M sequences space including both heavy and light chain variable sequences. These models were able to address the limitation of oversampling. However, substantial computational resources were utilized to train these models on a large antibodies dataset. Moreover, the dataset of 558M sequences included antibody sequences from different organisms' studies. Therefore, it creates potential bias problems when generating sequences for mice and humans that limit the ability to generate diverse humanized sequences. Accordingly, there is further requirement for techniques that can generate diverse humanized libraries of de novo antibody sequences and relevant models contributing to antibody design, discovery, and optimization [32].

2.4 Study Rationale

The design and development of antibodies involve intricate processes that integrate both conventional and emerging methodologies. Display technologies, such as hybridoma screening and phage or yeast display, remain fundamental in identifying lead candidates and enhancing antibody affinity. However, these approaches are resource-intensive and limited in throughput, constraining the ability to optimize multiple parameters simultaneously. Despite their successes, challenges such as low solubility, high viscosity, and immunogenicity often hinder the transition of lead candidates to clinical viability.

Innovative solutions like protein language modeling (PLM), which is a subfield of natural language processing (NLP) have recently emerged as promising alternatives for antibody sequence design. By leveraging large language models (LLMs), PLM enables the generation of diverse and affinity-matured antibodies that mimic natural sequences. Models like AntiBERTa, IgLM, and ProtGPT2 illustrate the potential of machine learning to address complex sequence-function relationships, streamline optimization, and create de novo protein sequences with therapeutic relevance.

While these models have demonstrated significant advancements, limitations persist. Issues such as dataset bias, computational resource demands, limited affinity maturation, and restricted humanization capabilities remain barriers to broader adoption. Addressing these challenges is critical to furthering the development of robust and clinically viable antibody libraries.

2.5 Aims and Objectives

The following objectives have been proposed for this study:

- To address the limitations of display technologies by Deep Learning and Generative AI as a parallel assistive technology.
- To address the limitations of unidirectional nature of Autoregressive Language Models.
- To achieve data efficient and less compute-intensive trained models and automated protein sequence data engineering frameworks.
- To achieve diversification in the antibody for affinity maturation towards the target while avoiding immunogenicity issues.

CHAPTER 3: METHODOLOGY

This chapter highlights the methodology focused on the source and criteria to select and retrieve human antibody sequences. Also highlighting the steps to prepare the dataset and tokenization method applied for unsupervised training of the antibody language models and formulation of antibody sequence design. Furthermore, the steps and methods used for validation of the generated antibody sequences and desirability analyses for the therapeutic criteria have been explained.

3.1 Antibody Dataset Preparation

The dataset was prepared to train two generative models, i.e. Masked Language Model (MLM) and Autoregressive Language Model (ALM). The human antibody sequences were retrieved from the Observed Antibody Space (OAS) [37], on 22nd October 2022. The antibody sequences were checked for any sequencing errors as directed by OAS. Sequences were annotated using AbRSA: which is a numbering tool used for antibody numbering and CDRs delimiting [38]. To annotate CDRs, Chothia numbering scheme was applied. The sequences were required to have more than 15 residues before CDR1 and about 8-10 residues following CDR3. The sequences were filtered out to contain 6-10 residues in CDR1, 2-8 residues in CDR2 and CDR3 containing 10-35 residues. The entire set of sequences comprised of 1.03M unique sequences containing 688,614 unpaired heavy chain sequences and 350,590 unpaired light chain sequences. The data split of 80:10:10 ratio was applied into disjoint training, validation and test sets respectively each set containing both unpaired heavy and light chains. Finally, the training set comprised of 831,360 sequences, validation and test sets containing 103,921 sequences each.

3.2 Antibody Sequence Tokenization

In the context of a Large Language Model (LLM), tokens are the fundamental units of input and output. In natural language processing tasks, tokens often correspond to words, subwords, or characters. During both training and inference, the LLM handles input text as a series of tokens, with each token representing a specific word or symbol. For an LLM, the input text is referred to as a 'prompt,' while the output text is termed a 'completion'. The entire set of tokens that models use during training and inference is called its vocabulary. To train the models, a vocabulary of 25 tokens was used: the standard 20 amino acid single letters and five special tokens were used. The special tokens are: <*s*>, <*/s*>, <*pad*>*,* <*unk*>*, and* <*mask*>*,* each amino acid residue works like a single token, therefore, no byte-pair encoding was used. The full-length antibody sequence was encoded as a sentence, each sequence was encoded with start <*s*> token and end <*/s*> token, to align sequences of different lengths, <*pad*> tokens were used for padding out tensors to maximum allowed length of the sequence which was set at 150 to cover complete Fv region of the antibody sequence along with <*s*> and <*/s*> tokens while minimizing the amount of unnecessary padding, the <*mask*> was used as the masking token that during the MLM. To handle ambiguous residues such as X, <*unk*> token was used.

3.3 Antibody Sequence Design Formulation

Antibody sequence design can be formulated into two tasks: **(i)** Sequence Infilling and **(ii)** Full-length Sequence Generation. The sequence infilling task involves designing spans of residues and is like the text-infilling in natural language processing. Whereas the sequence generation is characterized as full-length antibody sequence generation through autoregressive sampling. The antibody sequence is represented as $S = (r_{1,} r_2, ...r_n)$, and r_i here, is the residue at the position *i* in the antibody sequence.

3.3.1 Infilling Design

The sequence is split into two main parts: the prefix, which is the span of residues preceding the infilling span, and the suffix, which is the span of residues succeeding the infilling span. To generate an infilled span of length *m* at position *k* along the antibody sequence, the span of residues $V=(r_k,r_{k+1}, ...,r_{k+m-1})$ is removed. Then the infilled span is generated to replace the removed section, forming a sequence by concatenating the prefix, resulting in the infilled sequence $S_{\backslash V}$ with the infilled span $(r_k',r_{k}'_{+1}, ...,r_{k}'_{+m-1})$ and the suffix. The resulting sequence structure can be described as follows:

The original sequence can be written as:

$$Orignal\ Sequence\ =\ Prefix\ +\ Infilling\ Span\ +\ Suffix$$

$$S\ =\ (r_1,r_2,\dots r_{k-1})\ +\ (r_k,r_{k+1},\dots r_{k+m-1})\ +\ (r_{k+m},r_{k+m+1},\dots r_n)$$

$$S\ =\ (r_1,r_2,\dots r_{k-1},r_k,r_{k+1},\dots r_{k+m-1},r_{k+m},r_{k+m+1},\dots r_n)$$

The generated infilled sequence becomes:

$$Infilled\ Sequence\ =\ Prefix\ +\ Infilled\ Span\ +\ Suffix$$

Therefore, by replacing the term *(r_k, r_{k+1}, ... r_{k+m-1})* with *(r_k',$r_{k}'_{+1}$, …,$r_{k}'_{+m-1}$)* we get the equation as:

$$S_{\backslash V} \; = \; (r_1, r_2, \ldots r_{k-1}) \; + \; (r'_k, r'_{k+1}, \ldots r'_{k+m-1}) \; + \; (r_{k+m}, r_{k+m+1}, \ldots r_n)$$

$$S_{\backslash V} = (r_1, r_2, \ldots r_{k-1}, r'_k, r'_{k\,+\,1}, \ldots, r'_{k\,+m\,-\,1}, r_{k+m}, \ldots r_n)$$

Hence, a fully formed equation with sequence of tokens ***Y*** for Infilling Language Model is:

$$\boldsymbol{Y} \; = \; (< s >, r_1, r_2, \ldots r_{k-1}, r'_k, r'_{k\,+\,1}, \ldots, r'_{k\,+m\,-\,1}, r_{k+m}, \ldots r_n, </s >)$$

The inspiration for the infilling was derived from the Infilling Language Modeling (ILM) framework proposed in the natural language infilling [39] to learn the probability distribution *p(V|$S_{\backslash V}$)*, then a deep autoregressive model was trained with parameters q maximizing *p(**Y**|q)*, that can be broken down into a product of conditional probabilities:

$$\max_q p(\boldsymbol{Y}|q) \; = \; \max_q \prod_{\boldsymbol{i}} p(\boldsymbol{Y_i}|\boldsymbol{Y}_{<i}, q)$$

Another infilling model was also trained on a similar infilling task that is based on the Masked Language Modeling (MLM). The inspiration for this model was drawn from a similar model that was trained through unsupervised learning across 250 million protein sequences to understand deep contextual representation spanning evolutionary diversity [40]. The end model has acquired knowledge representations related to biological features. Therefore, a MLM has been pre-trained on the antibody sequences dataset and 15% of the residues in the sequences were set to perturbation. Out of these, 80% of the residues get replaced by <*mask*> token, 10% by the random residues and 10% by the original residues. The masking ratios, retained from the original paper of the Bidirectional Encoder

Representations from Transformers (BERT) [41], are demonstrated to be optimal. The model infills the masked residues at the perturbed positions, *m* in the sequences during pre-training. The calculated MLM loss for a sequence $S = (r_1, r_2, \ldots r_n)$ of a batch *B* is:

$$\mathcal{L}_{MLM} = -\frac{1}{|B|}\sum_{S \in B_i \in m}\sum log\,\hat{p}(r_k|S_m)$$

3.3.2 Full-length Sequence Design

The sequence infilling task requires a parent sequence that is infilled at an infilling span *V* at specified positions replaced by $(r_k{'}, r_{k'+1}, \ldots, r_{k'+m-1})$ and MLM-based infilling using <*mask*> token results in generation of infilled sequences. Full-length sequence generation involves Autoregressive Language Modeling (ALM), which is left to right unidirectional language generation task as opposed to infilling in natural language processing suggested in the original paper for Generative Pre-Trained Transformer (GPT2) [42]. The full-length antibody sequences can be autoregressively sampled from left to right in unidirectional context. However, unidirectional sequence generation is prone to oversampling in the case of antibody sequence generation and often results in generation of undesired residues [36]. To generate reasonable sequences of varying lengths, bidirectional context was introduced, such that the full-length sequence generation can occur from both ends, resulting in coherent antibody sequences. As the OAS database used for training features sequences starting from the N-terminus, a prompting strategy was devised. This strategy involved providing the initial 3-4 motifs distinctly for the generation of heavy and light chain sequences. The autoregressive modeling does not require <*mask*> token, as in the case of infilling task. The model autoregressively samples full-length sequence through $p(\boldsymbol{Y_i}|\boldsymbol{Y_{<i}})$,

hence breaking down the sequence prediction into next-token prediction, thus, sampling an entire new sequence, where $\boldsymbol{Y}_{<i}$ denotes the tokens preceding $\boldsymbol{Y}_i$, therefore, final softmax layer predicts the probability distribution in discrete tokens as:

$$p(\boldsymbol{Y}) = \prod_{\boldsymbol{i}}^{\boldsymbol{n}} p(\boldsymbol{Y_i}|\boldsymbol{Y_{<i}})$$

3.4 Models Implementation and Training

Two types of transformers-based LLMs were implemented to train on the antibody dataset. Two Masked Language Models and Two Autoregressive Language Models were implemented varying in their sizes and hyperparameters. The training was unsupervised as implemented in the HuggingFace Transformers library. The version **4.27.4** of Transformers was implemented with Python version **3.9** for this task. Two RoBERTa (Robust BERT) based MLMs were implemented to train for Masked Language Modeling. And for Autoregressive Language Modeling, one standard GPT2 and another modified version of GPT2 based ALMs were trained, the training batch size for RoBERTas was set at 96 trained at 225000 and 30,000 steps respectively for RoBERTa and RoBERTa-S (S for small) and for GPTs batch size was 64 with training steps of 225000 and 162,500 respectively for GPT2 and GPT2-S with 1 gradient accumulation step for both transformers. The models were trained on a single NVIDIA A10G GPU. The training hyperparameters for both models are mentioned in Table 3.1Table 3.1 .

Table 3.1: Hyperparameters used for models training

Hyperparameter	RoBERTa	RoBERTa-S	GPT2	GPT2-S
N of Layers	12	3	12	3
Embed Dimensions	768	768	768	768
Hidden Dimensions	768	768	152	768
N of Attention Heads	12	3	12	6
Feed Forward Layer	3072	2048	2048	2048
Total Parameters	85,784,857	17,745,433	85,191,936	21,874,176

3.5 Antibody Library Generation

The goal of this study is to train AI models that are capable of designing humanized, diverse and affinity matured antibodies. To achieve and test the task of antibody library generation, the task was divided into two parts:

I. Antibody CDR3 redesigning for binding optimization

II. Full-length antibody sequence generation.

Both parts provide key insights about antibodies behavior and properties and necessary information about desired qualities and features to consider when designing *de novo* antibodies. And how the trained models perceive information from the antibody sequence.

3.5.1 Antibody CDR3 redesigning for binding optimization

Antibody affinity maturation is the cornerstone in the field of immunotherapeutic. Antibody's strong binding to the target provides it with the quality of its specificity to that target. Therefore, to test the models output in terms of antibody optimization, CDR3 for the antibody 1E6J which was downloaded from Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (https://www.rcsb.org/), an antibody against HIV-1 strain: B_HXB2R capsid protein p24, used as a diagnostic antibody in HIV-1 screening tests. The p24 protein is considered as highly conserved protein in various HIV-1 strains due to its less susceptibility to mutations [43]. The complex was downloaded from the PDB, since the region of interest for this study were the variable fragments (Fv) sequences of the antibody, as all the target binding CDR regions are present in Fv regions, therefore, the Fv sequences for both heavy (H) and light (L) chains were extracted using Chothia numbering scheme. Since most of the variability and binding exist in the CDR3 regions, both chains were annotated through AbRSA [38]. The CDR3 for H and L chain were found between the residue 95 to 102 and 89 to 97 respectively.

3.5.2 CDR3 design through Masked Language Modeling (MLM)

To generate de novo sequences through MLM, the CDR3 sequences of both H and L chain, the CDR3 regions between 95 to 102 and 89 to 97 respectively, were masked with the token 'X'. Both Roberta-MLMs were employed to design de novo sequences in unpaired manner, the total number of designed antibodies after pairing all *de novo* H and L sequences was 200. The sequences were further processed for structural fold prediction.

3.5.3 CDR3 design through Infilling Language Modeling (ILM)

To generate the sequences through ILM, the infill_range function was defined for autoregressive GPT2-models. The infill_range directs the models to infill new tokens autoregressively from minimum to maximum range. Therefore, the infill_range here was set for CDR3 regions of H and L chain respectively, which were generated in unpaired manner as well. The total number of antibodies generated after pairing H and L *de novo* sequences was 200, that were processed to predict 3D structures.

3.5.4 Full-length Antibody Sequence Generation

To generate full-length antibody sequences, the autoregressive token sampling was employed. For this task, the GPT2-S was employed, the initializing tokens were extracted from the **Dataset_S02** provided in the paper [44]. The dataset is composed of 242 therapeutic antibodies for different diseases. The initial three or four tokens were extracted for both H and L chains of the antibodies, the tokens were checked for redundancy and made sure to only consist unique tokens. The sequence generation was achieved with a combination of sampling temperatures *($T \epsilon \{0.5, 0.75, 1.0\}$)* with corresponding nucleus probabilities *($P \epsilon \{0.6, 0.8, 1.0\}$)*. The temperature is the sampling parameter for the degree of randomization, its greater value represents more randomness whereas the nucleus probability indicates the selection of the tokens with the highest probabilities. A total of 3000 sequences were generated in an unpaired manner that comprised of 1500 H chains and 1500 L chains, which were further analyzed and evaluated for humanness which serves as a proxy for immunogenicity. BioPhi (https://biophi.dichlab.org/) webserver was employed with relaxed prevalence threshold to check OAS percentile for the humanness of the sequences [35].

3.6 Antibody Structure Fold Prediction

Proteins have the tendency to fold into 3D dimensional space, this is directly linked to the sequential arrangement of the amino acids due to their nature and the functions they perform. Therefore, antibodies are an important class of proteins that also exhibit protein folding properties. In order to observe foldability in antibodies, the structure prediction was employed for the 400 paired mutant antibodies that were generated through CDR3 redesign. To predict structures, tFold-Ab was utilized, as its reliability and precision surpasses AlphaFold and IgFold [45]. The structure of the extracted Fv of the wild-type antibody was also predicted through tFold-Ab. The structures were visualized in the PyMOL (https://www.pymol.org) tool, which is a protein visualization and analysis tool. For this study, the open-source version was used.

3.7 Antigen-Antibody Complex Prediction

Protein-Protein docking is the *in silico* method of predicting the structure of the protein-protein complex from the given individual structures. This technique provides insights into the binding interactions between the target and the receptor proteins and their binding affinity [46]. Antibodies are also a type of protein receptor that have specificity of binding with their respective targets. So, to predict the antigen-antibody complex structures, an open-source tool ClusPro webserver was used (https://cluspro.bu.edu/) [47], [48], [49], [50], [51]. ClusPro has a dedicated “Antibody Mode”, that is specifically conditioned to

Antigen-Antibody docking [52]. The complex models are ranked by the cluster size, the more members in the cluster interpret better complex model.

3.8 Antigen-Antibody Complex Binding Affinity Prediction

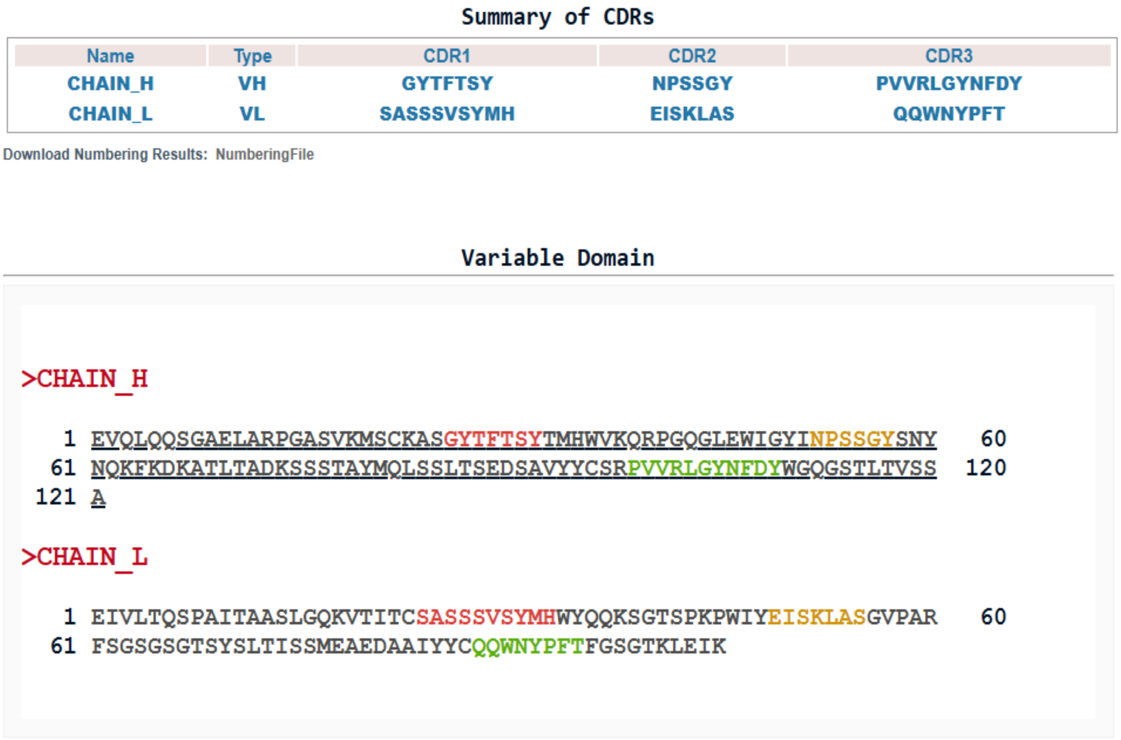
Protein-Protein complex binding affinity is a measure of the strength of the interaction of two proteins in a complex. This is the measure of the protein dissociation constant (K_d) during the formation of the binding complex represented in nanomolar (nM). The lesser value of this constant represents strong binding affinity. The other measuring unit is called Gibbs Free Energy (ΔG), which is the value of the energy released during the complex formation measured in $kcal/mol$, it is represented in negative value, the greater negative value represents stronger binding affinity, thus more stable complex. To predict the binding affinity, PRODIGY (**PRO**tein bin**DI**ng ener**GY**) tool was employed, available at (https://rascar.science.uu.nl/prodigy/) [53], [54]. The binding affinity was predicted for the 400 complex models docked though ClusPro.

CHAPTER 4: RESULTS

The major aim of this study is to develop a simple and cost-efficient framework to design affinity matured antibodies that can assist display technologies in selection and isolation of high-affinity antibodies. The results are divided into five sections, the first section provides the method for the antibody numbering and CDR delimiting performed through AbRSA. The second section provides insights into the antibody structure prediction through tFold-Ab and structure quality assessment that is followed by the third section which explains results for humanness/immunogenicity evaluation for the full-length generated antibodies. Lastly, the fourth section deals with the binding optimization and binding affinity prediction for the mutated antibodies.

4.1 AbRSA for CDR delimitation

AbRSA incorporates the biological information of the antibody features specific to the regions through dynamic programming that improves the robustness of antibody numbering. To annotate CDRs of 1E6J, AbRSA was employed, for this study the CDR3 for H and L chain were considered for design interest. Therefore, the CDR3-H was indicated from residue 95 to 102 sequence “PVVRLGYNFDY” where residue 89 to 97 for CDR3-L sequence “QQWNYPFT”. The Figure 4.1 shows CDRs for H and L chain indicated in the sequence.



Summary of CDRs

Name	Type	CDR1	CDR2	CDR3
CHAIN_H	VH	GYTFTSY	NPSSGY	PVVRLGYNFDY
CHAIN_L	VL	SASSSVSYMH	EISKLAS	QQWNYPFT

Download Numbering Results: NumberingFile

Variable Domain

```
>CHAIN_H

  1 EVQLQQSGAELARPGASVKMSCKASGYTFTSYTMHWVKQRPGQGLEWIGYINPSSGYSNY   60
 61 NQKFKDKATLTADKSSSTAYMQLSSLTSEDSAVYYCSRPVVRLGYNFDYWGQGSTLTVSS  120
121 A

>CHAIN_L

  1 EIVLTQSPAITAASLGQKVTITCSASSSVSYMHWYQQKSGTSPKPWIYEISKLASGVPAR   60
 61 FSGSGSGTSYSLTISSMEAEDAAIYYCQQWNYPFTFGSGTKLEIK
```

Figure 4.1: The figure displays the arrangement of FRs and CDRs in the 1E6J H and L chain, The FR are indicated in black, where CDR1s for both chains are indicated in red, CDR2s in yellow and CDR3s in green.

4.2 Antibody Structural Folding and Quality Assessment

Proteins have the tendency to fold into 3D dimensional space, this property provides them with their basic functional ability. Antibodies are also an important class of proteins; to perform their functions properly, antibodies are folded into 3D space. Proteins generally tend to lose or enhance their functional abilities when exposed to mutations, this might be caused by the misfolds due to mutations. To predict structures for the mutated library of the retrieved antibody 1E6J, tFold-Ab was employed. The total of 400 structures were predicted. The mutated structures showed foldability like naturally occurring antibodies that were compared with wild-type 1E6J through superimposing. The structures for GPTs and MLMs were also assessed for quality using VoroMQA [55], a webserver for protein quality assessment at (https://bioinformatics.lt/wtsam/voromqa), which uses interatomic contact areas to assess the protein structure quality and assigns a per residue score to each residue. The overall quality of the structure can be assessed through the global score [56]. The average global score for MLM infilled structures is 0.587, which is promising, whereas the average global score for the GPT infilled antibody structures stood at the promising value of 0.581. The Figure 4.2 and Figure 4.3 shows the structural visualization of the CDR3-H for the infilled antibodies generated through MLM and ILM compared with the wild-type 1E6J respectively. While Figure 4.4 and Figure 4.5 disclose predicted antibody structures quality assessments in terms high-quality X-ray crystallography through VoroMQA global scores for MLM and GPT infilled structures respectively

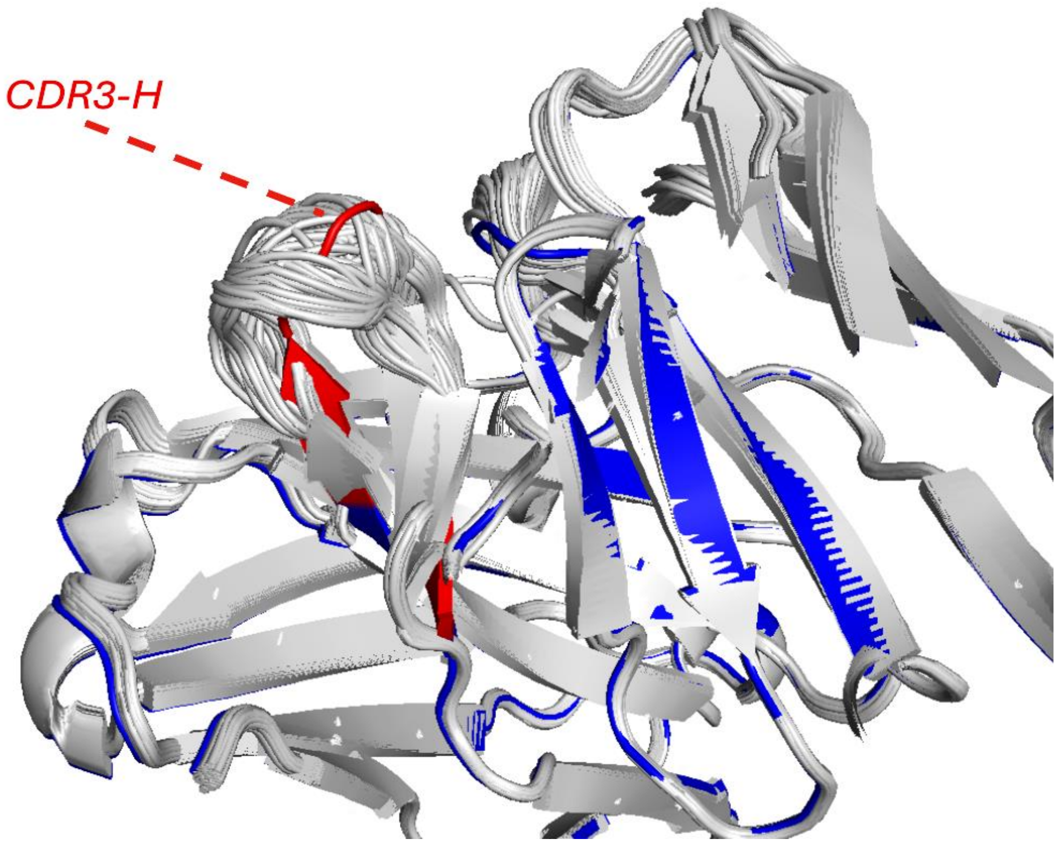


Figure 4.2: The figure displays the visualization of the superimposed structure of the wild-type 1E6J with mutant. The actual 1E6J is colored blue, the CDR3-H Loop (Unmutated) is highlighted in red. The whites are the structures that were predicted for the infilled sequences.

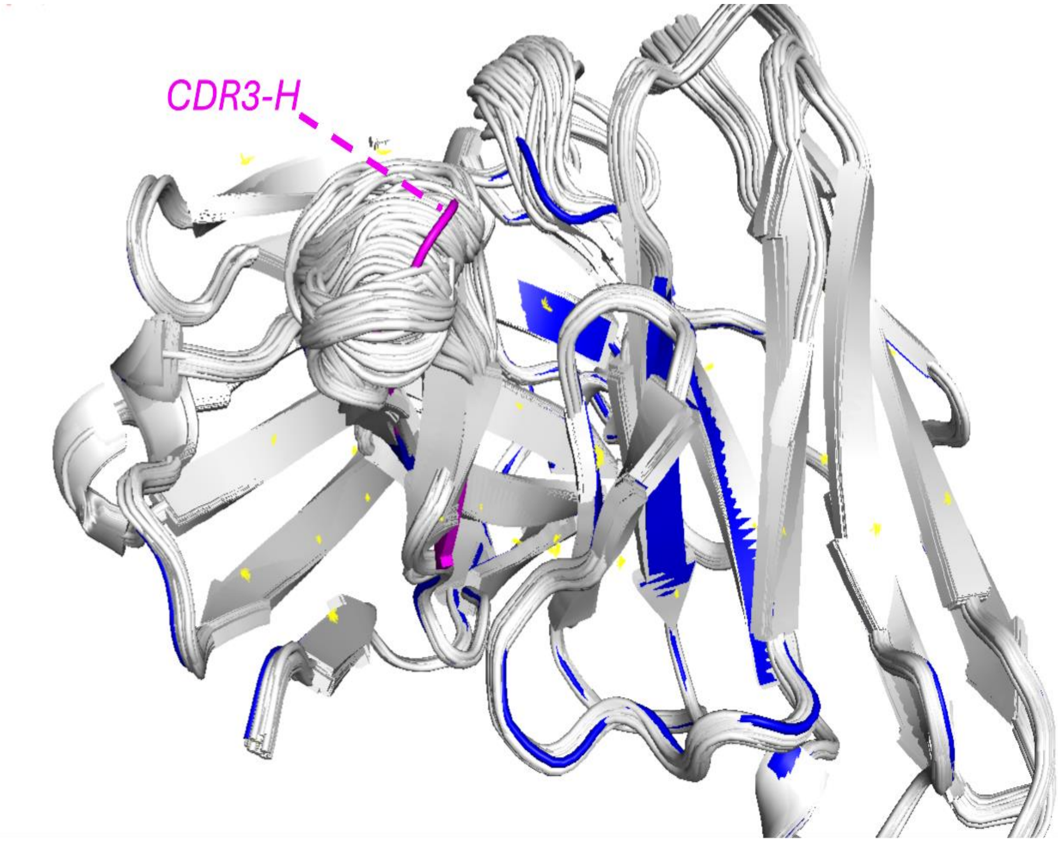


Figure 4.3: Alignment of 1E6J antibody with the de novo CDR3-H designed antibodies through ILM, the wild-type 1E6J is represented with blue and CDR3-H (wild-type) is highlighted in magenta color. The white structures represent infilled ones.

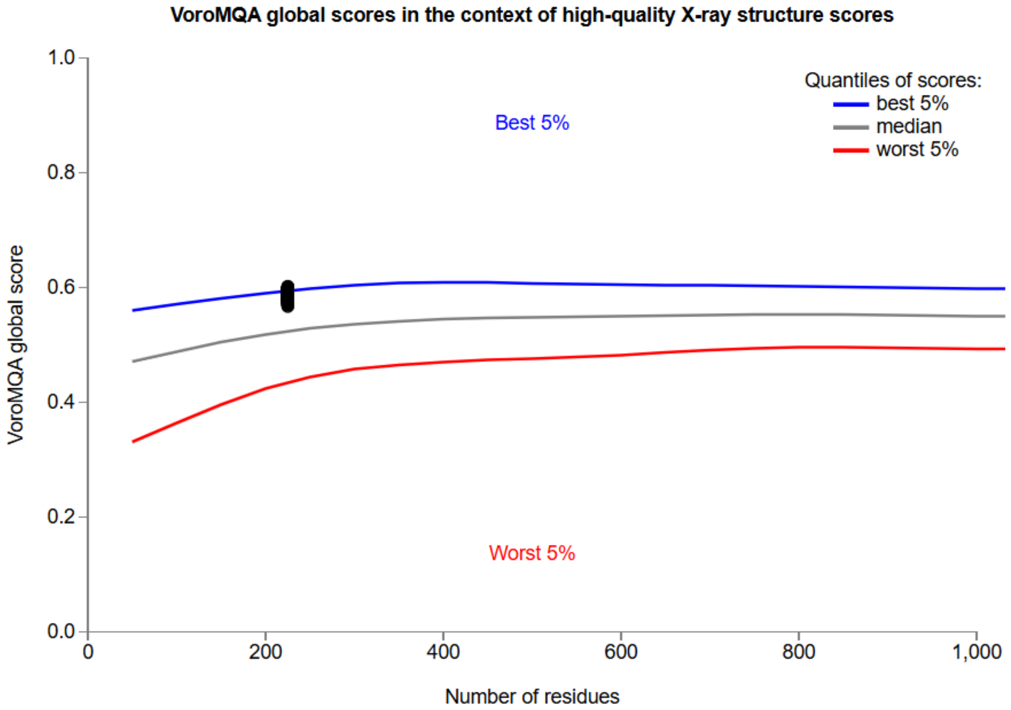


Figure 4.4: This plot shows the VoroMQA global scores for the MLM generated antibodies, a metric for structural quality, across a range of protein sizes (number of residues). The data is grouped into quantiles representing the best 5% (blue), median (gray), and worst 5% (red) of scores from high-quality X-ray structures. The black dots concentrated at the plot represent each structure occurring at the global score quantiles.

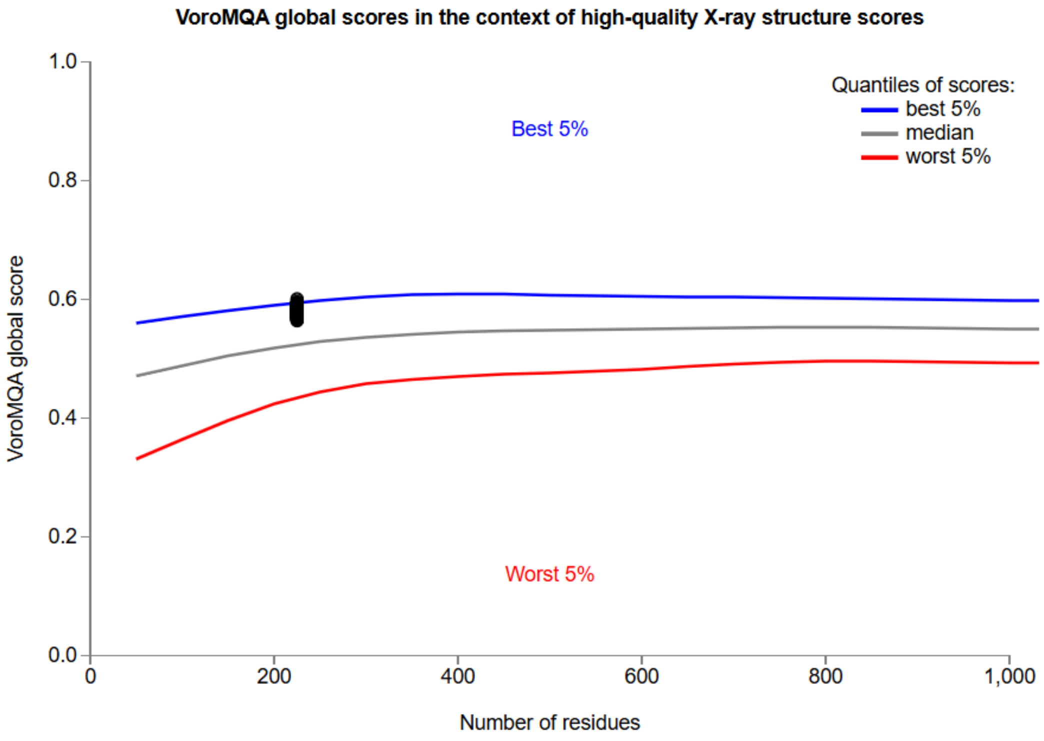


Figure 4.5: The plot displays the global scores for the infilling generated antibodies in the context of high-quality X-ray structures across differing protein lengths. The data is presented in quantiles showing the best 5% (blue), median (gray), and worst 5% (red) of scores. The clustering of black dots near the 200-residue mark represents individual structures at their respective global score quantiles. The result from the VoroMQA for the predicted structures of the GPT infilled antibodies.

4.3 Humanness Evaluation

Antibodies humanness provides the insights for the antibodies to their degree of human compatibility. Humanness metric can be used as a proxy to evaluate the immunogenicity of the antibodies. BioPhi (https://biophi.dichlab.org/) [35] platform to evaluate humanness for the 3000 de novo generated antibodies through bidirectional autoregressive modeling using the combination of three sampling temperatures corresponding with three nucleus probabilities that are *(T ϵ {0.5, 0.75, 1.0})* and *(P ϵ {0.6, 0.8, 1.0})* respectively. The H chain sequences show 96% OAS percentile in the 1st batch of 1000 sequences with *{T=0.5,* *P= 0.6}* where L chains show less variability at 99% OAS percentile with the presence of duplicate sequences which by nature L chains show less variability compared to their H chains counterpart. The 2nd batch of 1000 sequences with sampling parameters *{T=0.75, P= 0.8}* displayed 89% OAS percentile for H and 99% OAS percentile for the L chains. Therefore, 3rd batch sequences with sampling values *{T=1.0, P= 1.0}* exhibited a significant drop in H OAS percentile which stood at 57% whereas the L chain OAS percentile was 82% as displayed in the Figure 4.6. The OAS percentile indicates the sequences show greater degree of humanness with lesser sampling parameter values.

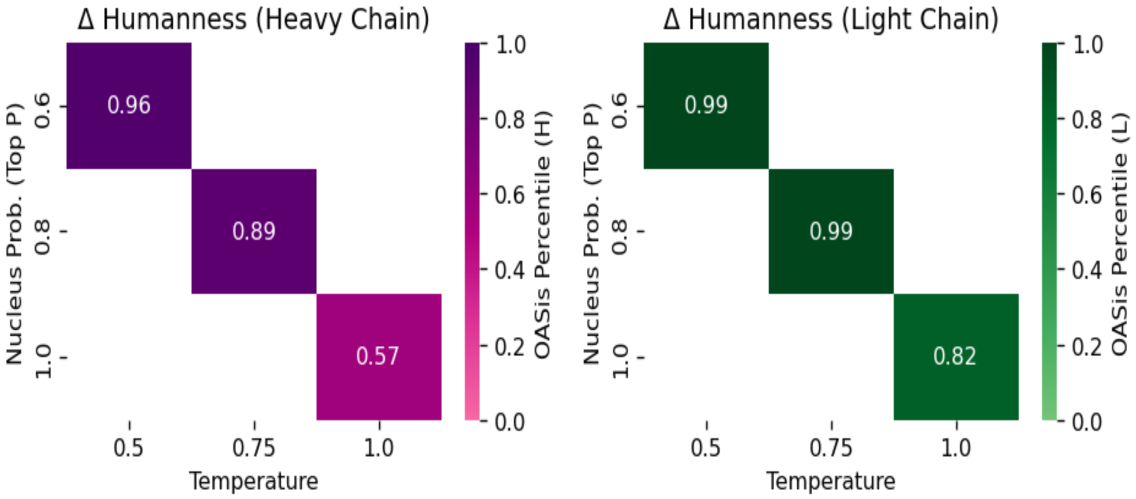


Figure 4.6: The plots indicate the change in the individual OAS percentiles with respect to the corresponding temperature and nucleus probabilities sampling values. The OAS percentiles (humanness) decrease with respect to increase in the sampling parameter values.

4.4 Antigen-Antibody Complex Binding Analysis

Antigen-Antibody Complex Binding Analysis involves the formation of docked complex from the individual components like antigen and antibody. To perform antigen-antibody interaction, ClusPro was employed using the “Antibody Mode” [52]. ClusPro uses the clustering method to rank the best docked model. Therefore, the cluster with greatest size i.e. that has maximum number of members is considered the best conformation and stable according to ClusPro. The docking was performed on 400 antibodies that were subject to CDR3 design, the docked models with max number of cluster members were retrieved, then to predict ΔG, PRODIGY was used. The results were compared with wild-type antibody that had the predicted ΔG = -13.0 $kcal/mol,$ where the experimentally reported K_d for 1E6J is 29 nM that is ≈ 2.9e-10 M. Therefore, 10 out of 400 antibodies showed significant improvement in ΔG values. Table 4.1, Table 4.2 and Figure 4.7, Figure 4.8 indicate the ΔG prediction results from the PRODIGY for the antibodies generated through MLM and ALM respectively. The highlighted cells indicate significant improvement in binding affinity values. The greater ΔG values indicate strongly binding antibodies.

Table 4.1: The results obtained from PRODIGY indicating improved Binding Affinity value at **Complex_106** using MLM.

Complex No.	K_d	B. Affinity (ΔG)	Contacts	Cluster Members
Parent (wildtype)	2.90E-10	-13.0	77	92
Complex_58	2.40E-10	-13.111	113	87
Complex_81	2.30E-10	-13.142	76	93
Complex_92	2.30E-10	-13.142	76	93
Complex_92	2.30E-10	-13.151	90	79
Complex_106	1.10E-11	-14.955	78	96

Complex_170	2.50E-10	-13.102	64	101

Table 4.2: Table indicating results for ALM generated antibodies complexes, the highlighted rows indicate significant improvement in the binding affinities compared with the Parent (wild-type). The increase in contacts number can be correlated with the increase in the specificity of antibodies.

Complex No.	K_d	**B. Affinity (ΔG)**	**Contacts**	**Cluster Members**
Parent (wildtype)	2.90E-10	-13	77	92
Complex_1	4.40E-11	-14.118	95	87
Complex_9	1.10E-10	-13.552	97	115
Complex_12	3.50E-11	-14.256	116	101
Complex_19	9.10E-13	-16.418	112	94
Complex_33	1.90E-10	-13.261	85	73
Complex_41	2.60E-11	-14.423	76	79
Complex_53	1.80E-10	-13.291	71	109
Complex_54	2.60E-10	-13.065	66	96
Complex_72	1.30E-10	-13.459	64	67
Complex_76	1.40E-10	-13.442	92	73
Complex_101	1.60E-10	-13.367	72	119

Complex_107	2.80E-13	-17.119	135	86
Complex_111	7.10E-14	-17.928	139	89
Complex_112	2.20E-12	-15.897	131	90
Complex_114	1.30E-12	-16.211	112	85
Complex_182	1.00E-10	-13.631	92	126

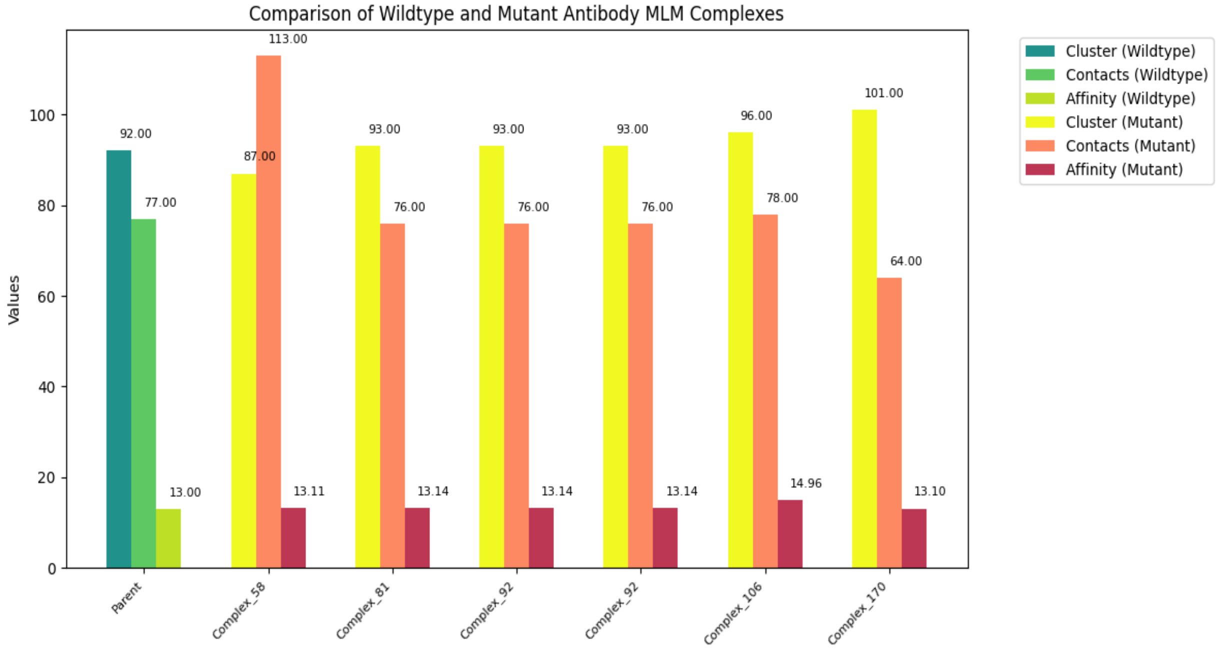


Figure 4.7: Grouped-bar plot highlighting the binding affinity, cluster size and contacts for the wildtype complex vs mutant antibodies complexes for the MLM, the unique colormap is used to indicate the wildtype values where the AI-designed antibodies are in same color scale.

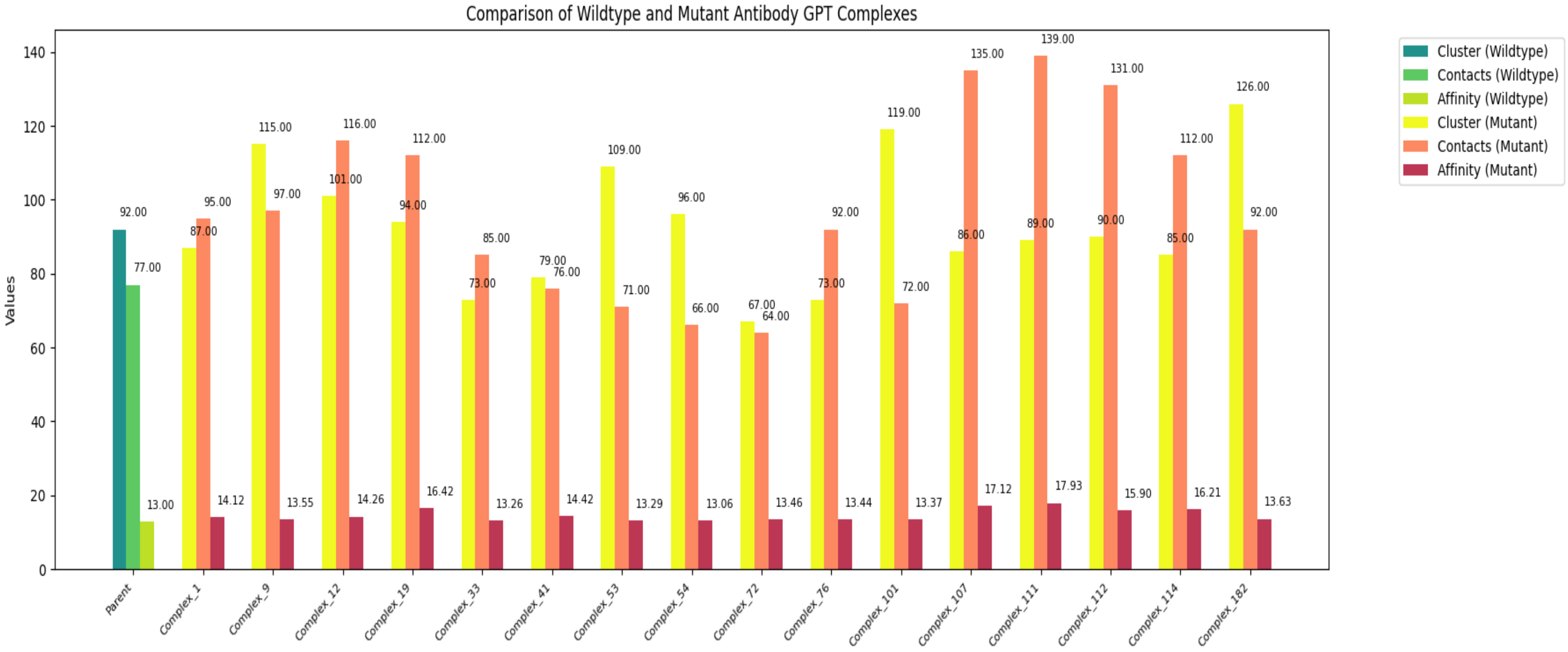


Figure 4.8: The plot indicates the comparison of binding affinities, cluster members and contacts in wildtype complex in the unique colormap vs mutant (AI-designed) antibodies in the same colormap, generated through GPT infilling.

CHAPTER 5: DISCUSSION

The human immune system is a complex network of different biological processes. The immune system is the major player in human biological defenses against a variety of pathogens. It is composed of different cellular and subcellular entities that can work coordinatively against pathogens and neutralize them. It has the sophisticated ability to recognize human cells and distinguish them from pathogens. And it can remember pathogens even after the infection is cured. This sophisticated system is divided into two main lines of defenses, first the innate immunity and second the adaptive immunity. The innate immunity is the first line of defense that activates in a short time and rapidly neutralizes the pathogen but cannot memorize the pathogen whereas the adaptive immunity uses specialized proteins called “antibodies”, that makes it possible to memorize pathogens if they invade in the future. Antibodies have the ability to bind against the specific antigen which makes them an important class of therapeutics.

The predictability of antigen-antibody interaction has always been the fundamental question in the field of immunology. Different techniques have been developed to discover high-affinity antibodies against particular antigen. For this purpose, techniques like display technologies have been developed to isolate high-affinity antibodies but these techniques rely on the target to hit approach for the lead discovery which makes complex and resource intensive and most of the antibody leads are discarded to isolate potential leads that have a high-specificity and high-affinity. And the factors like viscosity, pharmacokinetics and immunogenicity hinder the discovery of potential leads identification. To overcome these problems, computational methods have developed particularly with the use of generative

AI. Antibodies are a class of protein, and they have the tendency to exhibit protein-like properties and features. Therefore, researchers have employed various generative AI-driven transformer-based Large Language Models (LLMs) that were originally intended to generate human-like textual information but show promise in the field of antibody design and engineering. Particular examples include: IgLM and AntiBERTa [32], [34] with different serving purposes with respect to their intended use. The LLMs treat antibody sequence just like a textual language and can learn the patterns and information from the amino acids arrangements due to the attention mechanism. However, these models show limitations and biases when used to generate diversified human antibodies and have been trained on the dataset of different organisms with greater resources. This study proposed a class of four models, **AbSynth**, which includes two RoBERTa-based Masked Language Models (MLMs) and two GPT-based Autoregressive Language Models (ALMs).

The dataset used for this study was retrieved from OAS database, which was based on 1.03 million sequences of human antibodies 558 times less data than IgLM that were divided into 80, 10, and 10 ratios. GPT-based transformers are unidirectional decoders by nature which make them prone to oversampling problems resulting in undesired tokens in the sequence. To address this issue, bidirectionality was introduced in a GPT-S by modifying the architecture of the model to achieve variable lengths of antibody sequences. Furthermore, the study employed trained models to redesign CDR3 regions for 1E6J antibody and another 3000 *de novo* sequences were generated through combination of sampling parameters.

The study proposed that the generated sequences display foldable structures like naturally occurring antibodies. The study also showed that the generated antibodies show binding with the naturally existing target and 10 out of 400 generated antibodies show better binding affinity compared to the natural antibody. The study also showed that *de novo* generated sequences show average humanness of 96% at low sampling parameters in the H chains with diverse sequences and the diversity increases with decrease in humanness. Thus, reducing the risk of immunogenic response. The study also showed the L chains show less variability compared to the H chains.

It is seen that the study has achieved a suitable antibody library generation method with potential high-affinity antibody leads and can generate diverse humanized de novo full-length antibodies with high-affinity maturation. Currently, the trained models are antigen agnostic, therefore integrating the models with target antigen data will improve the models' specificity in antibodies lead generation.

CHAPTER 6: CONCLUSIONS AND FUTURE RECOMMENDATION

This study introduced a class of generative AI models, AbSynth, which are transformers based antibody language models, explaining its efficiency and effectiveness in antibody affinity maturation through CDR3 design and achieved diverse humanized antibodies through full-length de novo antibody design. The study shows that 10 out 400 AI-designed antibodies show greater affinity than the natural 1E6J antibody. Furthermore, the study showed that AbSynth autoregressive models can generate full-length diverse humanized antibodies with 96% humanness in H chains with low sampling parameter values.

AbSynth can significantly reduce resources and time to generate library of affinity matured antibodies with diverse humanization, this can serve as an effective in silico tool for antibody design with greater humanized precision equipped with the understanding of diverse context in human antibodies better thus, reducing the potential risk of immunogenicity.

Subsequent research on AbSynth will concentrate on enhancing antibody design accuracy and creating a portable design pipeline that integrates antigen data and permits sequence and structural co-design. In order to produce more cohesive, useful antibodies with better therapeutic qualities, this method can improve knowledge of antibody design contexts and integrate structural insights.

REFERENCES

[1] V. K. Vanguri, “The Adaptive Immune System,” in *Pathobiology of Human Disease*, Elsevier, 2014, pp. 1–4. doi: 10.1016/B978-0-12-386456-7.01101-1.

[2] S. McComb, A. Thiriot, B. Akache, L. Krishnan, and F. Stark, “Introduction to the Immune System,” in *Immunoproteomics*, vol. 2024, K. M. Fulton and S. M. Twine, Eds., in Methods in Molecular Biology, vol. 2024. , New York, NY: Springer New York, 2019, pp. 1–24. doi: 10.1007/978-1-4939-9597-4_1.

[3] J. S. Marshall, R. Warrington, W. Watson, and H. L. Kim, “An introduction to immunology and immunopathology,” *Allergy Asthma Clin. Immunol.*, vol. 14, no. S2, p. 49, Sep. 2018, doi: 10.1186/s13223-018-0278-1.

[4] J. M. Van Seventer and N. S. Hochberg, “Principles of Infectious Diseases: Transmission, Diagnosis, Prevention, and Control,” in *International Encyclopedia of Public Health*, Elsevier, 2017, pp. 22–39. doi: 10.1016/B978-0-12-803678-5.00516-6.

[5] G. A. Parker, “Cells of the Immune System,” in *Immunopathology in Toxicology and Drug Development*, G. A. Parker, Ed., in Molecular and Integrative Toxicology. , Cham: Springer International Publishing, 2017, pp. 95–201. doi: 10.1007/978-3-319-47377-2_2.

[6] R. Barrangou and L. A. Marraffini, “CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity,” *Mol. Cell*, vol. 54, no. 2, pp. 234–244, Apr. 2014, doi: 10.1016/j.molcel.2014.03.011.

[7] G. Altan-Bonnet, T. Mora, and A. M. Walczak, “Quantitative immunology for physicists,” *Phys. Rep.*, vol. 849, pp. 1–83, Mar. 2020, doi: 10.1016/j.physrep.2020.01.001.

[8] C. Janeway, Ed., *Immunobiology 5: the immune system in health and disease*, 5th ed. New York: Garland Pub, 2001.

[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, p. 55, Dec. 2019, doi: 10.3390/antib8040055.

[10] N. E. Weisser and J. C. Hall, “Applications of single-chain variable fragment antibodies in therapeutics and diagnostics,” *Biotechnol. Adv.*, vol. 27, no. 4, pp. 502–520, Jul. 2009, doi: 10.1016/j.biotechadv.2009.04.004.

[11] I. A. Wilson and R. L. Stanfield, “Antibody-antigen interactions: new structures and new conformational changes,” *Curr. Opin. Struct. Biol.*, vol. 4, no. 6, pp. 857–867, Jan. 1994, doi: 10.1016/0959-440X(94)90267-4.

[12] L. L. Lu, T. J. Suscovich, S. M. Fortune, and G. Alter, “Beyond binding: antibody effector functions in infectious diseases,” *Nat. Rev. Immunol.*, vol. 18, no. 1, pp. 46–61, Jan. 2018, doi: 10.1038/nri.2017.106.

[13] R. Dixit, J. Herz, R. Dalton, and R. Booy, “Benefits of using heterologous polyclonal antibodies and potential applications to new and undertreated infectious pathogens,” *Vaccine*, vol. 34, no. 9, pp. 1152–1161, Feb. 2016, doi: 10.1016/j.vaccine.2016.01.016.

[14] B. Valldorf *et al.*, “Antibody display technologies: selecting the cream of the crop,” *Biol. Chem.*, vol. 403, no. 5–6, pp. 455–477, Apr. 2022, doi: 10.1515/hsz-2020-0377.

[15] R. A. Norman *et al.*, “Computational approaches to therapeutic antibody design: established methods and emerging trends,” *Brief. Bioinform.*, vol. 21, no. 5, pp. 1549–1567, Sep. 2020, doi: 10.1093/bib/bbz095.

[16] 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.

[17] J. Graves *et al.*, “A Review of Deep Learning Methods for Antibodies,” *Antibodies*, vol. 9, no. 2, p. 12, Apr. 2020, doi: 10.3390/antib9020012.

[18] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, “Generative artificial intelligence: a systematic review and applications,” *Multimed. Tools Appl.*, Aug. 2024, doi: 10.1007/s11042-024-20016-1.

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

[20] J. A. Ruffolo, L.-S. Chu, S. P. Mahajan, and J. J. Gray, “Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies,” *Nat. Commun.*, vol. 14, no. 1, p. 2389, Apr. 2023, doi: 10.1038/s41467-023-38063-x.

[21] J. A. Ruffolo, J. J. Gray, and J. Sulam, “Deciphering antibody affinity maturation with language models and weakly supervised learning,” 2021, *arXiv*. doi: 10.48550/ARXIV.2112.07782.

[22] T. Jain *et al.*, “Biophysical properties of the clinical-stage antibody landscape,” *Proc. Natl. Acad. Sci.*, vol. 114, no. 5, pp. 944–949, Jan. 2017, doi: 10.1073/pnas.1616408114.

[23] D. Hu *et al.*, “Effective Optimization of Antibody Affinity by Phage Display Integrated with High-Throughput DNA Synthesis and Sequencing Technologies,” *PLOS ONE*, vol. 10, no. 6, p. e0129125, Jun. 2015, doi: 10.1371/journal.pone.0129125.

[24] A. B. Bos *et al.*, “Development of a semi-automated high throughput transient transfection system,” *J. Biotechnol.*, vol. 180, pp. 10–16, Jun. 2014, doi: 10.1016/j.jbiotec.2014.03.027.

[25] D. S. Tomar, S. Kumar, S. K. Singh, S. Goswami, and L. Li, “Molecular basis of high viscosity in concentrated antibody solutions: Strategies for high concentration drug product development,” *mAbs*, vol. 8, no. 2, pp. 216–228, Feb. 2016, doi: 10.1080/19420862.2015.1128606.

[26] E. M. Roth *et al.*, “Antidrug Antibodies in Patients Treated with Alirocumab,” *N. Engl. J. Med.*, vol. 376, no. 16, pp. 1589–1590, Apr. 2017, doi: 10.1056/NEJMc1616623.

[27] V. Greiff *et al.*, “Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires,” *J. Immunol.*, vol. 199, no. 8, pp. 2985–2997, Oct. 2017, doi: 10.4049/jimmunol.1700594.

[28] R. Fox *et al.*, “Optimizing the search algorithm for protein engineering by directed evolution,” *Protein Eng. Des. Sel.*, vol. 16, no. 8, pp. 589–597, Aug. 2003, doi: 10.1093/protein/gzg077.

[29] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” 2012, *arXiv*. doi: 10.48550/ARXIV.1211.5063.

[30] A. Vaswani *et al.*, “Attention Is All You Need,” 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.

[31] Z. Lin *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023, doi: 10.1126/science.ade2574.

[32] R. W. Shuai, J. A. Ruffolo, and J. J. Gray, “IgLM: Infilling language modeling for antibody sequence design,” *Cell Syst.*, vol. 14, no. 11, pp. 979-989.e4, Nov. 2023, doi: 10.1016/j.cels.2023.10.001.

[33] N. Ferruz, S. Schmidt, and B. Höcker, “ProtGPT2 is a deep unsupervised language model for protein design,” *Nat. Commun.*, vol. 13, no. 1, p. 4348, Jul. 2022, doi: 10.1038/s41467-022-32007-7.

[34] J. Leem, L. S. Mitchell, J. H. R. Farmery, J. Barton, and J. D. Galson, “Deciphering the language of antibodies using self-supervised learning,” *Patterns*, vol. 3, no. 7, p. 100513, Jul. 2022, doi: 10.1016/j.patter.2022.100513.

[35] D. Prihoda *et al.*, “BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning,” *mAbs*, vol. 14, no. 1, p. 2020203, Dec. 2022, doi: 10.1080/19420862.2021.2020203.

[36] J.-E. Shin *et al.*, “Protein design and variant prediction using autoregressive generative models,” *Nat. Commun.*, vol. 12, no. 1, p. 2403, Apr. 2021, doi: 10.1038/s41467-021-22732-w.

[37] T. H. Olsen, F. Boyles, and C. M. Deane, “Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences,” *Protein Sci.*, vol. 31, no. 1, pp. 141–146, Jan. 2022, doi: 10.1002/pro.4205.

[38] L. Li *et al.*, “AbRSA: A robust tool for antibody numbering,” *Protein Sci.*, vol. 28, no. 8, pp. 1524–1531, Aug. 2019, doi: 10.1002/pro.3633.

[39] C. Donahue, M. Lee, and P. Liang, “Enabling Language Models to Fill in the Blanks,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.05339.

[40] A. Rives *et al.*, “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” *Proc. Natl. Acad. Sci.*, vol. 118, no. 15, p. e2016239118, Apr. 2021, doi: 10.1073/pnas.2016239118.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018, *arXiv*. doi: 10.48550/ARXIV.1810.04805.

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners”.

[43] S. Monaco-Malbet *et al.*, “Mutual Conformational Adaptations in Antigen and Antibody upon Complex Formation between an Fab and HIV-1 Capsid Protein p24,” *Structure*, vol. 8, no. 10, pp. 1069–1077, Oct. 2000, doi: 10.1016/S0969-2126(00)00507-4.

[44] M. I. J. Raybould *et al.*, “Five computational developability guidelines for therapeutic antibody profiling,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 10, pp. 4025–4030, Mar. 2019, doi: 10.1073/pnas.1810576116.

[45] J. Wu, F. Wu, B. Jiang, W. Liu, and P. Zhao, “tFold-Ab: Fast and Accurate Antibody Structure Prediction without Sequence Homologs,” Nov. 13, 2022. doi: 10.1101/2022.11.10.515918.

[46] I. A. Vakser, “Protein-Protein Docking: From Interaction to Interactome,” *Biophys. J.*, vol. 107, no. 8, pp. 1785–1793, Oct. 2014, doi: 10.1016/j.bpj.2014.08.033.

[47] D. Kozakov *et al.*, “How good is automated protein docking?,” *Proteins Struct. Funct. Bioinforma.*, vol. 81, no. 12, pp. 2159–2166, Dec. 2013, doi: 10.1002/prot.24403.

[48] D. Kozakov *et al.*, “The ClusPro web server for protein–protein docking,” *Nat. Protoc.*, vol. 12, no. 2, pp. 255–278, Feb. 2017, doi: 10.1038/nprot.2016.169.

[49] S. Vajda *et al.*, “New additions to the C lus P ro server motivated by CAPRI,” *Proteins Struct. Funct. Bioinforma.*, vol. 85, no. 3, pp. 435–444, Mar. 2017, doi: 10.1002/prot.25219.

[50] I. T. Desta, K. A. Porter, B. Xia, D. Kozakov, and S. Vajda, “Performance and Its Limits in Rigid Body Protein-Protein Docking,” *Structure*, vol. 28, no. 9, pp. 1071-1081.e3, Sep. 2020, doi: 10.1016/j.str.2020.06.006.

[51] G. Jones *et al.*, “Elucidation of protein function using computational docking and hotspot analysis by *ClusPro* and *FTMap*,” *Acta Crystallogr. Sect. Struct. Biol.*, vol. 78, no. 6, pp. 690–697, Jun. 2022, doi: 10.1107/S2059798322002741.

[52] R. Brenke *et al.*, “Application of asymmetric statistical potentials to antibody–protein docking,” *Bioinformatics*, vol. 28, no. 20, pp. 2608–2614, Oct. 2012, doi: 10.1093/bioinformatics/bts493.

[53] A. Vangone and A. M. Bonvin, “Contacts-based prediction of binding affinity in protein–protein complexes,” *eLife*, vol. 4, p. e07454, Jul. 2015, doi: 10.7554/eLife.07454.

[54] L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, “PRODIGY: a web server for predicting the binding affinity of protein–protein complexes,” *Bioinformatics*, vol. 32, no. 23, pp. 3676–3678, Dec. 2016, doi: 10.1093/bioinformatics/btw514.

[55] K. Olechnovič and Č. Venclovas, “VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W437–W442, Jul. 2019, doi: 10.1093/nar/gkz367.

[56] K. Olechnovič and Č. Venclovas, “VoroMQA: Assessment of protein structure quality using interatomic contact areas,” *Proteins Struct. Funct. Bioinforma.*, vol. 85, no. 6, pp. 1131–1145, Jun. 2017, doi: 10.1002/prot.25278.