Department of Electronic and Telecommunication Engineering
University of Moratuwa

BM 4200 - Research Project

Deep Geometric Framework to Predict Antibody-Antigen Binding Affinity

Name	Index Number
P.M.N.S. Bandara	180066F
S.M. Chandanayake	180085L
S.S. Hettiarachchi	180237G
H.D.M. Prematilaka	180497C

Supervisors : **Dr. Subodha Charles**
Department of Electronic & Telecommunication Engineering,
University of Moratuwa.

Dr. Aravinda Munasinghe
Pfizer Incorporation,
USA.

Dr. Kaushalya Madhawa
Lily MedTech Incorporation,
Japan.

This report is submitted in partial fulfillment of the requirements for the module BM 4200 - Research Project

July 23, 2023

Approval of the Department of Electronic & Telecommunications Engineering

3 Oct. 2023

Head, Department of Electronic &
Telecommunication Engineering

This is to certify that I/we have read this project and that in my/our opinion it is fully adequate, in scope and quality, as an Undergraduate Graduation Project.

Supervisor: Dr. Subodha Charles

Signature: .....

Date: .....

Supervisor: Dr. Aravinda Munasinghe

Signature: Aravinda

Date: 03/10/2023

Supervisor: Dr. Kaushalya Madhawa

Signature: .....

Date: 2023/10/03

Declaration

This declaration is made on July 09, 2023.

Declaration by the Group Members

We declare that the dissertation entitled *Deep Geometric Framework to Predict Antibody-Antigen Binding Afffinity* and the work presented herein are our own. We confirm that:

- The study was conducted in candidature for the B.Sc. in Engineering (Hons) degree at the University of Moratuwa,
- where any part of this dissertation has previously been submitted for a degree or any other qualification at this university or any other institute, has been clearly stated,
- where we have consulted the published work of others, is always clearly attributed,
- where we have quoted from the work of others, the source is always given.
- with the exception of such quotations, this dissertation is entirely our own work,
- we have acknowledged all main sources of references,
- parts of this dissertation have been published.

3rd October 2023

.....

Date

.....

Bandara P.M.N.S. (180066F)

.....

Chandanayake S.M. (180085L)

.....

Hettiarachchi S.S. (180237G)

.....

Premathilaka H.D.M. (180497C)

Declaration by Supervisor

I/We have supervised and accepted this dissertation for the submission of the degree.

....

Dr. Subodha Charles Date

Aravinda 03/10/2023

....

Dr. Aravinda Munasinghe Date

2023/10/03

....

Dr. Kaushalya Madhawa Date

Abstract

DEEP GEOMETRIC FRAMEWORK TO PREDICT ANTIBODY-ANTIGEN BINDING AFFINITY

Groups Members : Bandara P.M.N.S, Chandanayake S.M., Hettiarachchi S.S., Prematilaka H.D.M.
Supervisors : Dr. Subodha Charles, Dr. Aravinda Munasinghe, Dr. Kaushalya Madhawa

In the field of drug development, molecules could be categorized as small molecules which are synthetic chemicals with simple structures and large molecules (biologics) which are complex proteins often produced by living organisms. Biologics are extremely target specific, unlike small molecules which implies that biologics result in fewer side effects compared to small molecules. The efficacy of a drug depends on the extent to which the constituent molecules interact with the target molecules. Thus, the strengths of those interactions must be evaluated during the drug design phase to achieve the desired efficacy levels. In literature, such protein-protein interactions are reflected by the binding affinity which is a qualitative parameter. The corresponding quantitative parameter is referred to as the binding energy. Therefore, accurate binding energy prediction is critical in designing drugs with higher affinity and specificity towards their target. Our study focused on Antibody (Ab)-Antigen (Ag) binding, which is a subclass of proteins. Currently, techniques such as Molecular Docking and Molecular Dynamics (MD) simulation are employed to determine the binding affinity at different binding poses. Molecular docking overlooks the temporal behaviour and hence, could be less accurate. On the other hand, MD simulations are accurate but computationally expensive and time-consuming in general, and these complexities rise exponentially with the number of atoms in the molecules.

Due to the dependency of MD on the number of atoms in the proteins of concern, researchers are now focusing on bypassing the MD simulations with Machine Learning (ML) models that could handle relatively large molecules such as Ab and Ag with less computational cost. However, the predictive performance of existing ML methods when calculating binding affinity is highly dependent on the quality of the Ab-Ag structures and they tend to overlook the importance of capturing the evolutionary details of proteins upon mutation. To overcome the said complexities and drawbacks, we developed a novel deep geometric network that consists of a geometric model that could process the 3D structures of the input proteins and a sequence model that could handle the amino acid sequences of the input proteins. We employed attention mechanisms in both models to ensure that atomistic level information as well as evolutionary information are incorporated into neighbour embeddings.

The proposed model was trained on a combined dataset which consists of multiple antigens and antibodies corresponding to common viruses such as human immunodeficiency virus (HIV), coronavirus disease (SARS-CoV-2), etc. to ensure sufficient generalizability and it was observed that the proposed model architecture surpassed the state of the art models by over 10% in terms of the mean absolute error.

DEDICATION

To the people who devote their time to the betterment of future generations

Acknowledgements

We extend our gratitude to Dr. Subodha Charles and Dr. Aravinda Munasinghe for their valuable guidance and support without which we would not have been able to achieve the milestones we reached over the last ten months. We are grateful to Dr. Aravinda Munasinghe for dedicating precious time from his tight schedule to teach us the fundamentals associated with antibodies and antigens. Moreover, the encouragement, flexibility and optimistic vibe Dr. Subodha Charles provided allowed the team members to bring out their best performance to this project. We were able to apply the theoretical knowledge acquired over the undergraduate years to an exciting field of research, which gives us humble satisfaction.

We would further like to mention Dr. Kaushalya Madhawa from Lily MedTech, Japan for sharing with the group members his valuable knowledge and experience on graph neural networks. Moreover, we would like to express our heartfelt gratitude to Mr. Vithushan Veranthirajah from the University of Colombo, who helped the team members clarify chemistry-related questions throughout the project and played a pivotal role in developing the protein structure generation pipeline.

In addition, we are grateful to Dr. Ranga Rodrigo, Head of the Department of Electronic and Telecommunication Engineering for allocating us GPUs which significantly fastened the protein structure generation process and training the deep learning models. We would also like to extend our gratitude to other Senior Lecturers for the valuable feedback given during feasibility, mid-review and final presentations.

Last but not least, we are grateful to all staff members, fellow colleagues, and family members for the continuous support given over the course of the project.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Ab	Antibody.
Ag	Antigen.
BLOSUM	Block Substitution Matrix.
CNN	Convolutional Neural Networks.
CPI	Compound-Protein Interaction.
DL	Deep Learning.
GAT	Graph Attention Networks.
GCN	Graph Convolutional Network.
HIV	Human Immuno-deficiency Virus.
IC50	Half-Maximal Inhibitory Concentration.
MAE	Mean Absolute Error.
MD	Molecular Dynamics.
MERS	Middle Eastern Respiratory Syndrome.
ML	Machine Learning.
mmCIF	Macromolecular Crystallographic Information File.
MSA	Multiple Sequence Alignments.
MSE	Mean Squared Error.
PDB	Protein Data Bank.
PPI	Protein-Protein Interactions.
SARS	Severe Acute Respiratory Syndrome.
SPM	Single Point Mutations.
SVM	Support Vector Machine.

Chapter 1

Introduction

The binding affinity and binding energy are important parameters in the context of drug design and development as they are direct indicators of how well two proteins bind to one another. In this chapter, we will delve into the details associated with the binding affinity of proteins, current techniques employed to quantify the antibody-antigen binding affinity, and their limitations. Further, we will discuss in detail, the objectives and scope of the proposed project.

1.1 Molecules in Drug Development

In the field of drug development, molecules can be categorized as large and small molecules, based not only on their size but also on how they are synthesized (in vitro/ in vivo), mode of action, binding sites, transportation mechanisms, etc. Small molecules are stable, synthetic chemicals with relatively simple structures. These have been around for decades and occupy more than 90% of the drugs currently in use [7]. Small molecules can be absorbed by the body through oral uptake. Common small molecule drugs include "medicine cabinet" drugs such as aspirin. Large molecules (also known as biologics) have complex structures and majorly consist of proteins produced by living cells. Their production processes are complicated and time-consuming. Examples of biologics in use include vaccines, blood/blood components, etc. and these are usually administered through injections.

The first biologic in the market was recombinant human insulin which was made available to the general public 100 years after its discovery. At present, the class of biologics comprises monoclonal antibodies, plasma proteins, recombinant proteins, etc. Recently, there has been a rise in research to check the possibility of using biologics in treating conditions such as cancer, Huntington's disease, diabetes, etc., owing to the specificity of the biologics. The other advantage of biologics over small molecules is their ability to trigger certain biological functions once administered into the body of the patient and this characteristic is important in treating the aforementioned conditions.

Biologics have high specificity in destination targeting, whereas small molecules may bind to off-targets and induce non-target harmful effects/side effects. Eight out of ten global best-selling drugs being biologics in 2018 indicates the increasing significance of biologics in the field of pharmaceuticals [8]. With advancements in recombinant technology, pharmaceutical companies have increased their investments in biologics-related research. In the midst of emphasizing the value of biologics, it is also important to state that small molecules are as nearly important. In fact, in 2015, Dr. Tu You won the Nobel Prize in medicine for discovering a small molecule drug for Malaria. Currently, large molecules (biologics) as well as small molecules are regulated by the Food and Drug Administration and biologics

are subjected to further scrutiny by the Center for Biologics Evaluation and Research in the United States of America.

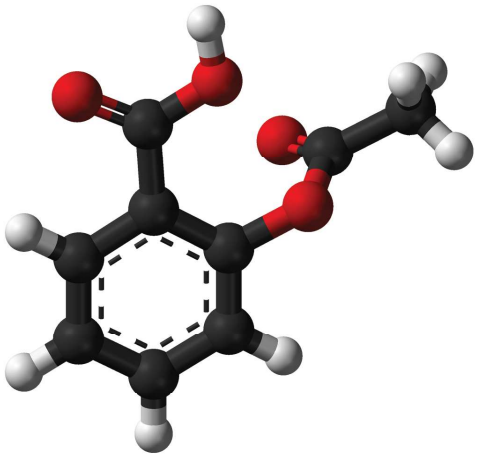

Figure 1.1.1: Small Molecule - Aspirin [1]

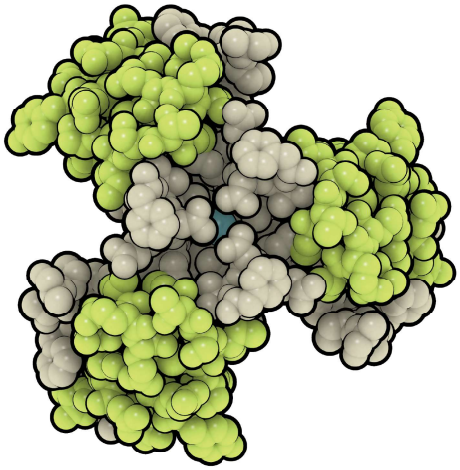

Figure 1.1.2: Large Molecule (Biologic) - Insulin [2]

1.2 Antibody-Antigen Interactions and Binding Affinity

In simpler terms, antigens are typically foreign substances that would induce an immune response in your body, whereas antibodies are a part of the immune response, produced to fight off such antigens. Antibodies (Ab) are large protective proteins produced by the immune system to identify and neutralize foreign objects such as pathogenic bacteria and viruses. The antibody recognizes a unique molecule of the pathogen called an antigen (Ag). Antibody molecules are roughly Y-shaped molecules and each tip of an antibody contains a paratope (analogous to a lock) that is specific for one particular epitope (analogous to a key) on an antigen, allowing these two structures to bind together with precision. This is one key step in neutralizing foreign objects. Therefore, Ab-Ag binding is one of the most essential protein-protein bindings and plays a unique role in the study of drug design. Accordingly, it is imperative to determine the extent of interactions between an Ab-Ag pair which in return would decide the suitability of the drug in subsiding the relevant pathogenic condition.

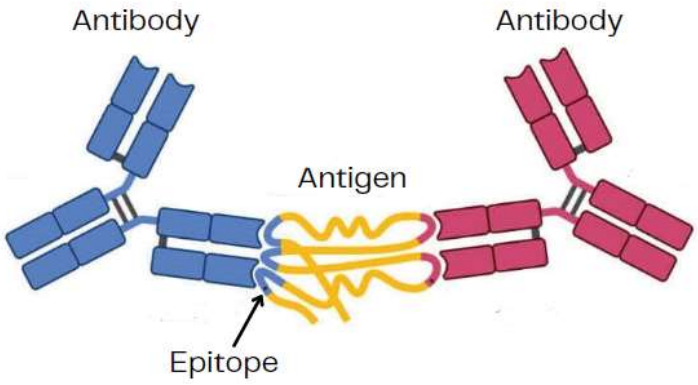

Figure 1.2.1: Antibody-Antigen Pairing

The strength of the Ab-Ag interactions is reflected by the binding affinity, and the free energy associated with the binding of two molecules is called the binding energy. As the name suggests, the binding energy is the energy released upon the binding of the two molecules. Accordingly, the higher the binding energy, the higher will be the binding affinity. Therefore, accurate binding energy prediction is a helpful tool for designing drugs with higher affinity and specificity toward their target. In addition to the binding energy, there are other quantitative parameters that could be used to evaluate the binding affinity. Some of them are dissociation constant, IC50, EC50 values, etc. The dissociation constant indicates the tendency of the Ab-Ag pair to detach from one another and the IC50 value measures the half-maximal inhibitory concentration (IC50) which in simple terms, is the concentration of the antibody required to reduce the activity of a sample of the antigen by 50%.

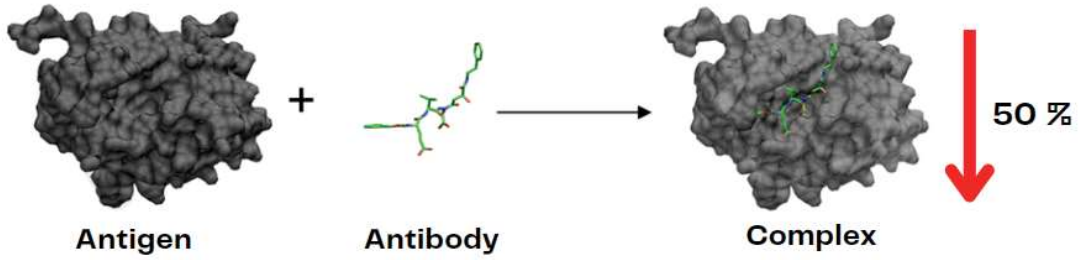

Figure 1.2.2: 50% Reduction in antigen activity upon binding with the antibody

1.3 Existing Simulation Techniques

The efficacy of drugs designed using small molecules and biologics depends on how well they can bind/interact with the target molecule(s). Thus, the strengths of those interactions must be evaluated during the drug design phase to achieve the desired efficacy levels. Currently, a technique known as 'Molecular Docking' is employed to study how two or more molecular structures fit together. Docking is a molecular modelling technique that is used to predict how molecular structures interact using computer simulations. Modelling interactions between two molecules is complicated as this involves a range of forces. To produce a stable bond between molecular structures, the binding naturally happens via the lowest energy pathway. The aim of molecular docking is to mimic these natural interactions that take place during the binding via the lowest energy pathway. To achieve this, molecular docking calculates affinities of multiple poses of binding before deciding the optimum binding poses. But, these decisions can overlook some predicted poses and may be less accurate. Hence, a more extensive evaluation is required.

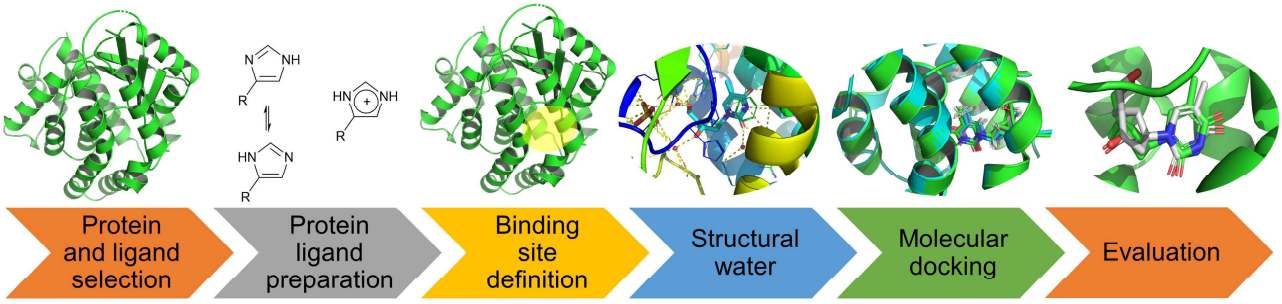

Figure 1.3.1: Basic steps involved in Molecular Docking

Any molecule has an inherent motion, and these motions play a major role in the context of antibody-antigen interactions. However, this critical factor is overlooked during molecular docking

simulations. On the other hand, Molecular Dynamics (MD) simulation is a more sophisticated simulation technique that extends the capabilities of molecular docking by including the temporal behaviour of the structures of concern.

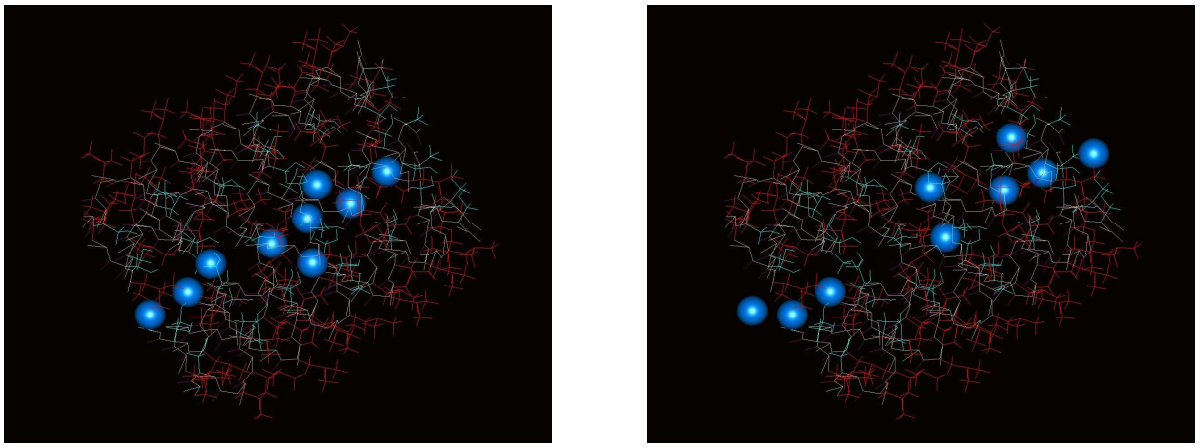

Figure 1.3.2: Molecular Dynamics simulation of ion propagation through a protein-ion channel [3]

MD simulation calculates the Coulomb force on each atom at a given instance and then determines the new position, velocity, and acceleration of the atoms in the next time step using Newton's kinematics equations. Thus it is clear that MD simulations involve heaps of statistical and mechanical calculations for each atom of the molecule at a given time step which poses computational challenges.

1.4 Problem Statement

Due to the over-dependency of MD on the number of atoms in the proteins of concern, conducting MD simulations for large molecules remains a daunting task even to this date. With recent advancements in deep learning (DL), researchers are now focusing on bypassing the MD simulations with ML models that could handle relatively large molecules such as Ab and Ag, with less computational cost. However, the predictive performance of existing ML methods when calculating binding affinity is highly dependent on the quality and resolution of the structure of the Ab-Ag complex. Further, characterizing biologics in these applications requires multiple analytical tools, and even with recent advancements, the structures cannot be completely defined. Therefore, in this work, we intended to develop a novel deep-learning pipeline to predict the Ab-Ag binding energy without requiring MD simulations or over-relying on conventional structure modelling.

1.5 High-Level Block Diagram

The main aspects of the study are a well-curated dataset, a deep-learning-based geometric and sequence model that processes input proteins, and a web platform that the community can use to process their own inputs. Accordingly, any end-user could upload the PDB files for the antigen and the antigen via the web platform. Then, the model pipeline would process the input files and yield the IC50 value that would indicate the extent of binding affinity between the Ab-Ag pair.

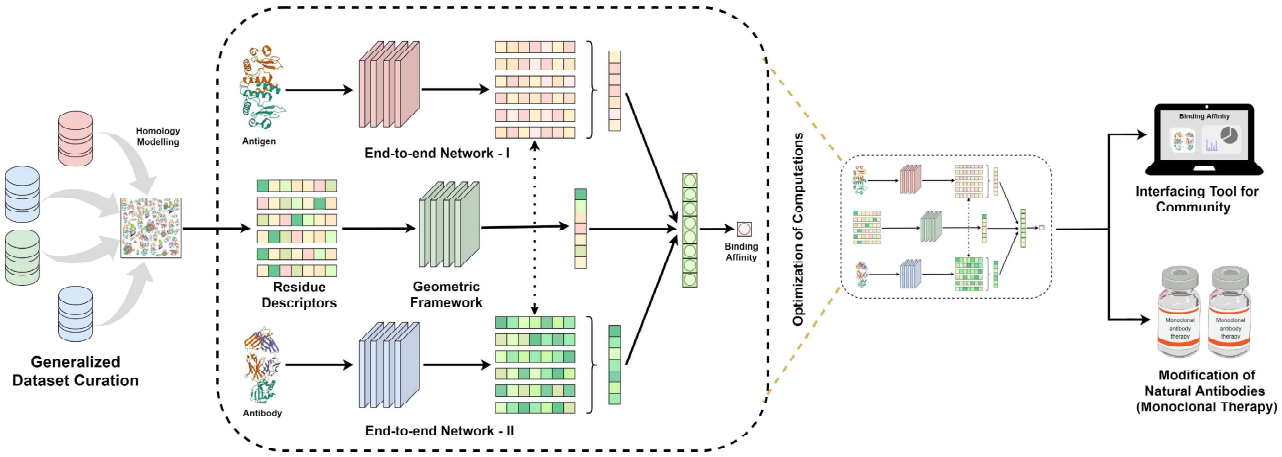

Figure 1.5.1: High-level block diagram

1.6 Objectives

1. **Curate a generalized dataset for protein representation**
   In the current context, many publicly available datasets contain data points corresponding to a specific class of antigen or antibody. However, as the expectation is to develop a generalized deep-learning model, it is important to expose the model to a variety of antibodies and antigens during training. Accordingly, one of the primary objectives was to curate a generalized dataset for public access.

2. **Design a predictive framework for antibody-antigen binding affinity**
   As mentioned in the problem statement, it is important to design and develop a robust, accurate deep learning-based model to predict the binding affinity of a given Ab-Ag pair in order to bypass the complexities associated with MD simulations. Further, the performance of the model should be reliable enough to be accepted by the community.

3. **Extend the model towards antibody therapy**
   Once the model performance is validated, the proposed pipeline could be extended toward antibody therapy which is widely used in treating conditions such as cancer, autoimmune diseases, etc.

4. **Publish the work in a well-known forum**
   We expect to conclude the study by publishing our work in a reputed forum/ journal/ conference with the intention of popularizing our study as well as the web platform.

1.7 Scope

1. Proposing a novel framework to predict the pico-level antibody-antigen binding affinity, which does not rely upon molecular dynamics simulations for structure modelling.

2. Curating a dataset for generalized representation of proteins and their key relationships

3. Exploring and presenting novel amino-acid descriptors to embed evolutionary details into the predictive framework.

4. Optimizing the computations for lesser inference times for real-world implementations

5. Applying the framework to the modification of natural antibodies to adapt them for targeted use cases: Towards antibody therapy
6. Developing a user-ergonomic interfacing tool for easy community access

In the subsequent sections, we will delve into the details of existing work, our methodology, results, and discussion.

Chapter 2

Literature Review

In this chapter, we explore the related work that focuses on deep learning-based antigen-antibody binding affinity prediction models, available curated datasets, and molecular descriptors for deep learning models.

2.1 Existing Databasees

Dunbar et al. [9] introduced SAbDab, a structural antibody database, which provides annotations for various properties such as accurate heavy and light chain combinations, antigen information, and the affinity of antigen-antibody binding.

Following this, Sarah et al. [10] presented AB-Bind, a database containing mutants and their experimentally determined changes in binding free energy. Yoon et al. [11] introduced CATNAP, an analysis tool based on the Los Alamos HIV Database, designed to address the latest advancements in HIV-neutralizing antibody research. This platform specifically focuses on neutralizing antibody potencies in conjunction with viral sequences. Recently, Rawat et al.[12] manually curated interaction profiles of 1780 coronavirus-related neutralizing antibodies into a database named Ab-Cov.

Engelhart et al. [13] introduced the Alphaseq dataset that comprises quantitative binding scores of scFv-format antibodies against a SARS-CoV-2 target peptide, aiding machine learning model development. This dataset was the largest publicly available dataset that includes an antibody, antigen sequences and quantitative measurement of binding scores.

Jankauskaite et al. [14] presented SKEMPI 2.0, an updated database of protein-protein interaction binding free energy changes upon mutation. It includes 7085 mutations, with expanded data on kinetics, enthalpy, entropy changes, and 440 mutations that abolish detectable binding.

2.2 Molecular Descriptors for Deep Learning Models

Various studies have employed molecular descriptors in their predictive models. One such study by Simone et al. [15] combined atomistic modelling and machine learning to estimate binding affinity. Their approach incorporated atomic descriptors, protein-protein scoring functions, protein stability scoring functions, and entropy models as features.

In subsequent work, Magar et al. [16] introduced a predictive framework that utilized atomistic descriptors, taking into account atoms' type, valency, and hybridization state, to predict neutralizing antibodies for SARS-CoV-2. A more recent direction in this research field was led by Junjie et al. [17], who employed statistical and combinatorial properties of spectrum information from Hodge Laplacian matrices for predicting protein-to-protein interaction binding affinity.

2.3 Deep Learning-based Predictive Frameworks

The demand for computational methods that predict interface contacts between proteins is growing rapidly, as they offer substantial enhancements over conventional techniques like molecular docking and protein function analysis tools. These advanced methods provide more accurate and efficient means of identifying the interactions between proteins at their interfaces, leading to a deeper understanding of their functions and potential applications in various fields of research. Despite the fact that many computational techniques have been suggested, none of them was accurate enough. In order to improve the accuracy, Kurumida et al. [18] developed a novel method by merging various machine learning-based predictors.

Following such previous works, Wang et al. [19] introduced a novel approach involving a topology-based network tree, which comprises a topology-based feature extraction stage and a convolutional neural network-based gradient boosting tree model. This innovative method aims to predict affinity changes, but these models lack generalizability due to their absolute dependence on the defined feature space.

Expanding the boundaries of research in this domain, Shan et al. [20] proposed a geometric attention network capable of generating geometric embeddings for both wild-type and mutation complexes of SARS-CoV-2 variants. Identifying key residue pairs near the protein interface that contribute to binding affinity sheds light on crucial interactions. To predict the effect of mutations on binding affinity, they cleverly incorporated a multi-layer perceptron. However, a critical limitation of their work is the inability to capture evolutionary details solely through 3D structures fully.

In our discussion, we will explore various machine-learning approaches used in previous research, including conventional ML and deep learning, for predicting the binding affinities of antigen-antibody interactions. We will also aim to devise a better solution that aligns with our research purpose.

2.3.1 Conventional Machine Learning-based Approaches

Abisi et al. [21] proposed a sequence-based method called ISLAND which is a sequence-only support vector regression model, for predicting antibody-antigen binding affinity. In this approach, the Smith-Waterman alignment method is utilized with the BLOSUM-62 substitution matrix and gap opening and extension penalties set at -11 and -1, respectively, to assess the level of homology between two protein complexes. Molina et al. [22] introduces PPI-Affinity, a tool that uses support vector machine (SVM) predictors to screen protein-protein and protein-peptide interactions datasets and rank mutants of a given protein structure. Existing predictors for small molecules' binding affinity are commonly used for peptides, but protein-protein interaction (PPI)-Affinity addresses this by specifically considering the complexity and heterogeneity of peptide interactions.

When it comes to the support vector regression function, which is a conventional machine-learning technique, has several drawbacks associated with respect to the task at hand. One major

limitation is that they are data-dependent, which means obtaining a generalized model is challenging. Additionally, this model does not effectively extract parallel information, such as antigen sequence information and antibody sequence information, even if they are concatenated during the information extraction process.

2.3.2 CNN-based Approaches

Convolutional Neural Network (CNN) is often considered better than SVM as a generalization model in binding affinity prediction. In this context, CNN exhibits superior performance due to its ability to automatically learn hierarchical features from raw input data, such as protein sequences or molecular structures. The CNN architecture allows it to identify and capture patterns at different levels, which is particularly advantageous when dealing with complex and high-dimensional data like protein-protein interactions. On the other hand, SVM, being a traditional machine learning technique, heavily relies on predefined features and may struggle to generalize well to unseen data. It often requires careful feature engineering to achieve good results, which can be laborious and time-consuming.

JunJie Wee and Kelin Xia propose PerSpect-EL [17], a novel approach for PPI binding affinity prediction. PerSpect-EL utilizes persistent spectral-based PPI representation and featurization, generating Hodge Laplacian matrices from a filtration process. The PerSpect attributes, capturing statistical and combinatorial properties of the spectrum information, are fed into a 1D CNN. These CNN networks are then stacked together to form the ensemble.

Muhao et al. [23] proposed a novel end-to-end framework, Protein-Protein Interaction Prediction Based on Siamese Residual RCNN (PIPR), for predicting PPI based on amino acid sequences. In this procedure antibody and antigen sequences are sent separately through two residual RCNNs and sequence embeddings are combined in element-wise multiplication. Binding affinity is obtained with previous results and regression model. The framework utilizes a residual RCNN within a Siamese learning architecture to effectively capture local and sequential features from primary protein sequences, enabling it to address various PPI prediction tasks without predefined features.

However, CNN-based approaches do suffer from their localized convolution operation and thereby, struggle to learn long-range dependencies and connections effectively. In addition, RNNs do lack the capability for extracting long-range evolutionary details from protein amino-acid sequences while suffering from the essential need for sequential processing.

2.3.3 Graph Learning-based Approaches

Puents et al. [24] proposed PLA-Net, a deep-learning approach for predicting target–ligand interactions in drug discovery. The method combines ligands' and targets' chemical information using a two-module deep graph convolutional network. Adversarial data augmentations are also employed to improve model interpretability and highlight relevant substructures each comprising a deep graph convolutional network (GCN) followed by an average pooling layer. These modules extract relevant features from their respective input graphs. The extracted representations are then concatenated and combined using a fully connected layer to predict the probability of target–ligand interactions.

Liu et al. [25] introduced a novel structure-based deep-learning framework, called GeoPPI, that predicts the impact of amino acid mutations on protein-protein interaction binding affinity. It first learns geometric representations from protein structures via self-supervised learning, which is then used to train gradient-boosting trees for affinity change prediction.

Wang et al. [19] introduced a novel approach called element- and site-specific persistent homology, a branch of algebraic topology, to simplify the structural complexity of protein-protein complexes. The method embeds crucial biological information into topological invariants. Additionally, a new deep learning algorithm called NetTree, combining convolutional neural networks and gradient-boosting trees, is proposed for predicting protein-protein interaction binding affinity changes following mutation.

Morehead et al. [26] introduced DeepInteract, the Geometric Transformer, a geometry-evolving model for protein structures, showcasing its efficacy in predicting residue-residue interactions in protein complexes. The study envisions various potential applications of the Geometric Transformer in protein deep learning, including quaternary structure quality assessment and residue disorder prediction. However, a notable drawback of the current design is its high computational complexity due to the dot product self-attention mechanism. To address this issue, the authors propose exploring efficient alternatives, such as the Nystromform.

Zhang et al. [27] introduced a novel approach to learning effective protein representations using 3D structures. Unlike existing methods that rely on pretraining with unlabeled amino acid sequences, this approach pre trains protein graph encoders using multiview contrastive learning and self-prediction tasks to capture the geometric features of proteins. By leveraging known protein structures, this method shows promise in improving protein property prediction tasks, which are crucial for understanding protein function and structure in biology.

K. Yugandhar and M. Michael Gromiha [28] presented a novel sequence-based method called PPA-Pred for predicting protein-protein binding affinity in the context of functional classification. By analyzing 135 protein-protein complexes and 642 sequence properties, they found that the correlation between binding affinity and sequence features varies with complex type. The proposed approach utilizes regression models specific to different classes, achieving promising correlation (0.739-0.992) in jack-knife tests. The method offers valuable insights into the role of binding site residues in governing binding affinity and can be a valuable tool for analyzing protein-protein interaction networks in specific diseases.

Nguyen et al. [29] introduced a computational model called GraphDTA, which represents drugs as graphs and utilizes graph neural networks and has shown promise in predicting drug–target affinity more accurately than non-deep learning models and other competing deep learning methods. By combining drug molecular graph representations with protein sequence encoding, GraphDTA efficiently estimates drug–target affinity, demonstrating the suitability of deep learning models in this domain and the potential for further advancements using graph-based representations. The authors highlight the three-stage process of the model for predicting drug-target affinities. The model combines drug SMILES code processing through graph representation learning with protein sequence encoding using 1D convolutional layers. By concatenating both representations and passing them through fully connected layers, the architecture accurately estimates drug-target affinity values.

Li et al. introduced BACPI [30], an end-to-end neural network model, for predicting compound-protein interactions (CPIs) and binding affinity. The model combines graph attention networks and convolutional neural networks to learn representations of toms from the compound structure graph and representations of residues from the protein sequence, using a bi-directional attention neural network to integrate these representations. By utilizing these integrated features, BACPI makes predictions for CPI or binding affinity.

Jha et al. proposed a method for predicting PPIs using GCN and graph attention network (GAT) [31]. It incorporates both structural information from protein 3D coordinates and sequence features using a protein language model. The protein graph is constructed from PDB files, representing an amino acid network. The proposed approach is evaluated on two PPI datasets and shows supe-

rior performance compared to previous methods, highlighting its effectiveness in predicting PPIs by leveraging both sequence and structural information. These GCN-based approaches typically struggle with identifying the most related nodes in the graph representation and thereby, smoothen the feature space where it would be beneficial to pay more attention to certain nodes in the graph representation.

Existing works in the literature that are based on graph learning for predicting antibody-antigen binding affinity, as presented above, are highly dependent on the atomistic or residual information encoded through graph representations and thereby overlook the evolutionary details present in the amino-acid sequences. Even if they consider both information streams, they tend to neglect the binding nature of antibodies and antigens and therefore, the need for cross-information sharing between the information flow of antibodies and antigens. Therefore, through our work, we intend to address these limitations in the literature by effectively incorporating both atomistic and evolutionary details of antibodies and antigens while sharing information imitating the binding nature in place.

Chapter 3

Methodology

In this section, we will delve into the details of the dataset curation process, sequence model, structure model, and combined model. We will discuss the mathematical concepts associated with deep-learning models and the justifications for employing specific techniques. In brief, the sequence model processes the amino acid sequence of the input proteins, whereas the structure model processes the 3D structures of the proteins after converting them to graphs.

3.1 Dataset Curation

Since publicly available datasets are tabulated in different formats, the first task associated with curating a generalized dataset is to process the datasets in a way that all the data points are identically formatted. In addition, as the models require the 3D structure of the proteins, suitable measures had to be taken to generate the 3D structures that were unavailable in the public datasets. Homology modeling and Google DeepMind's AlphaFoldV2 were employed in this regard. Before discussing the steps involved in dataset curation, we will first go through the summary of datasets.

Dataset	Datapoints	Mutations	Data Type	Numerical Value
AB-Bind	1 101	Available	Sequences	$\Delta\Delta G$
AB-Cov	1 964	Available	Structures	IC50, EC50
CATNAP	129 686	N/A	Names only	IC50, IC80, ID50
SAbDab	1 327	Available	Structures	$\Delta\Delta G$, Affinity
SKEMPI	7 086	Available	Structures	Affinity
AlphaSeq	1 259 700	N/A	Sequences	IC50

Table 3.1: Summary of the used publicly available datasets

3.1.1 Datasets and Their Curation Pipelines

AB-Bind

Antibody-Bind (AB-Bind) database [32] includes 1101 mutants with experimentally determined changes in binding free energies ($\Delta\Delta G$) across 32 complexes. Protein variants in the dataset can be grouped into those with only single-point mutations (SPM) and non-SPM (multiple mutations per variant) categories, and contain 645 and 466 variants in each category, respectively.

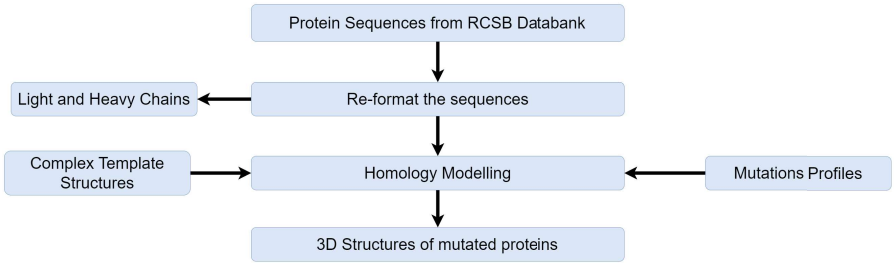

Figure 3.1.1: Developed specific pipeline for data curation using Ab-Bind dataset

Ab-CoV

Ab-CoV [33] contains manually curated experimental interaction profiles of 1780 coronavirus-related neutralizing antibodies. It contains more than 3200 data points on half-maximal inhibitory concentration, half-maximal effective concentration, and binding affinity. Further, the dataset contains the protein structures that are complemented with stability and binding affinity.

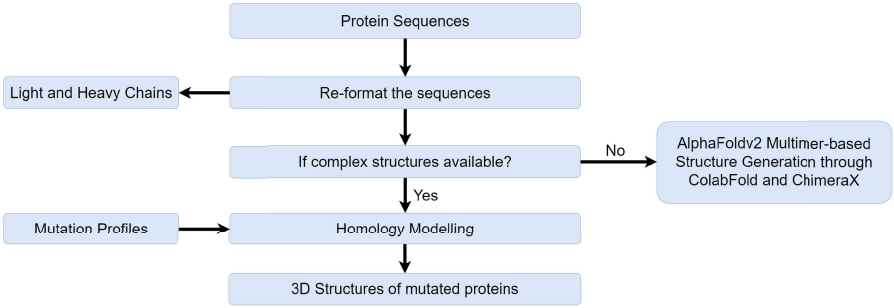

Figure 3.1.2: Developed specific pipeline for data curation using Ab-CoV dataset

CATNAP

CATNAP [34] includes neutralization data from published studies and the current collection comprises 172 antibodies and 722 HIV-1 viruses. The neutralization panel data includes in terms of IC50 and IC80 values for specific monoclonal antibodies and pseudo-typed viruses that were collected from 49 published neutralization studies.

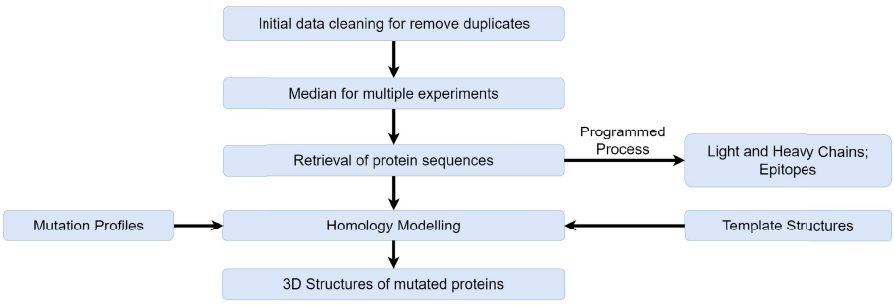

Figure 3.1.3: Developed specific pipeline for data curation using CATNAP dataset

SAbDab

SAbDab [35] contains all the antibody structures available in the protein data bank, annotated and presented in a consistent fashion. Each structure is annotated with a number of properties including experimental details, antibody nomenclature (e.g. heavy-light pairings), curated affinity data, and sequence annotations.

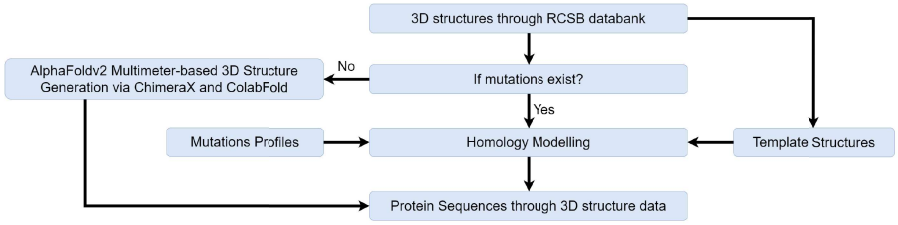

Figure 3.1.4: Developed specific pipeline for data curation using SAbDab dataset

SKEMPI

The SKEMPI database [36] contains data on the changes in thermodynamic parameters and kinetic rate constants upon mutation, for protein-protein interactions including manually curated binding data for 7085 mutations, and changes in kinetics for 1844 mutations, enthalpy, and entropy changes for 443 mutations, and 440 mutations, which abolish detectable binding.

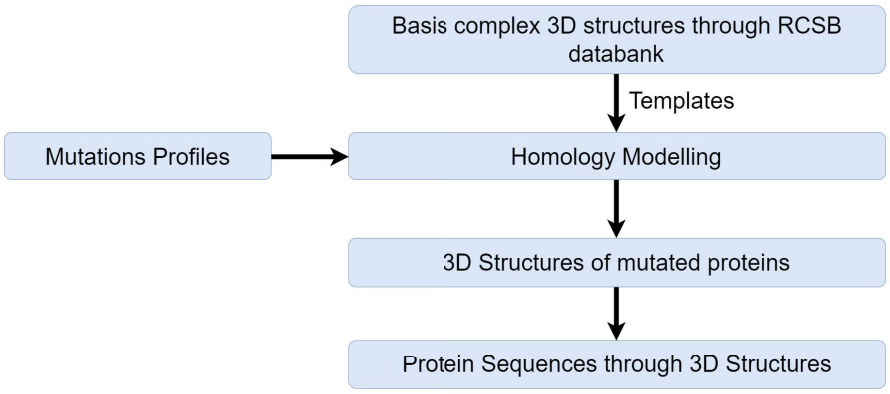

Figure 3.1.5: Developed specific pipeline for data curation using SKEMPI dataset

AlphaSeq

The dataset [37] contains quantitative binding scores of scFv-format antibodies against a SARS-CoV-2 target peptide collected via an AlphaSeq assay. Four sets of 29,900 antibodies were designed in silico by creating all k=1 mutations and random k=2 and k=3 mutations throughout the complementary-determining regions. Of the 119,600 designs, 104,972 were successfully built into the AlphaSeq library and target binding was subsequently measured with 71,384 designs resulting in a predicted affinity value for at least one of the triplicate measurements.

3.1.2 Generalized Dataset Curation Process

Generally, experimental uncertainties can produce multiple IC50 values for a given Ab-Ag pair over several trials. To negate the impact of outliers, we first had to take the median value of the provided IC50 values for repeated entries of Ab-Ag pairs. Based on the availability of the template structure and the mutation profile, we then decide whether to use Homology modelling or AlphaFoldV2. If the mutation profile along with the template structure is provided, then we opt for Homology modelling. If either of the two requirements is not given, then we use the AlphaFoldV2 pipeline. The following flowchart indicates the processing steps associated with the dataset curation.

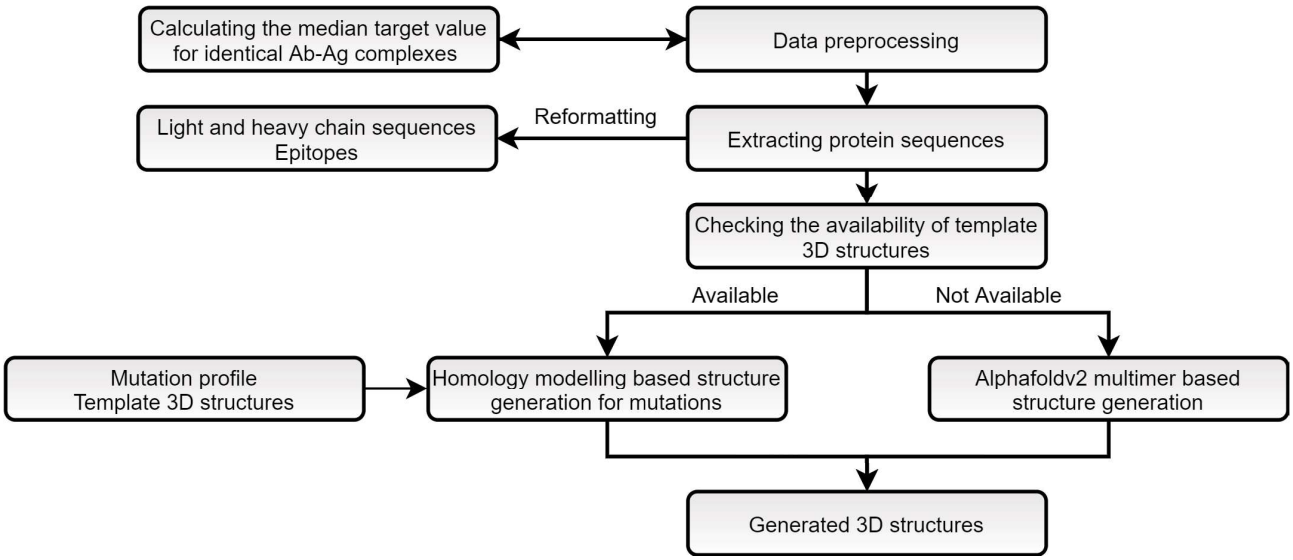

Figure 3.1.6: The generalized flowchart describing the steps of dataset curation for all datasets

3.1.3 Homology Modelling

In order to utilize homology modeling, it is essential to have the 3D structure of the reference protein which serves as the template structure, and a mutation profile indicating the mutant, type of mutation, and the location of the mutation that would be introduced to the reference protein.

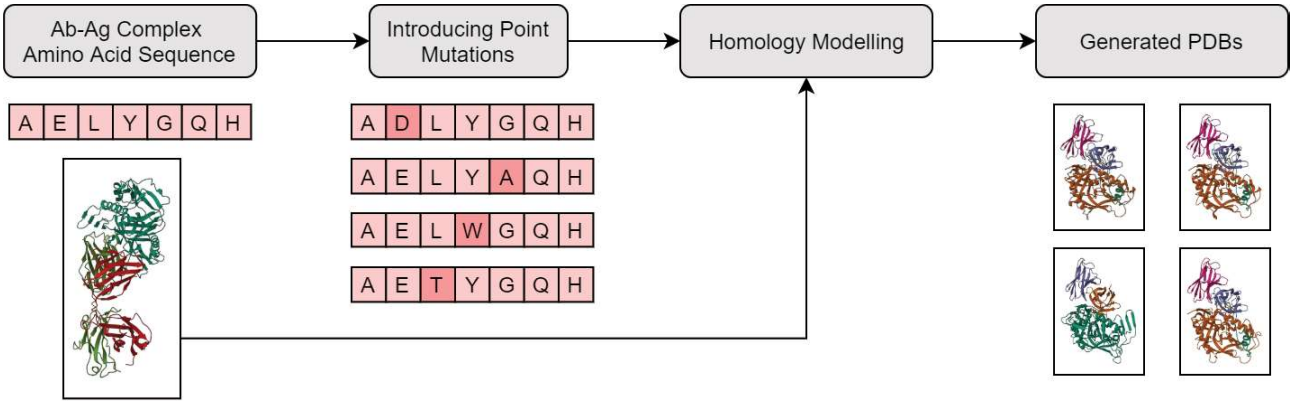

Figure 3.1.7: Basic overview of the homology modelling pipeline

Now, let's briefly go through the steps involved in homology modelling. Once the mutations are introduced using the provided mutation profile, the mutated sequence is aligned with the original

sequence in a way that the correspondence between the two sequences is maximized. In simpler terms, we want as many amino acids in the two sequences to be aligned as possible.

Needleman-Wunsch Algorithm for Sequence Alignment

Needleman-Wunsch alignment which is also referred to as global alignment is a popular technique used in sequence alignment. This algorithm consists of a penalty/ reward system for gaps(g), matches(m), and mismatches ($-m$). The two sequences would be compared by filling a scoring matrix using the defined penalty/ rewards.

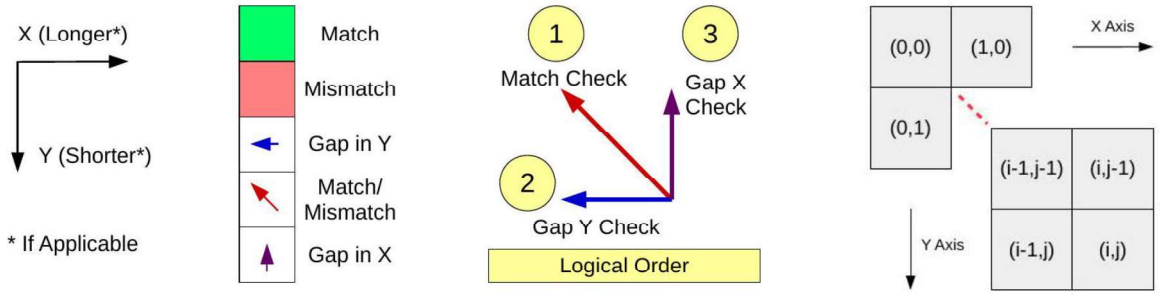

Figure 3.1.8: Notations for filling the substitution matrix

Using the aforementioned notation as a reference, the algorithm is implemented by initializing the first row and first column of the matrix with $-gi$ and $-gj$. Then, the scoring matrix is completed using the following recursive relation.

$$F(i,j) = \max \begin{pmatrix} F(i-1,j-1) \pm m & \text{Match or mismatch} \\ F(i-1,j) - g & \text{Gap in } x \\ F(i,j-1) - g & \text{Gap in } y \end{pmatrix} \tag{3.1}$$

Then, starting from the bottom right vertex, the traceback process is started until the top left vertex is reached. During the traceback process, the next step is taken toward the neighbour entry with the highest value. In the provided example below, we have used a gap of $g = 2$ and a match/mismatch of $m = 1$.

Figure 3.1.9: A simple example for the global alignment algorithm: (a) Initializing the scoring matrix (b) Completing the scoring matrix using Eq. 3.1 (c) Tracing back to derive the alignment

In addition to global alignment, other popular algorithms used to compare sequences are local alignment and semi-global alignment which are not discussed in this report.

Structure Generation

Once the alignment is complete, the basic structure for the mutated sequence will be derived. The first stage in structure derivation is backbone modeling which is responsible for estimating the positions of the amino group, α-Carbon atom, and the Carboxyl group of each amino acid in the polypeptide sequence. Backbone modeling is followed by loop modeling which performs necessary conformational adjustments to the modeled backbone. Using the fact that the torsion angles about the $C_a - C_b$ bond (ψ angle) are nearly similar for homologous proteins, side chains of the mutated sequence are modelled.

Model Optimization and Validation

With the intention of alleviating steric collisions, the estimated relative atomic locations are fine-tuned so as to minimize the potential energy of the conformation of the protein. This is done during the model optimization step. The energy of the protein includes stretching energy, torsion energy, bending energy, and non-bonding interaction energy.

At last, the optimized model needs to be evaluated to see if there is a negative impact owing to possible errors in the template structure and the percentage identity between the target and template sequence. In this context, certain aspects such as the bond angles, bond lengths, chirality, etc. would be used in the evaluation process.

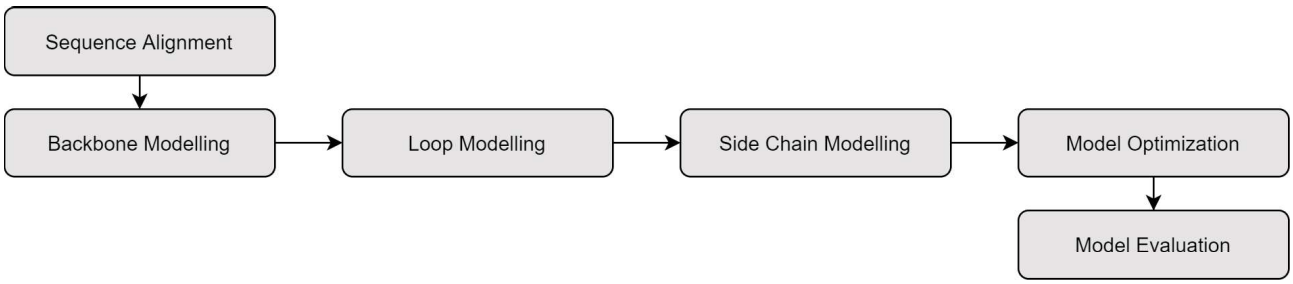

Figure 3.1.10: Key steps of homology modeling

The discussed pipeline was used to generate 3D structures for antibodies/antigens corresponding to the flu virus, HIV virus, etc. Some of the templates that were used during homology modelling are 3NGB, 3ZTJ, 5JW4, 1BRS, etc. (These are PDB codes/tags for the Ab/ Ag/ Ab-Ag complex).

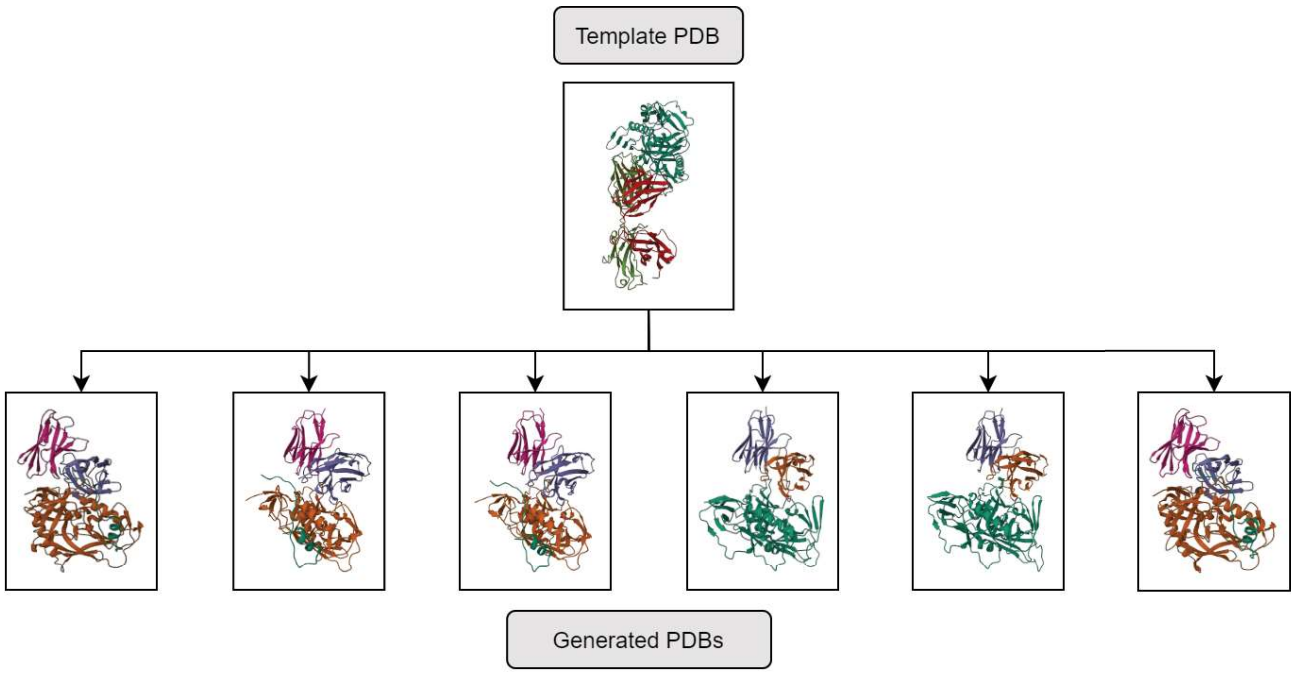

Figure 3.1.11: Sample outputs from the homology modelling pipeline for mutated versions of the VRC01 antibody that neutralizes different variants of HIV

3.1.4 AlphaFold-V2

AlphaFold-V2 multimer model [4] is a state-of-the-art model with atomistic level accuracy for protein structure prediction from amino acid sequences which is proved to be useful even in the absence of a homologous structure as shown by their superior performance at the 14^{th} critical assessment of protein structure prediction.

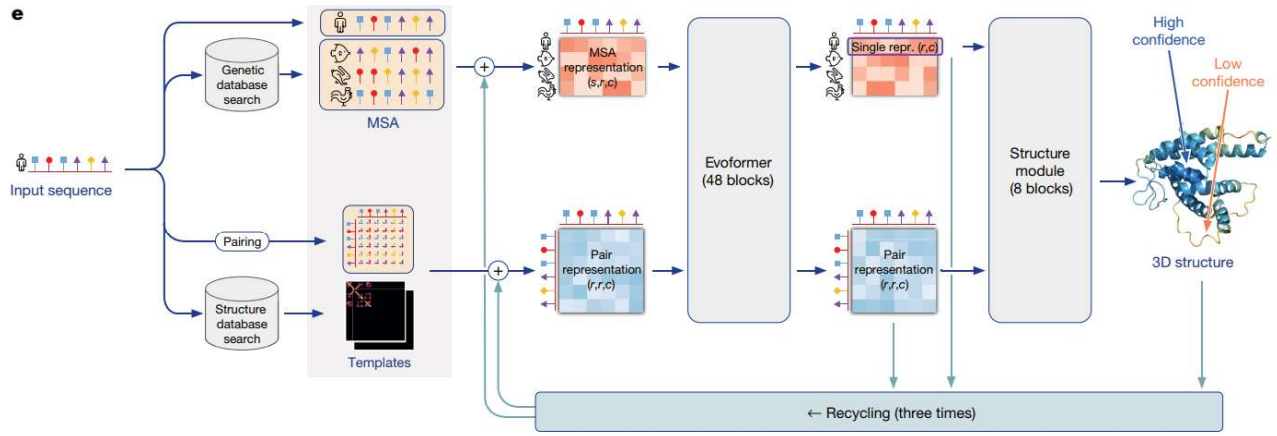

Figure 3.1.12: AlphaFold-V2 model architecture [4]: Here, array shapes are shown in parentheses and s, r, and c refer to the number of sequences, residues, and channels respectively. The input to the model is the protein amino acid sequence and then the input sequence will be fed into the generic and structure database searches to generate multiple sequence alignments (MSA) or find template structures. The resulting MSA or template structures will be fed into interconnected parallel transformer-based Evoformer (for enhanced MSA and pair representation through viewing the prediction problem as a graph inference problem in 3D space in which the edges are residues) and Structure (to generate 3D structure from the input enhanced MSA and pair representations) modules as shown in this figure and finally, the predicted 3D structure could be accessed as the output from the Structure module.

In our implementation for predicting the 3D structures of proteins where the template structures are absent, we followed the pipeline from ColabFold [38] to infer the AlphaFold-V2 with an accelerated combination of MMseqs2-based homology search.

3.2 Sequence-Based Model

Once the input protein files are provided, those files are processed to obtain the FASTA sequences of the proteins. In simpler terms, the FASTA sequences contain the amino-acid sequences of the proteins. Accordingly, each amino acid has a one-letter code and once the input files are processed, we will obtain a letter sequence corresponding to the amino-acid sequence. Since deep learning models are not suited to handle letter representations, the derived FASTA sequences are then encoded using a numerical scheme which will be discussed in the subsequent subsections.

Numerical encodings of the proteins are then fed to the sequence-based model and several self-attention, multi-head attention, and cross-attention layers are incorporated with the intention of passing information from adjacent amino acids from the same molecule or the other counterpart while allowing the model to learn which neighbor amino acids to pay attention to.

The antigen sequence and the antibody sequence are processed via identical pathways and information is cross-propagated through cross-attention layers. Accordingly, there are three pathways in the sequence model, each of which produces one embedding vector. The output embedding vectors represent the antigen, antibody, and antibody-antigen complex. The three embedding vectors which could be considered to be rich representations of the input proteins and the complex are then concatenated into a single feature vector. The vector is further processed to obtain the predicted IC50 value.

During training, mean squared error was used as the loss function to back-propagate the error, to fine-tune the weights of the sequence model. The following diagram summarizes the sequence model pipeline and in the subsequent subsections, we will discuss in detail the FASTA representation, encoding schemes, model layers, and loss function.

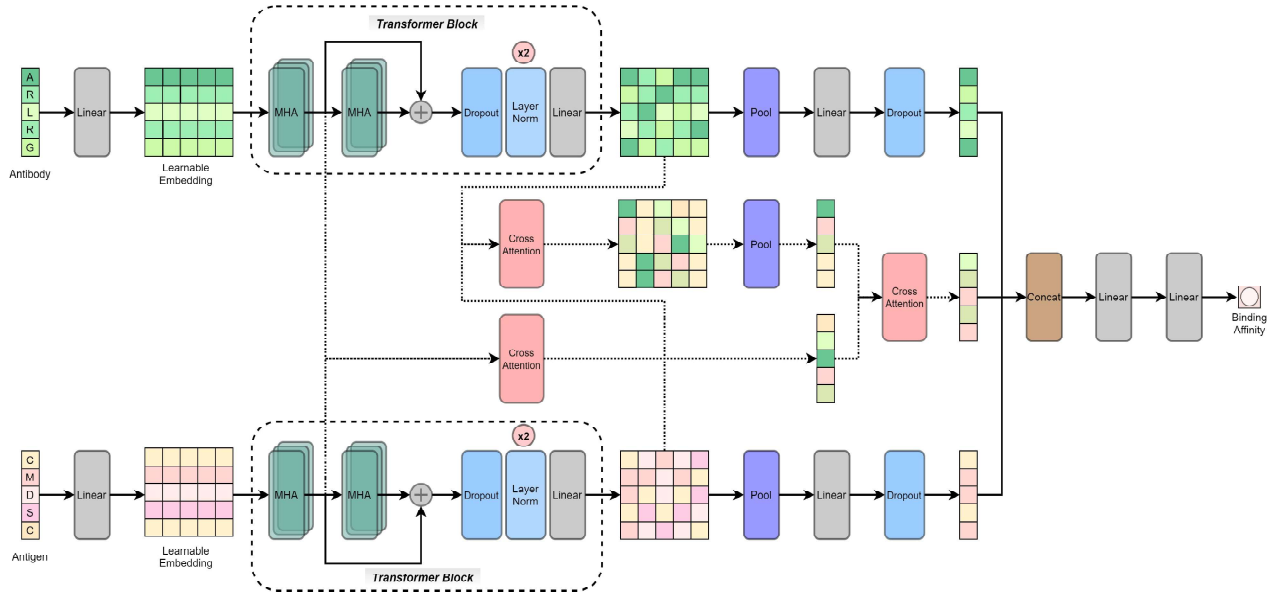

Figure 3.2.1: The network architecture for the sequence-based model

3.2.1 FASTA Format

When we are dealing with proteins, it is convenient to have a text-based format to represent the constituent amino acids. In bioinformatics and computational biology, the FASTA format is often used as the standard for representing amino-acid sequences where each amino acid is assigned a unique single-letter code. As there are 20 fundamental amino acids from which any protein could be made, 20 letters from the English alphabet are used in the FASTA format.

Amino Acid	FASTA Code
Alanine	A
Cysteine	C
Aspartic Acid	D
Glutamic Acid	E
Phenylalanine	F
Glycine	G
Histidine	H
Isoleucine	I
Lysine	K
Leucine	L
Methionine	M
Asparagine	N
Proline	P
Glutamine	Q
Arginine	R
Serine	S
Threonine	T
Valine	V
Tryptophan	W
Tyrosine	Y

Table 3.2: Basic amino acids and their corresponding one-letter FASTA code

3.2.2 Encoding Schemes

One-hot Encoding

In one-hot encoding, each amino acid will be assigned a vector with a dimension equal to the total no. of classes which is in this case, equal to 20 for the 20 amino acids. The vector would be a sparse vector with only one 1 entry at the allocated position for a given amino acid and the remaining elements of the vector would be zeros.

In Fig. 3.2.3, the blue boxes indicate the FASTA representation of the 20 amino acids. Accordingly, if the input protein is composed of 1000 amino acids, the one-hot encoded input protein would be a matrix with order 1000×20. However, the sparsity of the input matrix could impact the performance negatively, based on the context/ application.

A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3.2.2: One hot encoding scheme for the amino acids

VHSE-8 Encoding

VHSE stands for the vector of hydrophobic, steric, and electronic properties. VHSE-8 is a standard, non-sparse encoding scheme for natural amino acids. This encoding scheme contains eight principal components obtained from 50 descriptors representing each amino acid. The 50 descriptors contain 17 hydrophobic features, 18 steric properties, and 15 electronic properties. Furthermore, VHSE8-1 and VHSE8-2 are associated with the hydrophobic properties of the amino acid, and VHSE8-3 and VHSE8-4 describe the steric properties whereas the last 4 components indicate electronic properties.

If the input protein is composed of 1000 amino acids, the VHSE-8 encoded input protein would be a matrix with order 1000×8. Moreover, it is a richer representation of amino acids with lower dimensions compared to the one-hot encoding scheme.

A	0.15	-1.11	-1.35	-0.92	0.02	-0.91	0.36	-0.48
C	0.18	-1.67	-0.46	-0.21	0.00	1.20	-1.61	-0.19
D	-1.15	0.67	-0.41	-0.01	-2.68	1.31	0.03	0.56
E	-1.18	0.40	0.10	0.36	-2.16	-0.17	0.91	0.02
F	1.52	0.61	0.96	-0.16	0.25	0.28	-1.33	-0.20
G	-0.20	-1.53	-2.63	2.28	-0.53	-1.18	2.01	-1.34
H	-0.43	-0.25	0.37	0.19	0.51	1.28	0.93	0.65
I	1.27	-0.14	0.30	1.80	-0.30	-1.61	-0.16	-0.13
K	-1.17	0.70	0.70	0.80	1.64	0.67	1.63	0.13
L	1.36	0.07	0.26	-0.80	0.22	-1.37	0.08	-0.62
M	1.01	-0.53	0.43	0.00	0.23	0.10	-0.86	-0.68
N	-0.99	0.00	-0.37	0.69	-0.55	0.85	0.73	-0.80
P	0.22	-0.17	-0.50	0.05	-0.01	-1.34	-0.19	3.56
Q	-0.96	0.12	0.18	0.16	0.09	0.42	-0.20	-0.41
R	-1.47	1.45	1.24	1.27	1.55	1.47	1.30	0.83
S	-0.67	-0.86	-1.07	-0.41	-0.32	0.27	-0.64	0.11
T	-0.34	-0.51	-0.55	-1.06	0.01	-0.01	-0.79	0.39
V	0.76	-0.92	0.17	-1.91	0.22	-1.40	-0.24	-0.03
W	1.50	2.06	1.79	0.75	0.75	-0.13	-1.06	-0.85
Y	0.61	1.60	1.17	0.73	0.53	0.25	-0.96	-0.52

Figure 3.2.3: VHSE encoding scheme for the amino acids

3.2.3 Sequence Model Architecture

Suppose the encoded Ab and Ag sequences are of the dimensions $s_1 \times d$ and $s_2 \times d$, respectively. Initially, each d-dimensional vector would be projected onto another d-dimensional embedding space through a dense layer. Then, the projected matrix would be passed through separate, standard attention blocks whose component layers are multi-head attention (self), dropout, layer norm, and dense layers. The equation for the attention mechanism is given by, $\forall i = 1, 2, 3, \ldots, s$

$$\mathbf{z_i} = \sum_{j=1}^{s} \text{softmax}\left(\frac{< \mathbf{q_i}, \mathbf{k_j} >}{\sqrt{d}}\right) \mathbf{v_j} \tag{3.2}$$

where $\mathbf{z_i}$ is the output vector after the attention layer, s is the sequence length (Ab pathway, $s = s_1$ and Ag pathway, $s = s_2$), d is the embedding vector dimension. Furthermore, $< . >$ indicates the Euclidean inner product. Moreover, $\mathbf{q_i} = W_q \mathbf{x_i}$, $\mathbf{k_i} = W_k \mathbf{x_i}$ and $\mathbf{v_i} = W_v \mathbf{x_i}$ with $\mathbf{x_i}$ being the input

vector and $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$. The softmax operator is defined as given below.

$$\text{softmax}\left(\frac{< \mathbf{q_i}, \mathbf{k_j} >}{\sqrt{d}}\right) = \frac{\exp\left(\frac{\mathbf{q_i}^T \mathbf{k_j}}{\sqrt{d}}\right)}{\sum_{l=1}^{s} \exp\left(\frac{\mathbf{q_i}^T \mathbf{k_l}}{\sqrt{d}}\right)} \tag{3.3}$$

Along with self-attention/ multi-head attention layers, cross-attention layers are also implemented to pass information from the Ab-pathway to the Ag-pathway and vice versa. Moreover, we have included two pathways for cross-attention with the intention of mimicking hierarchical information sharing. The cross-attention layer has a similar equation as in the case of self-attention with a subtle difference in the limits of summation. Accordingly, for cross-attention, we can provide the following equation considering the flow of information from the Ag pathway to the Ab pathway. $\therefore \forall i = 1, 2, 3, \ldots, s_1$,

$$\mathbf{z_{i(Ab)}} = \sum_{j=1}^{s_2} \text{softmax}\left(\frac{< \mathbf{q_{i(Ab)}}, \mathbf{k_{j(Ag)}} >}{\sqrt{d}}\right) \mathbf{v_{j(Ag)}} \tag{3.4}$$

where $\mathbf{z_{i(Ab)}}$ is the output Ab vector after the cross-attention layer. Moreover, $\mathbf{q_{i(Ab)}} = W_q \mathbf{x_{i(Ab)}}$, $\mathbf{k_{j(Ag)}} = W_k \mathbf{x_{i(Ag)}}$ and $\mathbf{v_{j(Ag)}} = W_v \mathbf{x_{i(Ag)}}$ with $\mathbf{x_{i(Ab)}}$ and $\mathbf{x_{i(Ag)}}$ being the input Ab vector and Ag vector, respectively and $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$. Other symbols hold the same meaning as in the self-attention layer.

The outputs from the Ab pathway, Ag pathway, and the Ab-Ag pathways are then concatenated into a single vector and further processed through two dense layers to yield the predicted IC50 value. During the training process, mean squared error(MSE) was employed to adjust the weights.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_{true} - y_{pred}\right)^2 \tag{3.5}$$

3.2.4 Contrastive Learning-based Sequence Model

With the intention of increasing the sensitivity of the model to mutations, we trained the sequence model using a contrastive learning approach where we tried to maximize the disparity between predictions produced for the initial sequences and the mutated sequences. We employed a conventional MSE loss for the initial sequences and the reciprocal of the MSE loss for the mutated sequences. As the model tries to minimize the total loss, it would reduce the gap between the ground truth value and prediction for the initial sequences while increasing the gap between the ground truth and the prediction for the mutated sequences.

Accordingly, there are three replicas of the sequence model, two of which are only used during training and are not considered during inference to derive the predicted IC50. The three replicas process,

- Replica 1 - Original antibody and antigen
- Replica 2 - Original antibody and mutated antigen
- Replica 3 - Mutated antibody and original antigen

The types of mutations introduced to the protein sequences during the training process are insertion, deletion, substitution, random cropping, and/ or a combination of these mutations.

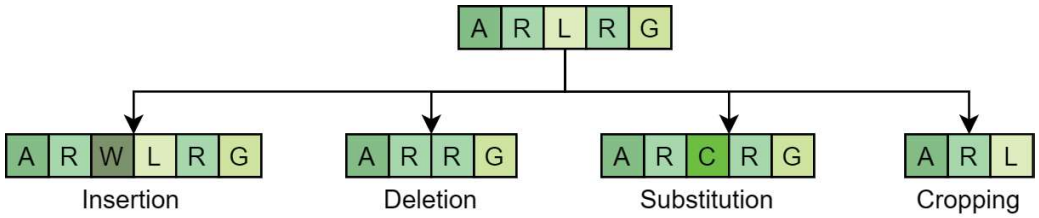

Figure 3.2.4: Types of mutations introduced to the protein sequences in the contrastive learning approach

- Insertion - A random amino acid would be inserted into a random position of the amino acid sequence of the protein
- Deletion - A random amino acid would be deleted from a random position of the amino acid sequence of the protein
- Substitution - A random amino acid would replace another amino acid that is already in the amino acid sequence of the protein
- Cropping - A segment of the amino acid sequence would be removed from the original sequence
- Combination - Any combination of the aforementioned mutations

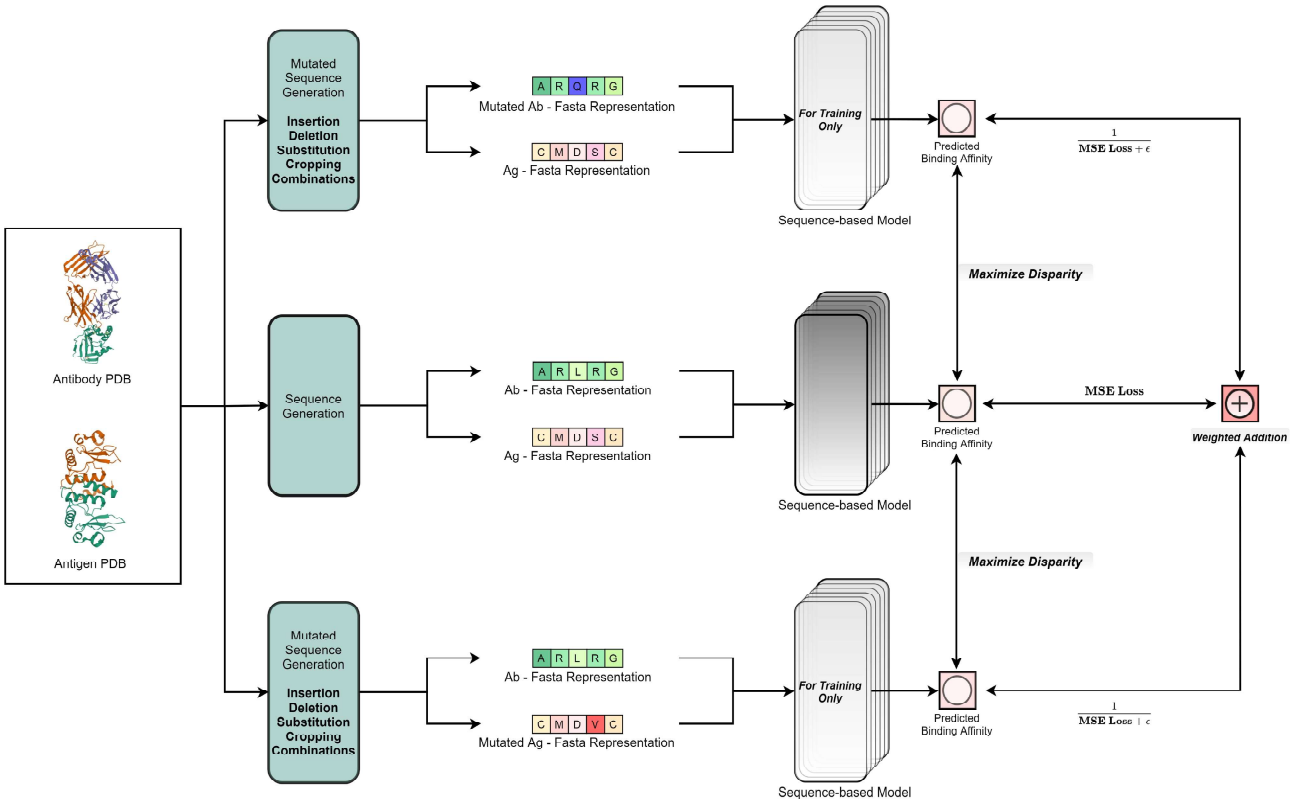

Figure 3.2.5: Overview of the adjusted sequence model for the contrastive learning process

As per the MSE loss defined in Eq. 3.5, the following loss was used to train the sequence model along with the contrastive learning approach. In the following equation, $\epsilon = 1e-5$ was used

to avoid zero division error.

$$\text{Total Loss} = \alpha \times \text{MSE}_{\text{Ab-Ag}} + \beta \times \frac{1}{\text{MSE}_{\text{Ab}_{\text{mutated}}-Ag} + \epsilon} + \gamma \times \frac{1}{\text{MSE}_{\text{Ab-Ag}_{\text{mutated}}} + \epsilon} \tag{3.6}$$

where α, β, and γ are adjustable hyperparameters that change the weight of the loss from each pathway.

3.3 Structure Based Model

In this section, we describe the structure-based model that takes molecular graph representations of antigens and antibodies as input.

3.3.1 Protein Data Bank Format

PDB file format which stands for protein data bank format is the standard format for storing a protein and its related details in a text-based file. The PDB file of a given protein includes the details required to fully define the 3D structure of the protein or in other words, it contains the relative atomic coordinates of each atom in the protein and a cartesian coordinate system is used in this context. In addition, it provides the details pertaining to the secondary structure of the protein of concern while summarizing certain details related to the bonds in the protein. These bond details reflect the atomic connectivity of the structure. The other key feature of this file format is that it enables the storage of experimental metadata. Recently, a newer format was proposed by the Protein Data Bank and it is referred to as Macromolecular Crystallographic Information File (mmCIF) format. The content of the PDB file format is summarized below.

- Header, Title, and Author Details: This section includes details of the individual/ group who defined the protein structure through experiments. Further, it includes the mode used to determine the 3D structure (i.e. X-ray diffraction, Nuclear Magnetic Resonance Imaging, etc.). In addition, the date is mentioned along with an appropriate title for the protein.
- Remarks: This section indicates miscellaneous information as well as standard information corresponding to the atomic coordinate calculations.
- Sequence Details: This section includes the details of the peptide chains (primary structure of the protein)
- Atom Details: This is the key section of the PDB file as it includes the 3D cartesian coordinates of the constituent atoms. The units associated with distance measurements are Angstroms. The other details provided in this section are element name (symbol), temperature factor, and occupancy.
- Hetero Atom Details: This section includes certain details of the non-constituent atoms of the protein.

There are multiple software tools that are capable of visualizing and analyzing the PDB file formats. Some standard tools are Chimera, Molekel, Houdini, PyMol, VMD, BioBlender, etc.

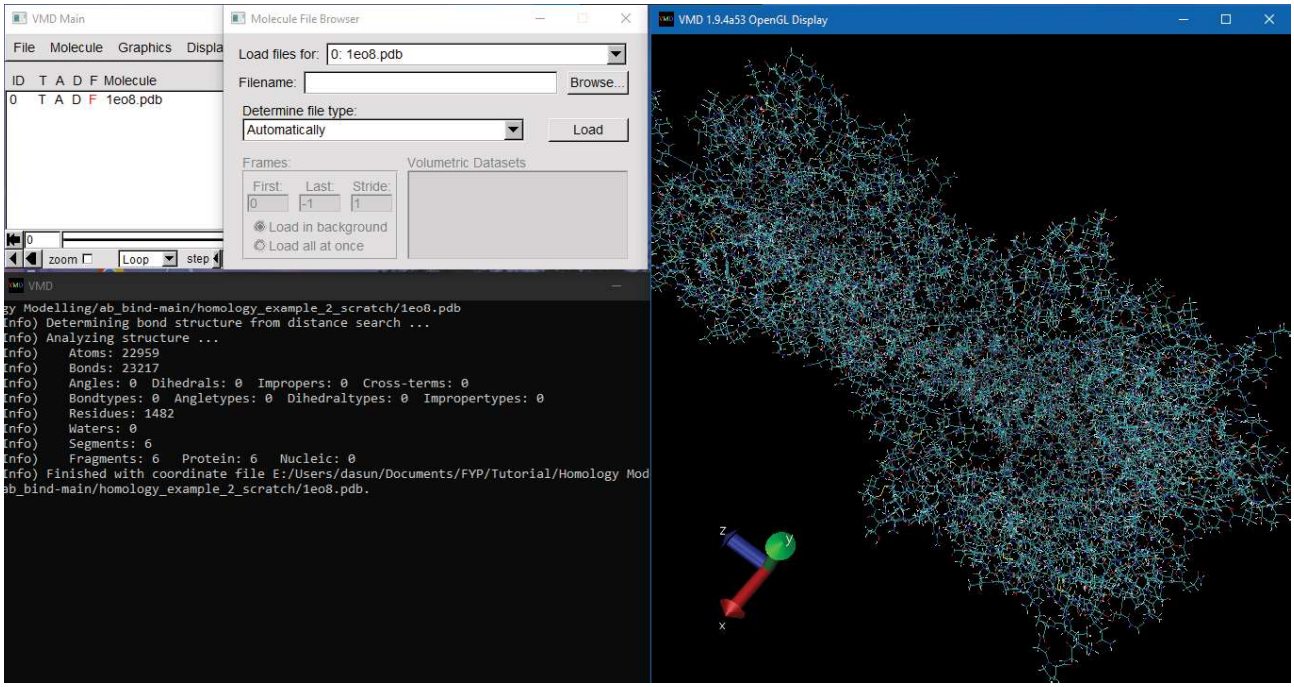

Figure 3.3.1: Analysis performed using VMD visualization tool on the 1E08.pdb which includes the details of the 3D structure of the Iron-Hydrogenase/cytochrome C553 complex

3.3.2 Graph Representation of Molecules

We represent both antibody and antigen molecules as graphs $G = (V, E)$ that use atoms as nodes with their respective 3D coordinates denoted as $X \epsilon \mathbb{R}^{3\times n}$, and initial atomic features $F \epsilon \mathbb{R}^{4\times n}$. Edges include all atom pairs within a distance cutoff of $5\mathring{A}$. Initial atomic features include atomic number, implicit valance, charge list, and the degree of the atom. Edge features include the bond strength and we calculate the bond strength as the reciprocal of bond distance.

3.3.3 Structure Model Architecture

The structure model architecture comprises 2 parallel branches that simultaneously process antibody and antigen graph representations. Each branch consists of 4 graph convolutional layers [39], followed by 4 graph attention layers [40] and a graph average pooling layer. Toward the goal of aggregating the features of neighboring atoms, we feed the graph representations of antibody and antigen molecules into a stack of graph convolutional layers. The core property of the graph convolutional layer is that it takes the weighted average of all neighbors' node and edge features, including itself. In general, the graph feature aggregation and updates can be represented as follows:

$$v_i^{(t+1)} = g^t(\rho(\{e_{ij}^t | (i,j) \in E\}), v_i^t) \tag{3.7}$$

where e_{ij}^t refers to the edge features at t^{th} iteration while v_i^t refers to the node features at t^{th} iteration. $\rho(.)$ aggregates the edge features to nodes and $g(.)$ updates the node feature vector and assigns to the new node feature vector at $(t+1)^{th}$ iteration, $v_i^{(t+1)}$.

Here, a single graph convolutional layer, followed by a ReLU non-linear layer is denoted as follows.

$$Z' = \text{ReLU}\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}ZW\right) \tag{3.8}$$

where $Z \epsilon R^{8\times n}$ denotes the input feature matrix that includes coordinate values, node, and edge features of n nodes. Here, A is the adjacency matrix, $\tilde{A} = A + I_N$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

Following the stack of graph convolutional layers, a stack of 4 graph attention layers is applied to identify the atoms that should be given more priority. The graph attention operator works as follows.

$$Z_i' = \alpha_{i,i} W Z_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j} W Z_j \tag{3.9}$$

$$\alpha_{i,j} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top \left[\boldsymbol{W}Z_i \,\|\boldsymbol{W}Z_j\| \,\boldsymbol{W}_e e_{i,j}\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}} \exp\left(\text{LeakyReLU}\left(a^\top \left[\boldsymbol{W}Z_i \,\|\boldsymbol{W}Z_k\| \,\boldsymbol{W}_e e_{i,k}\right]\right)\right)} \tag{3.10}$$

where Z_i is the input feature vector of i th node and W is the weight matrix. Afterwards, we feed the graph attention layer output to a graph average pooling layer to reduce the spatial dimension of the output. After concatenating the final embeddings of antigen and antibody, the final output on the binding affinity is predicted after passing the concatenated embedding through 2 linear layers.

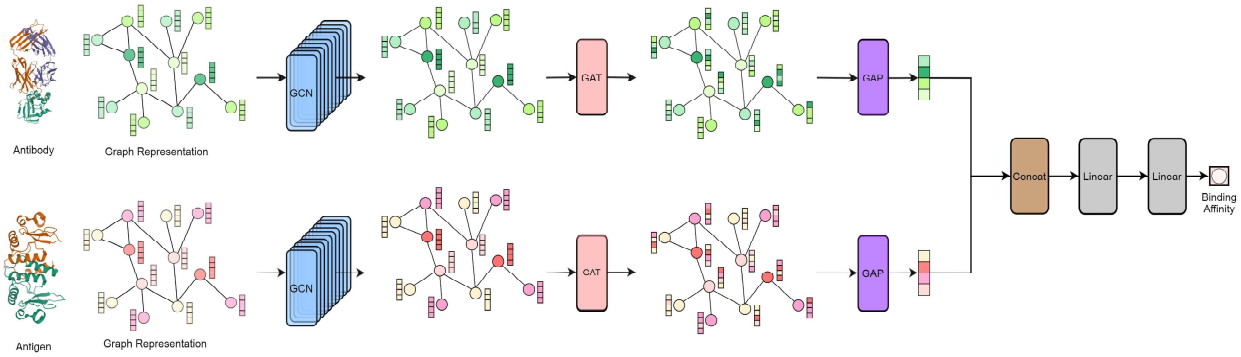

Figure 3.3.2: Structure-based Model Architecture

3.4 Combined Model

With the expectation of capturing evolutionary details through protein amino acid sequences and atomistic details through protein 3D structures to enhance the prediction performance in binding affinity, the final combined model consists of the sequence and structure model mentioned in section 3.2.3 and 3.3.3, respectively. We apply cross-attention to capture the complementary information between sequence and structure domains. Cross-attention output is again concatenated with the two separate embeddings from parallel models and fed into 2 linear layers to get 2 predictions for binding affinity. Final binding affinity is a weighted average between those two predictions and the weights are decided based on the conducted ablation study.

$$Final\ BA = \delta_1 \times BA\ from\ Structure\ Model + \delta_2 \times BA\ from\ Sequence\ Model \tag{3.11}$$

BA denotes the predicted binding affinity. Here, we use a weighted loss function, through the hyperparameters of δ_1 and δ_2 with the same weights mentioned previously for the prediction. Through ablations, δ_1 is selected to be 0.45, and δ_2 is selected to be 0.55.

In training, a combined MSE loss function is introduced to incorporate the influence of two binding affinity predictions from parallel and interconnected sequence and structure paths into the synchronous training process.

$$\text{Combined loss} = \alpha \times MSE(StP, T) + \beta \times MSE(SeP, T) + \gamma \times MSE(StP, SeP) \tag{3.12}$$

where StP, SeP and T refer to the predictions from the structure-based model path, sequence-based model path, and the target true value respectively. Furthermore, α, β, and γ are hyperparameters to be selected through ablation studies.

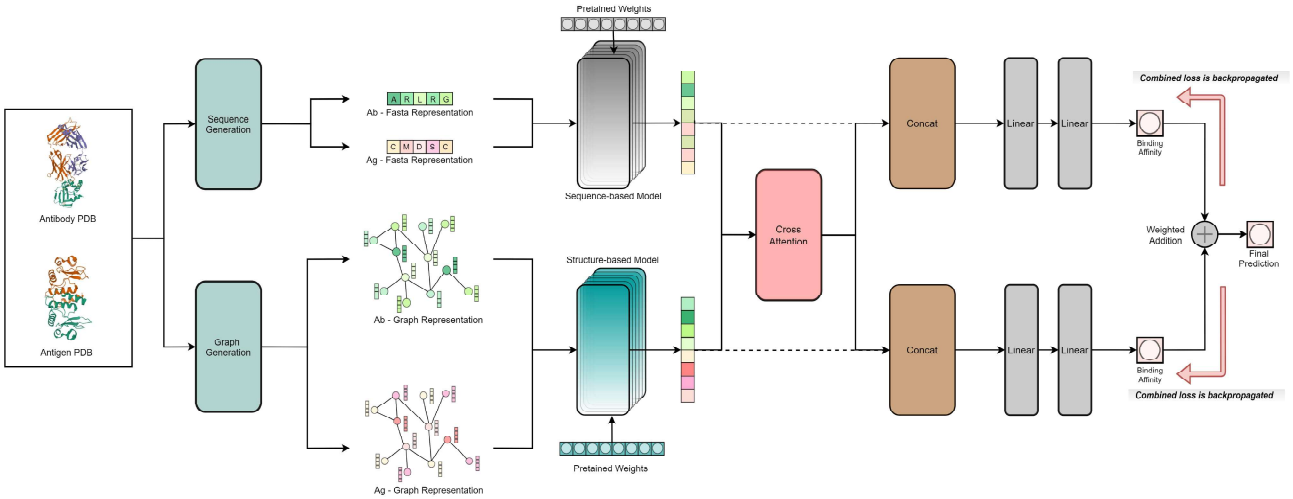

Figure 3.4.1: Combined Model Architecture consisting of sequential and structural branches

3.5 Community Access Tool: Web-based Platform

The website serves as a community access tool and is structured with three servers: one for the front end and two for backend connectivity. The frontend is developed using the Angular framework, while the backend is built using Flask for one server and Node.js with Express.js for the other server. Azure cloud platforms are utilized to establish connections with cloud services. The deployment is achieved through Docker implementation. The diagram illustrates the connectivity of each component within the web-based platform.

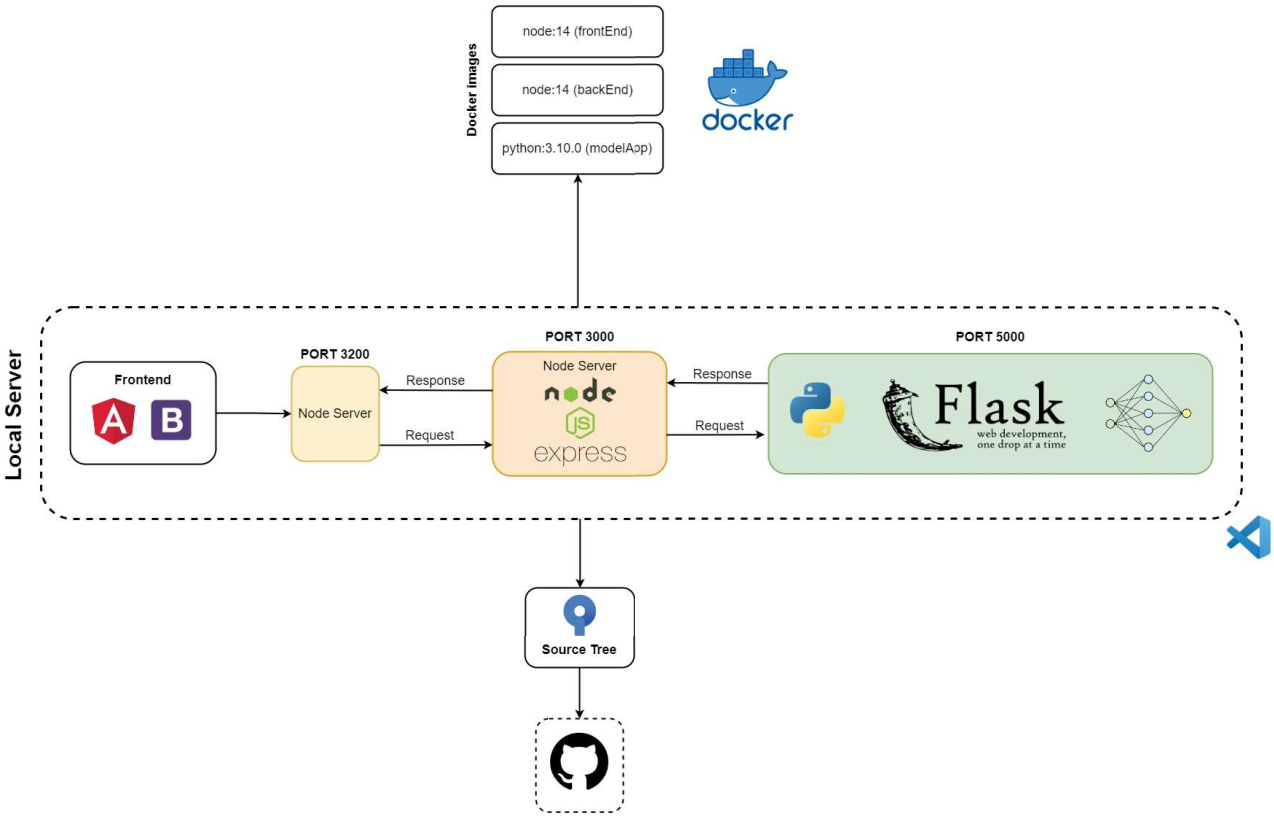

Figure 3.5.1: Web development architecture

In the community access tool, users are provided with the option to upload PDB files of antigens and antibodies. These files are stored in an Azure cloud account blob storage and subsequently processed in a virtual machine on the cloud server. After processing, the binding affinity value (IC50 value) of the antigen-antibody pair will be displayed on the web page.

Chapter 4

Experiments and Results

In this chapter, we discuss in detail our experiments and results in comparison with several baseline works in the literature. The results are presented and discussed in the following subsections: dataset curation, sequence-based models, ablations on features, structure-based models, combined models, contrastive learning against mutations, and ablations on model parameter selections.

4.1 Dataset Curation

The utilized publically available datasets were preprocessed uniquely through their own preprocessing pipeline as presented in the section 3.1 and the resulting final curated datasets are denoted as **P2PXML-Seq** and **P2PXML-PDB** are compared with the publicly available datasets in the Table 4.1. P2PXML-Seq only contains antibody-antigen pairs in protein amino acid letter sequence format while the P2PXML-PDB only contains antibody-antigen pairs in PDB format which thereby represent the protein 3D structures.

	Datapoints	Usable Datapoints*	Mutations	Type	Numerical Values
AB-Bind	1 101	1 101	✓	Sequence	$\Delta\Delta G$
Ab-Cov	1 964	1 420	✓	Sequence	IC50, EC50
CATNAP	129 686	11 208	✗	Names only	IC50, IC80, ID50
AlphaSeq	**1 259 700**	352 139	✗	Sequence	IC50
SAbDab	1 327	N/A	✓	Structure	$\Delta\Delta G$, affinity
Skempi	7 086	N/A	✓	Structure	affinity
P2PXML-Seq**	365 868	**365 868**	✓	Sequence	IC50
P2PXML-PDB**	129 143	**129 143**	✓	Structure	IC50

Table 4.1: Dataset comparison. *Here, usable datapoints refer to the remaining datapoints after a defined set of preprocessing steps including the duplicate removal and the removal of datapoints with no numerical value for binding affinity. **P2PXML-Seq is our curated protein sequence dataset and P2PXML-PDB is our curated protein structure dataset.

As per our understanding, our curated datasets are the largest datasets specifically developed for the antibody-antigen binding affinity prediction as highlighted through Table 4.1. Furthermore, our datasets have a consistent format with respect to the antibody-antigen pairs and the corresponding numerical values (i.e. IC50) and express sufficient generalizability, unlike most of the other datasets

in the literature where only one antigen is considered, since our datasets contain pairs of numerous antigens such as SARS-CoV-2, HIV, MERS, flu and their related antibodies.

4.2 Evaluation Parameters

Throughout the experiments, the following parameters are utilized to evaluate and compare the results between the performance of each model as in the literature due to the nature of the prediction task in the regression fashion.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.1}$$

where the evaluation parameter: mean absolute error is denoted as MAE and y_i and $\hat{y}_i$ refer to the true value and prediction value respectively.

In certain occasions, the mean squared error is also considered as an evaluation parameter and as a loss function in an ad-hoc manner to further validate the results.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

4.3 Sequence-based Models

As descriptively discussed in section 3.2, the sequence-based model is developed to capture evolutionary details of the amino acid sequences of both antigens and antibodies while hierarchically sharing the information learned from parallel pipelines through cross-attention. Based on this intuition, several approaches were tested and then compared with the state-of-the-art approaches [41, 42] which utilize protein sequences for the binding affinity prediction.

As per the results in Table 4.2, the final sequence-based model is selected to be the parallel transformer model with hierarchical cross-attention of which the model architecture is extensively discussed in section 3.2. The results show that the proposed final sequence-based model surpasses the state-of-the-art methods in all three datasets by MAE margins of 4%, 44%, and 1% for VirusNet, P2PXML-Seq, and P2PXML-PDB datasets respectively and therefore, it is utilized to extract the information from protein amino acid sequences in the final combined model as well.

Model	Parallel Pipeline	Cross attention	VirusNet	P2PXML-Seq	P2PXML-PDB
SVR with RBF kernel [41]	✗	✗	102.1267	10.8236	4.0053
1D CNN with attention [42]	✓	✓	3.5124	0.0964	1.8536
LSTM	✓	✗	5.2853	0.2460	2.3931
LSTM	✓	✓	3.6288	0.0926	1.8676
Transformer	✓	✗	3.4995	0.0898	1.8510
Transformer	✓	✓	**3.3707**	**0.0542**	**1.8358**
Transformer with multiple cross-attention*	✓	✓	3.4744	0.0553	1.8359
Transformer with distogram**	✓	✓	3.3882	0.0551	1.8400
Transformer with protein language embeddings***	✓	✓	3.4263	0.0623	1.8433

Table 4.2: Results comparison between the protein sequence-based models with MAE as the performance parameter. The datasets used are VirusNet, P2PXML-Seq and P2PXML-PDB. Apart from the SVR and 1D CNN models, all other models are proposed by this research for the task. *Here, the transformer model with multiple cross-attention refers to having cross-attention blocks in each stage of the parallel transformer blocks other than the hierarchical two cross-attentions as in the parallel transformer model with cross-attention. **In the transformer model with distogram, the distograms are calculated following [5] with the aim of feeding distograms instead of the encoded protein sequences. ***The pre-trained protein language embeddings are generated from protein sequences through ProtT5-XL-BFD [6] (without fine-tuning the model in [6] to our task) and then fed into the transformer model instead of the encoded protein sequences.

The implementation details of the final sequence-based model are summarized in Table 4.3 after selecting the model and training parameters through a rigorous ablation study.

Parameter	Value/Method
Training	
Learning rate	0.0001
Learning rate scheduler*	LambdaLR
Optimizer	Adam
Loss function	MSE
Batch size	4
Dropout rate	0.05
Training: Validation split	4:1
Early stopping	Validation loss
Early stopping patience	5
Limitations	
Maximum length of antibody	300 amino acids
Maximum length of antigen	1200 amino acids

Table 4.3: The implementation details of the final sequence-based model after selecting parameters through ablation studies. *Here, the learning rate is set constant until a pre-defined epoch and then, made exponentially decaying with a rate of 0.01. The training is performed until converged.

4.4 Ablations on Sequence Encoding Schemes

As discussed in the section 3.2.2, it is required to convert the letter sequences of protein amino-acid representations into a numerical format before feeding into the sequence-based models and three encoding schemes are utilized for experiments for the task: one-hot, VHSE-8 and BLOSUM.

Encoding scheme	MAE
One-hot	0.9757
VHSE-8	**0.9682**
BLOSUM	0.9774

Table 4.4: Results comparison between different encoding schemes. Here, P2PXML-Seq dataset is utilized and the model is a parallel multi-layer perception model.

Through our results as presented in Table 4.4, VHSE-8 encoding scheme is utilized for all the consequent model implementations (including sequence-based models and the combined models) due to its superior performance over the other two encoding schemes. We hypothesize the evolutionary details embedded within the VHSE-8 encoding scheme are the reason for this better performance since the binding affinity prediction is considerably dependent on such evolutionary details which could be extracted through the protein amino-acid representations.

Specifically mentioned in certain occurrences, the protein language embeddings from [6] are utilized to convert the protein amino acid sequences to a numerical format instead of the VHSE-8 encoding scheme to evaluate the performance of such language embeddings, which has been learned from over billions of protein sequences, in compared with the utilized encoding schemes for our task.

4.5 Structure-based Models

As descriptively discussed in section 3.3, the structure-based model is developed to capture the atomistic-level or residue-level structural information of the antibodies and antigens and thereby, enhance the antibody-antigen binding affinity prediction. Based on this intuition, several approaches were tested and then compared with the state-of-the-art approaches [43, 44] which utilize protein 3D structures for the binding affinity prediction.

As per the results in Table 4.5, the final structure-based model is selected to be the parallel graph convolution (GCN) + graph attention (GATConv) model in which the model architecture is extensively discussed in section 3.3. The results show that the proposed final structure-based model surpasses the state-of-the-art methods in our benchmark PDB dataset by a slight MAE margin of 0.08% and an MSE margin of 0.04% and therefore, it is then utilized to extract the information from protein 3D structures in the final combined model as well.

Model	MAE	MSE
Parallel GNN [43]	1.9268	6.8929
Parallel GCN [44]	1.8795	6.6507
Parallel GAT* (Ours)	1.9525	7.0562
Parallel GCN + cross-attention** (Ours)	1.8896	6.6953
Parallel GCN + GATConv (Ours)	**1.8779**	**6.6504**

Table 4.5: Results comparison between the protein structure-based models with MAE and MSE as the performance parameters. The dataset used is our P2PXML-PDB dataset. *Here, the parallel GAT model refers to a full graph attention network which is developed on the hypothesis that identifying the most needed nodes through attention would be sufficient for better binding affinity prediction, but the performance of the model indicates that information is insufficient for better performance. **The parallel GCN model with cross-attention is developed following the success of our final sequence-based model with the expectation that the information sharing between the parallel paths would be beneficial for a better prediction. However, surprisingly, it is not as useful as for the sequence-based model and we hypothesize that the local aggregation of the node features in the intermediate stages lacks or does not represent sufficient global information space which is essential for the graph-level prediction task.

The implementation details of the final structure-based model are summarized in Table 4.6 after selecting the model and training parameters through a rigorous ablation study.

Parameter	Value/Method
Training	
Learning rate	0.01
Learning rate scheduler*	LambdaLR
Optimizer	Adam
Loss function	MSE
Batch size	8
Dropout rate	0.05
Training: Validation split	3:1
Early stopping	Validation loss
Early stopping patience	5
Limitations	
Number of hidden channels in the model	128
Number of layers in the model	32

Table 4.6: The implementation details of the final structure-based model after selecting parameters through ablation studies. *Here, the learning rate is set constant until a pre-defined epoch and then, made exponentially decaying with a rate of 0.001. The training is performed until converged.

4.6 Ablations on Graph Node and Edge Features

One of the most crucial steps in geometric deep learning, specifically graph-based deep learning, is the graph representation which is then fed into the graph learning model since all the node and edge level aggregation and information flow are dependent on the node and edge features fed at the initial state of the graph representation. Furthermore, the nodes and edges are defined at the beginning using the available protein structure data in PDB format based on the determined graph representation as well.

Therefore, to find an optimal set of node and edge features for the graph representation of

proteins and thereby strengthen the chemistry-based intuition of the final model, an extensive ablation study is performed using two datasets: Ab-CoV and Ab-Bind, and the results are presented in Table 4.7.

	Feature	Experiments						
Node	x-coordinate	✓	✓	✓	✓	✓	✓	✓
	y-coordinate	✓	✓	✓	✓	✓	✓	✓
	z-coordinate	✓	✓	✓	✓	✓	✓	✓
	atomic number	✓	✓	✓	✓	✓	✓	✓
	chirality	✓	✓	✓	✗	✗	✗	✗
	implicit valance	✓	✓	✗	✓	✗	✗	✗
	charge list	✓	✓	✗	✓	✗	✗	✗
	degree	✓	✓	✗	✓	✓	✗	✗
	number of atoms in a ring	✓	✓	✗	✗	✓	✗	✗
	radical electrons	✗	✓	✗	✗	✓	✗	✗
	hybridization	✗	✓	✗	✗	✓	✗	✗
Edge	interatomic distances encoded with Gaussian basis functions	✗	✓	✗	✗	✗	✓	✗
	bond strength	✗	✗	✗	✓	✗	✗	✓
Ab-CoV	MAE	8.0124	7.2343	8.3865	**7.0127**	8.1923	9.3845	9.8632
Ab-Bind	MAE	0.0004	0.0009	0.0008	**0.0001**	0.0003	0.0035	0.0055

Table 4.7: Evaluation of node and edge features for the protein structure-based model using the protein structures generated (using AlphaFold-V2 multimer model) for the protein sequences in Ab-Cov and AB-Bind datasets

Through the results, the following set of node features: x-coordinate, y-coordinate, z-coordinate, atomic number, implicit valence, charge list, degree, and edge features: bond strength and atomic distance is used for the remaining experiments and final model implementations of both the structure-based model and the combined model.

4.7 Combined Models

The intuition behind the combined model is to incorporate both the evolutionary details of antigens and antibodies through the final sequence-based model and the atomistic level features of the antibodies and antigens through the final structure-based model while sharing the information learned through both pipelines to imitate the chemical binding potential.

The combined model is developed in the following stages and the resulting performance after the implementation of each stage is conveyed in Table 4.8.

- Combined-B: This refers to the combined base model where the outputs from the final sequence-based model and the structure-based models are collectively considered for calculating the combined loss function as presented in section 3.4.
- Combined-V1: The Combined-B model is modified such that the last latent vectors from sequence-based and structure-based models are concatenated and then, the concatenated vectors are passed through their parallel paths to individually predict the outputs as suitable for the combined loss function.

- Combined-V2: Instead of the latent vector concatenation in the Combined-V1 model, the resulting vector from cross-attention is concatenated with the latent vectors from each model and then, that concatenated vector is passed through their parallel paths to individually predict the outputs as suitable for the combined loss function.
- Combined-V2 + pre-trained weights: Instead of randomly initializing the weights of the Combined-V2 model at the start of the training, the pre-trained weights are utilized for weight initialization which is obtained from separately training our final protein sequence-based model (using the P2PXML-Seq dataset). Here, pre-trained weights are not utilized for the structure-based counterpart in the Combined-V2 model since both the structure-based model and the combined models are trained on the P2PXML-PDB dataset and in the implementation, the protein 3D structure files in the PDB format are used to generate both graph representation and amino acid sequences which are then separately fed into their respective pipelines.

Model	VHSE8 encoding	Protein language embeddings (from ProtT5-XL-BFD)
Combined-B	1.7503	1.8567
Combined-V1	1.6628	1.8200
Combined-V2	1.6576	1.8139
Combined-V2 + pre-trained weights	**1.6572**	**1.8094**

Table 4.8: Results comparison between the VHSE8 encoding and pre-trained protein language embeddings from ProtT5-XL-BFD [6] for protein amino-acid sequences. The dataset used is our P2PXML-PDB dataset and the performance parameter is MAE.

Through the results presented in Table 4.8, it is evident that the introduced combined loss function expresses enhanced performance in binding affinity prediction than the individual sequence-based and structure-based models (5.6% improvement than the final sequence-based model and 6.8% improvement than the final structure-based model) establishing the importance of the combined loss function in the Combined-V2 model.

The Combined-V1 model outperforms the Combined-B model by a margin of 5% which highlights the importance of incorporating both evolutionary and atomistic details through sequence-based and structure-based pipelines respectively for enhanced binding affinity prediction. The cross-attention towards sharing information between learned representations utilized in Combined-V2 further improves the performance of Combined-V1 validating the importance of sharing learned representations from sequence-based and structure-based pipelines as we initially hypothesized. The utilization of pre-trained weights slightly improved the results further to the lowest MAE of 1.6572.

Through Table 4.8, it could be deduced that the same pattern of performance improvement is perceived when utilizing the protein language embeddings from [6] instead of the VHSE-8 encoding scheme to encode the protein amino-acid sequences, but the superior performance of VHSE-8 encoding over protein language embedding in every combined model variant suggests that it is better to use an encoding scheme rather than deep learning-based protein language embeddings given the fact that we utilize the language embeddings from a pre-trained model rather than a model which is fine-tuned to the task in hand. We further hypothesize that if we fine-tune the protein language model to our task, it may have a comparable or superior performance than the traditional encoding scheme we utilized since it could learn an enriched protein sequence representation tailored to the task rather than an ill-posed generalized representation which is not sufficient to acquire better performance.

Model	MAE
1D CNN with attention [42]	1.8536
GCN [44]	1.8795
Parallel Transformer with cross-attention (Ours)	1.8358
Parallel GCN + GATConv (Ours)	1.8779
Combined-V2 + pre-trained weights (Ours)	**1.6572**

Table 4.9: Overall results comparison. Here, the pre-trained weights are obtained from separately training our final protein sequence-based model (using the P2PXML-Seq dataset) and utilized for the weights initialization of the Combined-V2 model in the training. The dataset used is our P2PXML-PDB dataset.

Through Table 4.9, the summarized overall results are presented which shows that our final Combined-V2 model outperforms all the considered state-of-the-art approaches at least by a margin of 10.6% in both sequence-based and structure-based predictions and improved the results of our own final sequence-based and structure-based models as well.

The implementation details of the final combined model are summarized in Table 4.10 after selecting the model and training parameters through a rigorous ablation study.

Parameter	Value/Method
Training	
Learning rate	0.001
Learning rate scheduler*	LambdaLR
Optimizer	Adam
Loss function	Combined loss (3.12)
α in Combined loss	0.49
β in Combined loss	0.49
γ in Combined loss	0.02
Batch size	8
Dropout rate	0.05
Training: Validation split	3:1
Early stopping	Validation combined loss
Early stopping patience	5
Limitations	
Maximum length of antibody	300 amino acids
Maximum length of antigen	1200 amino acids
Number of hidden channels in the structure path	128
Number of layers in the structure path	32

Table 4.10: The implementation details of the Combined-V2 model after selecting parameters through ablation studies. *Here, the learning rate is set constant until a pre-defined epoch and then, made exponentially decaying with a rate of 0.001. The training is performed until converged.

4.8 Contrastive Learning towards Enhanced Mutation Dependancy

Since mutations play a vital role in determining the binding affinity such that the binding affinities should reflect the variations in binding affinity even under a single point mutation, it is required that our models should be sensitive enough for any kind of possible mutations such that the predictions from the model should deviate for the original antibody-antigen pair than the mutated

antibody-antigen pair. In the current implementation, the final sequence-based model is utilized for enhancing the sensitivity of the model against mutations since it is comparatively ergonomic to add mutations to protein amino-acid sequences, and in this context, it should be noted that the contrastive learning-based proposed approach can be easily extended to the structure-based and combined models through mutating the protein 3D structures via homology modelling.

First, the sensitivity of our final sequence-based model is evaluated as presented in Figure 4.8.1. Here, for each antibody-antigen pair in the test set in the P2PXML-Seq dataset, the absolute difference between the binding affinity prediction for the original pair and the one randomly mutated pair (the implementation and types of mutations are presented in section 3.2.4 and the percentage mutation refers to the degree of mutation: as an example if the percentage is 8% and the length of protein amino acid is 200, then, 16 amino acids in the protein sequence will be randomly mutated) is calculated and then, the continuous population density is plotted against the calculated absolute deviation.

Through the results in Figure 4.8.1, it is evident that the model is sensitive to the implemented random multiple-point mutations since when the degree of percentage mutation is getting higher from 8% to 40%, the density distribution against the absolute deviation is getting dispersed, suggesting the deviation between the predictions for original pair and the mutated pair is getting higher, which is as expected.

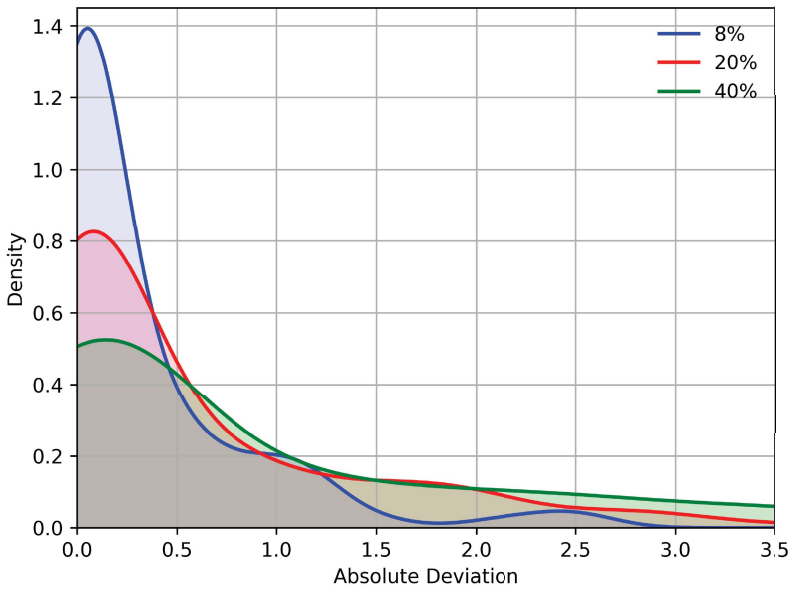

Figure 4.8.1: The sensitivity analysis of the final sequence-based model considering the population density against the absolute deviation of the model's prediction before and after the percentage mutation. Here, the absolute deviation is calculated as the absolute difference between the prediction for the original antibody-antigen pair and the prediction for the randomly mutated antibody-original antigen pair.

However, as in Figure 4.8.1, the model is lacking in mutation sensitivity especially for a lower degree of percentage mutation when the percentage is 8%. However, the ideal scenario is that the model should be sensitive enough even for a single point mutation, and therefore, to further enhance the model's sensitivity against mutations, we explore a contrastive learning approach in which the MSE loss between the predictions for the original pairs and the mutated pairs are forced to be increased

contrastively while the MSE loss between the prediction for original pair and the true value is forced to be reduced as presented in section 3.2.4.

In training the contrastive learning-added sequence-based model, the pipeline presented in Figure 3.2.5 and in inference, only the path of original pairs in the pipeline is utilized (i.e. for inference, all the predictions for both original pairs and mutated pairs are obtained from the middle path in Figure 3.2.5, neglecting the output predictions from upper and lower paths). As per the results shown in Figures 4.8.2 and 4.8.3 for 8% percentage mutation degree and 20% percentage mutation degree respectively, it is evident that the proposed contrastive learning approach significantly improves the mutation sensitivity of the sequence-based model by further dispersing the population density vs absolute deviation curves for each case as we hypothesized.

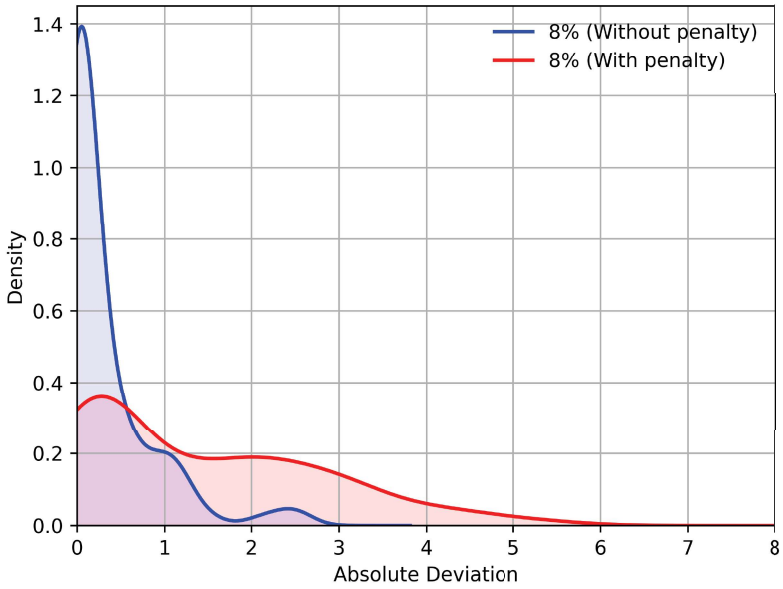

Figure 4.8.2: The sensitivity of the final sequence-based model before (i.e. without penalty) and after the contrastive learning approach (i.e. with penalty) to induce the model to be robust against random multiple point mutations. Here, the percentage level of mutation is 8%.

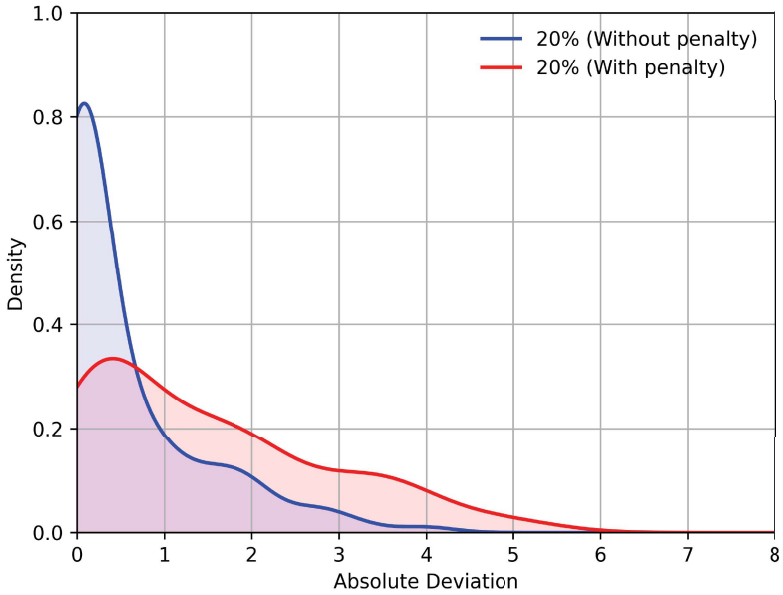

Figure 4.8.3: The sensitivity of the sequence-based model before (i.e. without penalty) and after the contrastive learning approach (i.e. with penalty) to induce the model to be robust against random multiple point mutations. Here, the percentage level of mutation is 20%.

4.9 Inference on Flu Virus Sequences

To evaluate the performance of our models in real-world scenarios, we inferred an unresolved set of flu viruses collected from [45] and performed over 2 million antibody-antigen combination evaluations and found 25 best combinations as suitable candidates for wet lab experiments. These wet lab experiments will be conducted as a joint effort in the future.

The antigens for which we found the best candidate antibodies are: H1_NC_99, H3_SG_93, H2_JP_57, H6_AB_85, H1_CA_2009, H1_SZ_95, H5_VT_2004, H1_SD_2007, H3_VC_75, H1_PR_34, H3_VC_2011, H3_PA_99, H1_SI_2006, H3_WH_95, H3_PT_2009, H3_WI_2005, H1_WSN_33, H7_BC_2004, H3_CA_2004, H3_HK_68, H1_BR_2010, H1_FM_47, H9_HK_97, H3_SY_97 and H1_BJ_95.

4.10 Timing Analysis

One of the most critical challenges in traditional wet-lab-based or molecular dynamics-based experiments for antibody-antigen binding affinity is the excessive time they take to completion of the expected output whereas wet-lab-based will typically take months to years and the molecular dynamics usually take weeks to months to produce their result. However, through deep learning-based approaches, it is possible to obtain the predicted binding affinity within minutes or even seconds with satisfactory accuracy and performance. As such both our Combined-V2 model and the sequence-based model are able to predict the binding affinity of a given antibody-antigen pair within 1 minute on a GPU and four minutes on a CPU in inference as shown in Table 4.11 which highlights the lack of need for further optimization for timing, sacrificing the performance.

Machine	Elapsed Time (s)	Average Inference Time per Ab-Ag Combination (s)
Combined-V2 model		
GPU (NVIDIA T4)	2694	22
CPU	18621	149
Sequence-based model*		
GPU (Tesla P100)	1052	50
CPU	4735	225

Table 4.11: Timing analysis of the Combined-V2 model and the final sequence-based model. *Here, the sequence-based model refers to the parallel transformer model with cross-attention

4.11 Ablations on Model/Training Parameter Selections

Learning Rate	Optimizer	Dropout	Embeddings	MAE	MSE
0.001	ADAM	0.05	✗	0.179275	0.049301
		0.1	✗	0.335331	0.116954
		0.2	✗	0.130520	0.039792
0.0001	ADAM	0.05	✗	**0.054212**	0.036044
		0.1	✗	0.088051	0.036068
		0.2	✗	0.058286	0.035924
		0.05	✓	0.103310	0.036906
		0.1	✓	0.066851	**0.035740**
		0.2	✓	0.062289	0.035802
	ADADELTA	0.05	✗	0.089726	0.036105
		0.1	✗	0.091582	0.036193
		0.2	✗	0.094721	0.037684

Table 4.12: Optimal hyperparameter value determination through ablation by exposing the proposed sequence model to the AlphaSeq dataset

It is important to run extensive ablations for different combinations of hyperparameters as we were trying to heuristically determine the parameters that would maximize the performance of the model. Accordingly, as per the table 4.12, it is evident that the best MAE is given by a learning rate of 0.0001, and ADAM optimizer with a dropout of 0.05. The set of hyperparameters that exhibit the best MSE is almost similar to those of the best MAE except for the dropout which is 0.10 in the case of best MSE. Nevertheless, under the hyperparameters that produce the best MSE, we obtain an MAE that is relatively worse than the best MAE. Moreover, note that the MSE corresponding to the best MAE is quite close to the best MSE. Accordingly, it was decided that a learning rate of 0.0001, and ADAM optimizer with a dropout of 0.05 are appropriate hyperparameter values for the training process of the designed models.

4.12 Website, Project Page and Code Availability

The web platform is developed as a community access tool where interested personnel can access the outputs from our Combined-V2, final sequence-based, and structure-based models for their input protein sequences and/or PDB files. The input format of each model on the website is as follows:

Model	Input format	Output format
Combined-V2	PDB files	Predicted binding affinity in $\mu g/ml$
Final sequence-based model	Text sequence (Pasted)	
Final structure-based model	PDB files	

Table 4.13: The input and output format for each model hosted on the website. The validity of the inputs is checked before feeding to our models through rule-based conditioning.

The website is accessible through this link: `https://p2pxml.azurewebsites.net/` while the project page and the codes will be available on this link: `https://github.com/aravinda1879/ab_bind`.

Figure 4.12.1 shows a snapshot of the developed web platform.

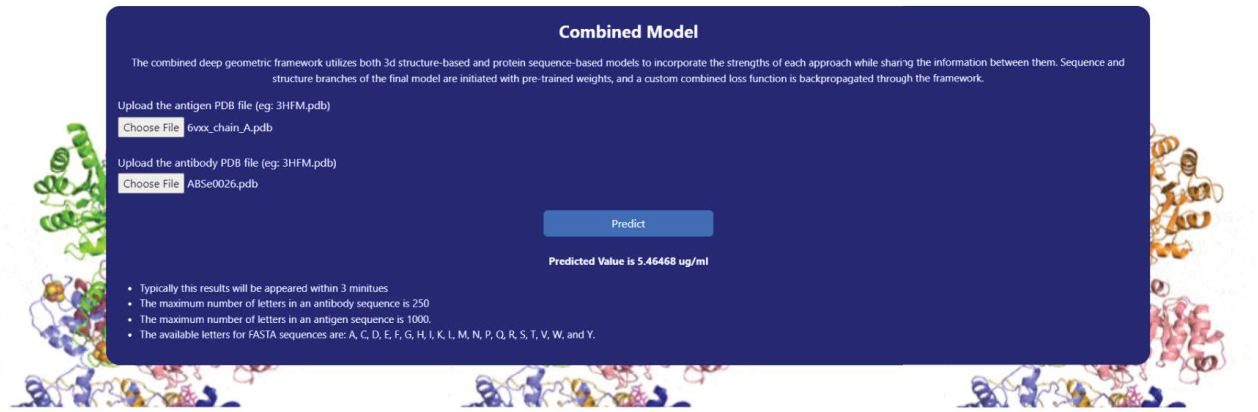

Figure 4.12.1: Designed Web interface

Chapter 5

Discussion and Conclusion

In the field of drug development, the efficacy of a drug depends on the extent to which the constituent molecules interact with the target molecules. Thus, the strengths of those interactions must be evaluated during the drug design phase to achieve the desired efficacy levels. In literature, such protein-protein interactions are reflected by the binding affinity which is a qualitative parameter. The corresponding quantitative parameter is referred to as the binding energy. Therefore, accurate binding energy prediction is critical in designing drugs with higher affinity and specificity towards their target. Our study focused on antibody-antigen binding, which is a subclass of proteins. Currently, techniques such as molecular docking and molecular dynamics simulations are employed to determine the binding affinity at different binding poses but they either overlook the temporal behaviour leading to lesser accuracy (in molecular docking) or are computationally expensive and time-consuming (in MD). Even though there exists an emerged interest in utilizing machine learning to predict binding affinity with lesser computational cost, the predictive performance of existing ML methods when calculating binding affinity is highly dependent on the quality of the Ab-Ag structures and they tend to overlook the importance of capturing the evolutionary details of proteins upon mutation.

To overcome the said complexities and drawbacks, we developed a novel deep geometric network that consists of a geometric model that could process the 3D structures of the input proteins and a sequence model that could handle the amino acid sequences of the input proteins. We employed attention mechanisms in both models to ensure that atomistic-level information, as well as evolutionary information, are incorporated into neighbour embeddings or between information flows of antibodies and antigens effectively and sufficiently. The proposed model was trained on our curated dataset which consists of sequences and structures of multiple antigens and antibodies including a diverse set of common viruses such as HIV and SARS-CoV-2 and their corresponding antibodies to ensure sufficient generalizability within the dataset.

After extensive ablation studies which were performed to select the encoding schemes, node and edge features, model training and inferring parameters, through the final results, it was observed that the proposed sequence-based model was able to surpass the existing sequence-based state-of-the-art methods in all benchmark datasets while our structure-based model outperforms the works in the literature as well. The Combined-V2 model, which is our final model integrating both final sequence-based and structure-based models, surpassed the state-of-the-art approaches at least by a margin of 10.6% in terms of the mean absolute error.

An additional contrastive learning approach is developed to improve the mutation sensitivity of the sequence-based model and the results demonstrate that it was instrumental in significantly enhancing the expected sensitivity in the sequence-based model. The proposed contrastive approach

can easily be extended to the structure-based and Combined-V2 models as well through homology modelling. Our work was further utilized to infer a real-world flu dataset and the wet-lab experiments are scheduled to further validate the performance of the framework. Furthermore, we have developed a website as a community access tool to allow the interested community to get prediction results for their input proteins through our hosted models.

Therefore, in summary, we believe that this work for be instrumental for accelerating the accurate prediction of the binding affinity of antibody-antigen pairs in the domain of drug development, especially towards monoclonal antibody therapy.

Bibliography

[1] Wikimedia commons, "Aspirin," July 2023, https://commons.wikimedia.org/wiki/File:Aspirin-B-3D-balls.png.

[2] Wikipedia, "Insulin degludec," May 2023, https://en.wikipedia.org/wiki/Insulin_degludec.

[3] TechWeb RSS, "Molecular dynamics simulation of ion propagation through a protein ion channel," https://www.bu.edu/tech/support/research/whats-happening/highlights/ion/.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug 2021.

[5] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zıdek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis, "Alphafold: Improved protein structure prediction using potentials from deep learning," 2020.

[6] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost, "ProtTrans: Toward understanding the language of life through self-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7112–7127, Oct. 2022.

[7] Editorial Team and Tina Tran, "Small and large molecules: Pharmaceutical drug development," May 2023, https://blog.contractlaboratory.com/pharmaceutical-drug-development-small-molecules-large-molecules/.

[8] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu, "Development of therapeutic antibodies for the treatment of diseases," *J. Biomed. Sci.*, vol. 27, no. 1, pp. 1, Jan. 2020.

[9] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane, "SAbDab: the structural antibody database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1140–D1146, nov 2013.

[10] Sarah Sirin, James R. Apgar, Eric M. Bennett, and Amy E. Keating, "scpAB/scp -bind: Antibody binding mutational database for computational affinity predictions," *Protein Science*, vol. 25, no. 2, pp. 393–409, nov 2015.

[11] Hyejin Yoon, Jennifer Macke, Anthony P. West, Brian Foley, Pamela J. Bjorkman, Bette Korber, and Karina Yusim, "CATNAP: a tool to compile, analyze and tally neutralizing antibody panels," *Nucleic Acids Research*, vol. 43, no. W1, pp. W213–W219, jun 2015.

[12] Puneet Rawat, Divya Sharma, R Prabakaran, Fathima Ridha, Mugdha Mohkhedkar, Vani Janakiraman, and M Michael Gromiha, "Ab-CoV: a curated database for binding affinity and neutralization profiles of coronavirus-related antibodies," *Bioinformatics*, vol. 38, no. 16, pp. 4051–4052, jun 2022.

[13] Emily Engelhart, Ryan Emerson, Leslie Shing, Chelsea Lennartz, Daniel Guion, Mary Kelley, Charles Lin, Randolph Lopez, David Younger, and Matthew E. Walsh, "A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide," *Scientific Data*, vol. 9, no. 1, oct 2022.

[14] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal, "SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation," *Bioinformatics*, vol. 35, no. 3, pp. 462–469, jul 2018.

[15] Simone Conti and Martin Karplus, "Estimation of the breadth of CD4bs targeting HIV antibodies by molecular modeling and machine learning," *PLOS Computational Biology*, vol. 15, no. 4, pp. e1006954, apr 2019.

[16] Rishikesh Magar, Prakarsh Yadav, and Amir Barati Farimani, "Potential neutralizing antibodies discovered for novel corona virus using machine learning," *Scientific Reports*, vol. 11, no. 1, mar 2021.

[17] JunJie Wee and Kelin Xia, "Persistent spectral based ensemble learning (PerSpect-EL) for protein–protein binding affinity prediction," *Briefings in Bioinformatics*, vol. 23, no. 2, feb 2022.

[18] Yoichi Kurumida, Yutaka Saito, and Tomoshi Kameda, "Predicting antibody affinity changes upon mutations by combining multiple predictors," *Scientific Reports*, vol. 10, no. 1, nov 2020.

[19] Menglun Wang, Zixuan Cang, and Guo-Wei Wei, "A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation," *Nature Machine Intelligence*, vol. 2, no. 2, pp. 116–123, feb 2020.

[20] Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, Xuanling Shi, Qi Zhang, Bonnie Berger, Linqi Zhang, and Jian Peng, "Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization," *Proceedings of the National Academy of Sciences*, vol. 119, no. 11, mar 2022.

[21] Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas, "ISLAND: in-silico proteins binding affinity prediction using sequence information," *BioData Mining*, vol. 13, no. 1, nov 2020.

[22] Sandra Romero-Molina, Yasser B. Ruiz-Blanco, Joel Mieres-Perez, Mirja Harms, Jan Münch, Michael Ehrmann, and Elsa Sanchez-Garcia, "Ppi-affinity: A web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity," *Journal of Proteome Research*, vol. 21, no. 8, pp. 1829–1841, 2022, PMID: 35654412.

[23] Muhao Chen, Chelsea J T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang, "Multifaceted protein–protein interaction prediction based on siamese residual RCNN," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, jul 2019.

[24] Paola Ruiz Puentes, Laura Rueda-Gensini, Natalia Valderrama, Isabela Hernández, Cristina González, Laura Daza, Carolina Muñoz-Camargo, Juan C. Cruz, and Pablo Arbeláez, "Predicting target–ligand interactions with graph convolutional networks for interpretable pharmaceutical discovery," *Scientific Reports*, vol. 12, no. 1, may 2022.

[25] Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng, "Deep geometric representations for modeling effects of mutations on protein-protein binding affinity," *PLOS Computational Biology*, vol. 17, no. 8, pp. e1009284, aug 2021.

[26] Alex Morehead, Chen Chen, and Jianlin Cheng, "Geometric transformers for protein interface contact prediction," 2021.

[27] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang, "Protein representation learning by geometric structure pretraining," 2022.

[28] K. Yugandhar and M. Michael Gromiha, "Protein–protein binding affinity prediction from amino acid sequence," *Bioinformatics*, vol. 30, no. 24, pp. 3583–3589, aug 2014.

[29] Thomas P Quinn, "Graphdta post-hoc analysis scripts," 2020.

[30] Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li, "BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction," *Bioinformatics*, vol. 38, no. 7, pp. 1995–2002, jan 2022.

[31] Kanchan Jha, Sriparna Saha, and Hiteshi Singh, "Prediction of protein–protein interaction using graph neural networks," *Scientific Reports*, vol. 12, no. 1, may 2022.

[32] Sarah Sirin, James R. Apgar, Eric M. Bennett, and Amy E. Keating, "Ab-bind: Antibody binding mutational database for computational affinity predictions," *Protein Science*, vol. 25, no. 2, pp. 393–409, 2016.

[33] Puneet Rawat, Divya Sharma, R Prabakaran, Fathima Ridha, Mugdha Mohkhedkar, Vani Janakiraman, and M Michael Gromiha, "Ab-CoV: a curated database for binding affinity and neutralization profiles of coronavirus-related antibodies," *Bioinformatics*, vol. 38, no. 16, pp. 4051–4052, 06 2022.

[34] Hyejin Yoon, Jennifer Macke, Jr West, Anthony P., Brian Foley, Pamela J. Bjorkman, Bette Korber, and Karina Yusim, "CATNAP: a tool to compile, analyze and tally neutralizing antibody panels," *Nucleic Acids Research*, vol. 43, no. W1, pp. W213–W219, 06 2015.

[35] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane, "SAbDab: the structural antibody database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1140–D1146, 11 2013.

[36] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal, "SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation," *Bioinformatics*, vol. 35, no. 3, pp. 462–469, 07 2018.

[37] Yulin Shao, Soung Chang Liew, and Taotao Wang, "Alphaseq: Sequence discovery with deep reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3319–3333, 2020.

[38] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger, "Colabfold: making protein folding accessible to all," *Nature Methods*, vol. 19, no. 6, pp. 679–682, Jun 2022.

[39] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[40] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, "Graph attention networks," 2018.

[41] Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas, "Island: in-silico proteins binding affinity prediction using sequence information," *BioData Mining*, vol. 13, no. 1, pp. 20, Nov 2020.

[42] Min Li, Zhangli Lu, Yifan Wu, and YaoHang Li, "BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction," *Bioinformatics*, vol. 38, no. 7, pp. 1995–2002, 01 2022.

[43] Kanchan Jha, Sriparna Saha, and Hiteshi Singh, "Prediction of protein–protein interaction using graph neural networks," *Scientific Reports*, vol. 12, no. 1, pp. 8360, May 2022.

[44] Paola Ruiz Puentes, Laura Rueda-Gensini, Natalia Valderrama, Isabela Hernández, Cristina González, Laura Daza, Carolina Muñoz-Camargo, Juan C. Cruz, and Pablo Arbeláez, "Predicting target–ligand interactions with graph convolutional networks for interpretable pharmaceutical discovery," *Scientific Reports*, vol. 12, no. 1, pp. 8434, May 2022.

[45] Simone Conti, "ppdx," https://github.com/SimoneCnt/ppdx/tree/master/example-flu/ppdb.