Deep Geometric Framework to Predict Antibody-Antigen Binding Affinity

Nuwan Bandara[a,b,*], Dasun Prematilaka[a,**], Sachini Chandanayake[a,**], Sahan Hettiarachchi[a,**], Vithurshan Varenthirarajah[c], Aravinda Munasinghe[d], Kaushalya Madhawa[e], Subodha Charles[a]

[a]*Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka*
[b]*School of Computing and Information Systems, Singapore Management University, Singapore*
[c]*Faculty of Science, University of Colombo, Sri Lanka*
[d]*Independent,*
[e]*Graduate School of Bioengineering, University of Tokyo, Japan*

Abstract

In drug development, the efficacy of an antibody depends on how the antibody interacts with the target antigen. The strength of these interactions, measured through "binding affinity", gives an indication of how successful an antibody is in neutralizing an antigen. Due to the high computational complexity of traditional techniques for binding affinity quantification, deep learning is recently employed for the task at hand. Despite the commendable improvements in deep learning-based binding affinity prediction, such approaches are highly dependent on the quality of the antibody-antigen structures and they tend to overlook the importance of capturing the evolutionary details of proteins upon mutation. Further, most of the existing datasets for the task only include antibody-antigen pairs related to one antigen variant and, thus, are not suitable for developing comprehensive data-driven approaches. To circumvent the said complexities, we first curate the largest and most generalized (i.e., including a wide array of antigen variants) datasets for antibody-antigen binding affinity prediction, consisting of more than $100K$ sequence pairs, $8K$ structure pairs and the corresponding continuous bind-

*Corresponding author
**Authors contributed equally
Email address: pmnsribandara@gmail.com (Nuwan Bandara)

ing affinity values. Subsequently, we propose a novel deep geometric neural network comprising a structure-based model, which is to account atomistic-scale structural features, and a sequence-based model, which is to attribute sequential and evolutionary information, while sharing the learned information from each model through cross-attention blocks. Further, within each parallel model, we mimic the interaction space of antibodies and antigens through a set of multi-scale hierarchical attention blocks and the final latent vectors of each model are obtained by considering antibody and antigen representative vectors and the interaction vector. The proposed framework exhibited a 10% improvement in mean absolute error compared to the state-of-the-art models while showing a strong correlation (> 0.87) between the predictions and target values. Additionally, we extensively discuss the model optimization strategies, weight space analysis, and interpretability in a post-hoc fashion. We release our datasets and code publicly to support the development of antibody-antigen binding affinity prediction frameworks for the benefit of science and society.

Keywords:
Proteins, Antibody, Antigen, Binding Affinity, Deep Geometric Framework

1. Introduction

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, mode of action, transportation, etc. Small molecules are stable, synthetic chemicals with relatively simple structures. These have been around for decades and constitute the majority of the drugs currently in use. The body can absorb small molecules through oral uptake. Common small molecule drugs include "medicine cabinet" drugs such as aspirin and penicillin. 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. Biologics have high specificity in destination targeting, whereas small molecules may bind to off-targets and induce non-target harmful effects (i.e. 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 [1].

The efficacy of drugs designed using biologics (and even using small molecules) depends on how well they can bind or 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. The *binding affinity* reflects the strength of the interactions. The free energy associated with the binding of two molecules is called the binding free energy. Binding energy, which is generally a negative figure, is taken as a quantitative measure of the binding affinity. 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 towards their target.

In this work, we focus on antibody-antigen binding, which is a type of biologics interaction. Antigens are foreign substances that induce an immune response in a body, whereas antibodies are a part of the immune response produced to fight off such antigens. The extent of interactions between an antibody-antigen pair, evaluated through the binding affinity, decides the suitability of the drug in subsiding the relevant pathogenic condition. Here, we primarily consider the IC50 value, which is the concentration of the antibody required to inhibit the antigen activity by 50%, as the quantitative measure for the binding affinity. Further, we only focus on antibodies and antigens that are classified as proteins in this work.

Currently, molecular docking is used to study how two or more molecular structures fit together [2] through computer simulations. However, modeling 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. Molecular docking aims to mimic these natural interactions that take place during the binding via the lowest energy pathway. To achieve this, molecular docking calculates binding affinities at different binding poses to determine the optimum binding pose and binding affinity. However, the accuracy of molecular docking heavily depends on the ability of the utilized potential function to describe the forces in the system in a precise manner. In contrast, 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. Due to the over-dependency of MD simulations on the number of atoms in the proteins of concern, conducting MD simulations for large molecules remains a daunting task even to this date. To this end, recent advancements in deep learning have led researchers to use deep

learning models as a faster alternative to time-consuming MD simulations [3, 4]. But, the predictive performance of existing deep learning methods when calculating the binding affinity is highly dependent on the quality and resolution of the three-dimensional structures of the antibody and antigen.

To overcome the challenges of the above traditional methods, numerous deep learning-based methods have recently been utilized to predict the binding affinity. To this end, Wang et al. [5] introduced a method involving topology-based feature generation using element and site-specific persistent homology to capture the structural characteristics of 3D protein structures. This method employs a combination of convolutional neural network (CNN) and gradient boosting tree (GBT) model in which the CNN is utilized to extract more concise features from the topological descriptors, followed by the GBT as the prediction model. However, this approach is limited to predicting the effects of mutations within protein-protein complexes. In [6], the authors presented a geometric attention network that generates embeddings for mutation and wildtype 3D complexes, capturing residue information based on atom proximity while the attention is employed to identify crucial residue pairs at the protein interface for binding affinity. However, in this approach, capturing evolutionary details solely through 3D structures is ineffective and presents poor performance. Li et al. [7] proposed a bi-directional attention neural network for predicting compound-protein interactions and binding affinity. This method models compound-protein interactions as a continuous value prediction problem and employs graph neural networks (GNNs) to process structural data, alongside convolutional neural networks (CNNs) for processing sequence data. The network integrates both representations but has limitations in efficiently capturing residue features due to fixed windowing over protein sequences. More recently, DG-affinity [8] proposed to utilize large-language models to process amino-acid sequences to predict binding affinity, which still lacks the capability to capture the atomistic information through the structures for affinity prediction.

In this work, we introduce a novel deep-learning model that combines a geometric model utilizing graph convolution and graph attention operations to process the antibody-antigen structures and a sequence model utilizing self-attention and cross-attention to model the amino-acid sequences of the antibody-antigen pair. One major goal of the study was to develop a general deep model that is not confined to a specific family of antigens. Accordingly, the proposed network was trained on a curated dataset comprising antibody-antigen pairs for HIV, MERS, flu virus, etc. We observed a significant im-

provement in the mean absolute error compared to existing state-of-the-art models. Therefore, the key contributions of this work can be summarized as follows:

- We curate the largest and most generalized datasets for antibody-antigen binding affinity prediction in the literature, including both protein sequences and structures.
- We propose an end-to-end deep learning framework for antibody-antigen binding affinity prediction that combines a geometric model, which processes the atomistic-level structural details of the antibody-antigen pairs, and a sequence model, which processes the evolutionary details of the antibody-antigen pairs. Through extensive evaluations on the existing datasets and our curated datasets, we show that the proposed framework consistently outperformed the state-of-the-art methods by significant margins.
- We present a web-based platform where the users can obtain predictions for the binding affinity of their desired antibody-antigen pairs via uploading protein structure files. Further, we release the codes and datasets openly to support the development of antibody-antigen prediction frameworks for the benefit of science and society.

2. Methods

In this section, we delve into the details of the dataset curation process and our proposed models: sequence model and structure model followed by the model that combined both sequence and structure models. We 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 sequences of the input proteins, whereas the structure model treats the 3D structures of the proteins as graphs with nodes and edges. In addition to the sequence and structure models, the combined model has attention layers to share information between the said models.

2.1. Dataset Curation

Since publicly available datasets are tabulated in different formats, the first task associated with curating a generalized dataset was to process the datasets to have a similar format. 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 some public datasets. Homology modeling [9] and AlphaFoldV2 [10] were employed in this regard. The Table 1 presents a summary of publicly available datasets.

Dataset	**Raw 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	Available	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	K_d

Table 1: Summary of the used publicly available datasets. Here, $\Delta\Delta G$, IC50, EC50, IC80, ID50 and K_d refer to the change in the change in Gibbs free energy, half-maximal inhibitory concentration, half-maximal effective concentration, 80% inhibitory concentration, 50% inhibitory dose and protein-protein dissociation constant respectively.

Generally, experimental uncertainties can produce multiple IC50 values for a given antibody-antigen pair over several trials. To negate the impact of outliers, we first considered the median value of the provided IC50 values for repeated entries of antibody-antigen pairs. In contrast, we extensively discuss the structural modelling details in section 2.2 and Fig. 1 summarizes the complete set of processing steps associated with the dataset curation.

2.2. Structural Modelling

We aim to either retrieve (directly from the datasets mentioned in Table 1) or model (using the methods mentioned in sections 2.2.1 and 2.2.2) the 3D structures of both antibodies and antigens wherever applicable. However, it is to be noted that both experimental, such as from X-ray crystallography and cryo-EM, and computationally predicted structures represent static snapshots, often outside physiological conditions, and may not fully capture the conformational dynamics present in vivo [11]. Crystallographic structures, for instance, are typically determined in the presence of detergents and under non-physiological conditions, which can influence protein confor-

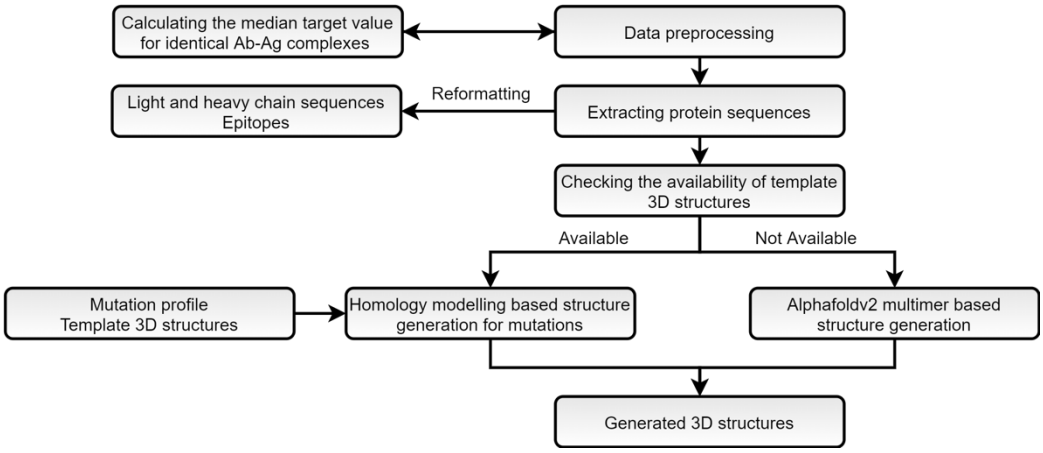


Figure 1: The generalized flowchart describing the steps followed to curate the mentioned datasets.

mation. Similarly, computational models are limited by the quality of input data and the assumptions of the algorithms used.

In our computational workflow, based on the availability of the template structure and the mutation profile, we then decided whether to use Homology Modelling or AlphaFoldV2 to generate the 3D structure. If the mutation profile, along with the template structure, was provided, then we utilized homology modeling. If either of the two requirements was not satisfied, then we used the AlphaFoldV2 pipeline.

2.2.1. Homology Modelling

Homology modeling is a reasonably accurate comparative protein structure modeling technique in which an atomic-resolution model of the target protein is constructed using its amino-acid sequence and an experimental 3D structure of a related homologous protein [12]. We utilized a homology modeling-based pipeline when the template sequence (original amino acid sequence) and its 3D structure (PDB format) were available along with the mutation profile. First, we created the mutated protein sequence based on the mutation profile provided for the template sequence using a package called MDAnalysis [13]. Homology Modelling was then employed to generate the 3D structure (in PDB format) of a mutated sequence (i.e. target sequence) using the template PDB structure. For the sequence alignment task between the target sequence and template sequence, a Python package called Bio-python [14] was used instead of the default sequence alignment

function available in the MODELLER for better results.

The first stage in structure derivation was 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 was followed by loop modeling, which performs necessary conformational adjustments to the modeled backbone. Finally, side chains were modeled. To alleviate steric collisions, the estimated relative atomic locations were fine-tuned to minimize the potential energy of the conformation of the protein. This was done during the model optimization step. At last, the optimized model was evaluated. These steps were conducted using 'MODELLER' [15] freeware program.

Further, to improve the accuracy and reliability of the generated structures, we consider multiple templates, wherever possible, which is particularly useful in regions where individual templates may lack coverage or accuracy. Further, to ensure both local and universal structural quality, we carefully align and rank the models using the well-accepted DOPE score to select the most reliable and accurate structure [16].

2.2.2. AlphaFold-V2

AlphaFold-V2 multimer model [10] is a state-of-the-art model with atomistic level accuracy for protein structure prediction from amino acid sequences, which is proven to be useful even in the absence of a homologous structure, as shown by their superior performance at the 14^{th} Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction.

In our implementation for generating the 3D structures of proteins where the template structures were absent, we followed the pipeline from ColabFold [17], which runs inference using the AlphaFold-V2 along with an accelerated combination of MMseqs2-based homology search where MMseqs2 [18] refers to Many-against-Many-searching which is widely used to search and cluster sequences.

Similar to homology modelling pipeline, we implement the following steps to improve the accuracy and reliability of structural modelling in AlphaFold-V2 pipeline. In summary, we leverage multiple templates, generate several models and select the top-ranked model based on the defined scores. To be specific, here, using the default settings of ColabFold [17], we generate multiple models (i.e., five per prediction) which provide per-residue and global confidence metrics: pLDDT and pTM scores, to assess model reliability. We

select the top-ranked model based on these scores, as they have been shown to correlate well with experimental accuracy [10].

2.3. Sequence-based Model

The input protein files were processed to obtain the FASTA sequences. Since deep learning models require a numerical input we encode the complete FASTA sequence (i.e., to be specific, not just complementarity-determining regions) using a numerical scheme such as one-hot, BLOSUM, and VHSE8 Our motivation behind the selection of these encoding schemes is in the section, *Ablations on Sequence Encoding Schemes* in the appendix.

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

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

where $\mathbf{z_i} \in \mathbb{R}^d$ is the output vector after the attention layer, s is the sequence length (antibody pathway, $s = s_1$ and antigen 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 $\mathbf{x_i} \in \mathbb{R}^d$ being the input vector and $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$. The softmax operator is defined as,

$$\text{softmax}\left(\frac{< \mathbf{q_i}, \mathbf{k_j} >}{\sqrt{d}}\right) = \frac{\exp\left(\frac{\mathbf{q_i}^T \mathbf{k_j}}{\sqrt{d}}\right)}{\sum_{l=1}^{s} \exp\left(\frac{\mathbf{q_i}^T \mathbf{k_l}}{\sqrt{d}}\right)} \tag{2}$$

Along with self-attention/multi-head attention layers which are supposed to extract a learned latent representation of the antibody and antigen separately, cross-attention (CA) layers were also implemented to pass information from the antibody pathway to the antigen pathway and vice versa to imitate the antibody-antigen interaction space. As shown in Fig. 2, we

included two pathways for cross-attention to mimic the 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 antigen pathway to the antibody pathway. $\therefore \forall i = 1, 2, 3, \ldots, s_1$;

$$\mathbf{z_{CA,i(Ab)}} = \sum_{j=1}^{s_2} \text{softmax}\left(\frac{< \mathbf{q_{CA,i(Ab)}}, \mathbf{k_{CA,j(Ag)}} >}{\sqrt{d_{CA}}}\right) \mathbf{v_{CA,j(Ag)}} \tag{3}$$

where $\mathbf{z_{CA,i(Ab)}} \in \mathbb{R}^{d_{CA}}$ is the output antibody vector after the cross-attention layer. Moreover, $\mathbf{q_{CA,i(Ab)}} = W_{CA,q}\mathbf{x_{CA,i(Ab)}}$, $\mathbf{k_{CA,j(Ag)}} = W_{CA,k}\mathbf{x_{CA,i(Ag)}}$ and $\mathbf{v_{CA,j(Ag)}} = W_{CA,v}\mathbf{x_{CA,i(Ag)}}$ with $\mathbf{x_{CA,i(Ab)}}$, $\mathbf{x_{CA,i(Ag)}} \in \mathbb{R}^{d_{CA}}$ being the input antibody latent vector and antigen latent vector, which are obtained after the first few non-linear layers, respectively and $W_{CA,q}, W_{CA,k}, W_{CA,v} \in \mathbb{R}^{d_{CA} \times d_{CA}}$. Other symbols hold the same meaning as in the self-attention layer. A similar equation can be provided for the flow of information from the antibody pathway to the antigen pathway by interchanging 'Ab' (which refers to antibody) and 'Ag' (which refers to antigen) subscripts in Equation 3 and changing the upper limit of summation to s_1 instead of s_2.

The outputs from the antibody pathway, antigen pathway, and antibody-antigen pathways were then concatenated into a single vector and further processed through two dense layers to get the output.

2.4. Structure-based Model

Even though antibodies undergo conformational changes, upon binding with antigens, leading to complex structures capturing the bound state (i.e., induced fit) [19], most antibody-antigen complexes lack experimentally resolved structures, thus, a limited information on binding sites for many datapoints, necessitating reliance on high-confidence predicted models of unbound components [20] for a data-driven study. Therefore, despite the potential information loss, here we utilized individual structure details of antibodies and antigens rather than the antibody-antigen complex structure details. Our decision is further motivated by the following findings in the literature as well: (1) unbound antibody and antigen structures are often used to simulate real-world scenarios where pre-binding conformations are unknown, like in docking tools with ensemble of unbound structures [20]. Metrics like root mean square deviations in complementarity determining region and epitope

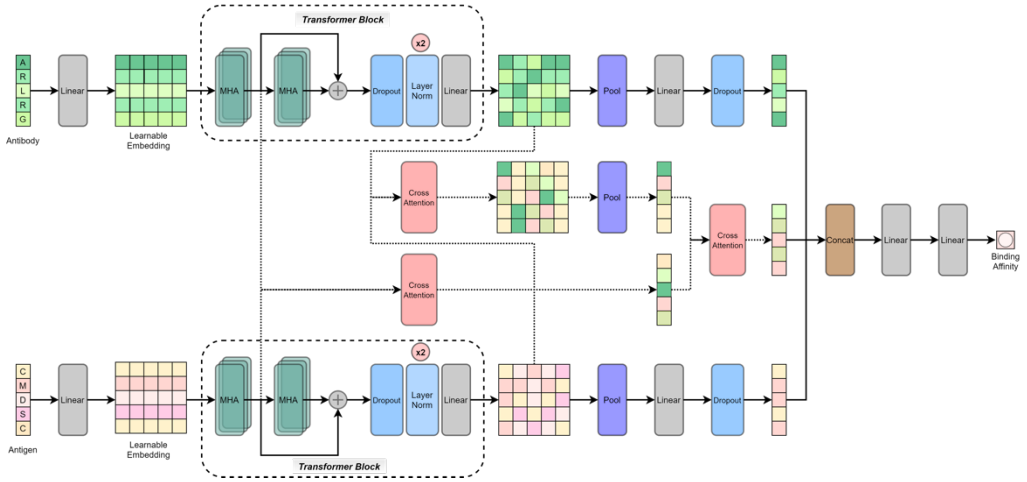


Figure 2: The network architecture for the sequence-based model. *MHA refers to Multi-head Attention and other layers in the diagram hold conventional meanings as used in deep learning.

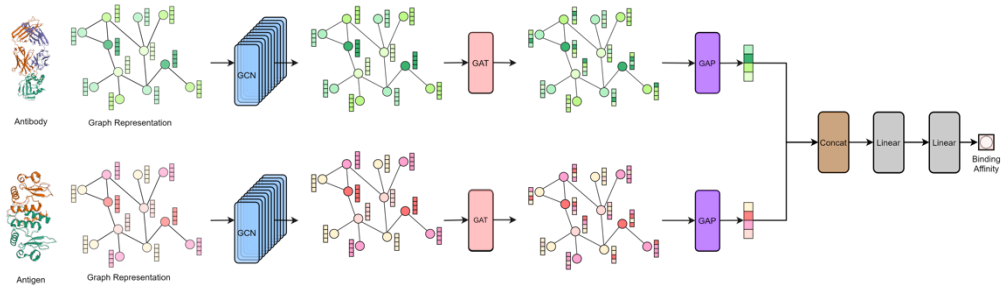


Figure 3: The network architecture for the structure-based model. *Here, GCN, GAT, and GAP refer to Graph Convolution, Graph Attention, and Graph Average Pooling, respectively.

in these unbound models correlate with docking success, emphasizing the sufficiency of high-quality individual structures for capturing key interaction motifs [21], and (2) simple interaction features such as contact counts derived from unbound structures achieve classification performance comparable to complex structural descriptors, suggesting the information loss is mitigated by feature engineering focused on interaction hotspots [22].

In our implementation, we represented both antibody and antigen molecules as graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that use atoms as nodes with their respective 3D coordinates denoted as $X \in \mathbb{R}^{3\times n}$, and initial atomic features $F \in \mathbb{R}^{4\times n}$. 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 which we calculate as the reciprocal of bond distance.

The structure model architecture comprises two parallel branches that simultaneously process antibody and antigen graph representations, as depicted by Fig. 3. Each branch consists of 4 graph convolutional layers [23], followed by four graph attention layers [24] and a graph average pooling layer. To aggregate the features of neighboring atoms, we passed the graph representations of antibody and antigen molecules through a stack of graph convolutional layers. The core property of the graph convolutional layer is that it takes the weighted average of neighbours' node and edge features, including itself. A single graph convolutional layer, followed by a ReLU non-linear layer, is denoted as follows.

$$Z' = \mathrm{ReLU}\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}ZW_{GCN}\right) \tag{4}$$

where $Z \in \mathbb{R}^{8\times n}$ 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 $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. Moreover, W_{GCN} is the weight matrix and I_N is the identity matrix. Following the stack of graph convolutional layers, a stack of four graph attention layers was applied to identify the atoms that should be given more priority. The graph attention operator works as follows.

$$Z'_i = \alpha_{i,i}W_{GAT}Z_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}W_{GAT}x_j \tag{5}$$

$$\alpha_{i,j} = \frac{\exp\left(\mathrm{LR}\left(a^\top\left[W_{GAT}Z_i \,\|W_{GAT}Z_j\|\, W_e e_{i,j}\right]\right)\right)}{\sum\limits_{k\in\mathcal{N}(i)\cup\{i\}} \exp\left(\mathrm{LR}\left(a^\top\left[W_{GAT}Z_i \,\|W_{GAT}Z_k\|\, W_e e_{i,k}\right]\right)\right)} \tag{6}$$

where Z_i is the input feature vector of the i^{th} node, W is the weight matrix and LR denotes the LeakyReLU non-linear function. The coefficients computed by the attention mechanism, $\alpha_{i,j}$, in Eq. 5 are computed by following the original GAT paper [24] as shown in Eq. 6. To this end, a denotes the attention mechanism which is a single-layer feedforward neural network parametrized by a weight vector $\in \mathbb{R}^{2F'}$ where F' is the cardinality of node features after the layer in consideration. W_e and $e_{i,j}$ are the edge feature weight matrix and edges in consideration.

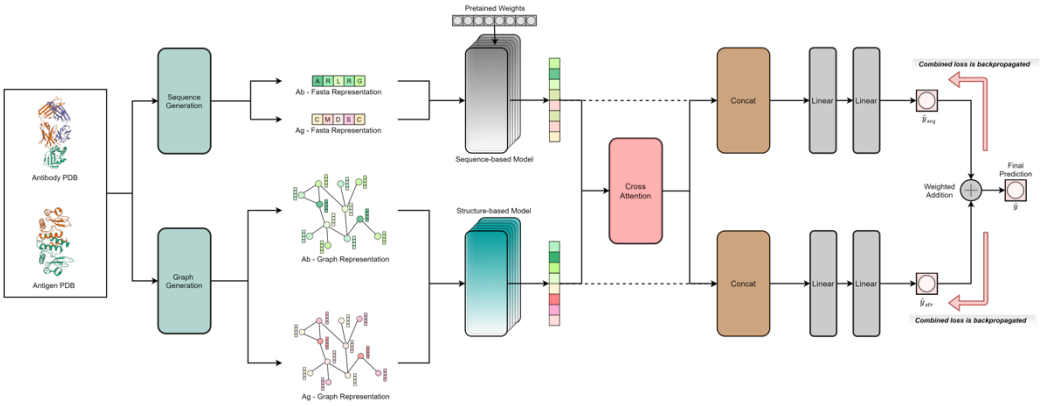


Figure 4: The network architecture for the combined model.

Afterwards, we fed the graph attention layer output to a graph average pooling layer to reduce the spatial dimension. Subsequently, embeddings of the antibody and antigen obtained via each pathway were concatenated and processed further through two linear (dense) layers to retrieve the binding affinity prediction from the structure model.

2.5. Combined Model

With the expectation of capturing evolutionary details through protein amino acid sequences and atomistic details via protein 3D structures to enhance the prediction performance, the final combined model constitutes the aforementioned sequence and structure models. We applied cross-attention to propagate complementary information between the two domains. As indicated by Fig. 4, the cross-attention output was concatenated separately with the embeddings from each constituent model and passed through linear (dense) layers to obtain two intermediate binding affinity predictions ($\hat{y}_{seq}$ and $\hat{y}_{str}$). The final binding affinity prediction ($\hat{y}$) is a weighted average between those two predictions, and the weights were decided based on the conducted ablation study.

$$\hat{y} = (\delta_1 \times \hat{y}_{str}) + (\delta_2 \times \hat{y}_{seq}) \tag{7}$$

Here, we used a weighted average function with hyperparameters; δ_1 and δ_2. Through ablations, δ_1 and δ_2 were selected to be 0.45 and 0.55, respectively. During the training phase of the combined model, for a given ground truth value (y), the following combined loss function ($\mathcal{L}_{total}$) was calculated, and the

error was backpropagated, updating the weights in the constituent sequence and structure models, simultaneously.

$$\begin{aligned}\mathcal{L}_{total} = \lambda \times MSE(y, \hat{y}_{str}) + \mu \times MSE(y, \hat{y}_{seq}) \\ + \nu \times MSE(\hat{y}_{str}, \hat{y}_{seq})\end{aligned} \tag{8}$$

where λ, μ and ν are hyperparameters selected through ablation studies. The intermediate binding affinity predictions from the sequence and structure models are denoted by $\hat{y}_{seq}$ and $\hat{y}_{str}$, respectively.

2.6. Evaluation Metrics

Throughout the experiments, the following metrics were utilized to evaluate the performances and compare the results between the models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

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

The mean squared error (MSE) was mainly used as a loss function and on certain occasions, it was also considered as an evaluation metric to further validate the results.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

Further, to evaluate the correlation between the true values and predicted values from our models, concerning linearity and ranks, Pearson correlation coefficient (r) and Spearman's correlation coefficient (ρ) were utilized respectively.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y_i}) \times (\hat{y}_i - \bar{\hat{y}_i})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \times \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}_i})^2}} \tag{11}$$

$$\rho = 1 - \frac{\sum_{i=1}^{n} 6(d_i^2)}{n(n^2-1)} \tag{12}$$

In Equation 12, d_i refers to the difference between the two ranks of each true and predicted value pair.

3. Results

In this section, we discuss in detail our experiments and results in comparison with several baseline models from the literature. The impact of incorporating both evolutionary and atomistic-level details through sequence-based and structure-based models, respectively, is exhibited in the results. The results are presented and discussed in the following order of subsections: Dataset Curation, Combined Models, Sequence-based Models, and Structure-based Models.

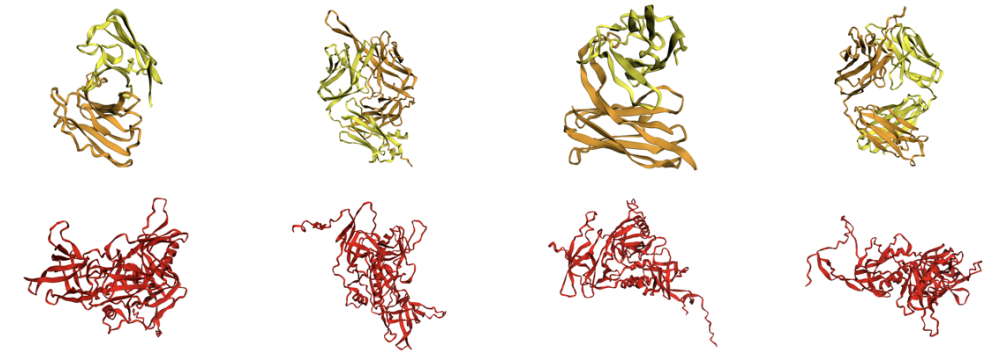
3.1. Dataset Curation

As publicly available datasets are compiled in different formats, they were initially preprocessed through custom pipelines and the resulting curated sequence and structure datasets were named *P2PXML-Seq* and *P2PXML-PDB*, respectively. A detailed comparison of the curated datasets with the publicly available datasets is provided in Table 2. P2PXML-Seq only contains antibody-antigen pairs in the respective amino acid one-letter code (FASTA format [25]) while the P2PXML-PDB contains 3D structures of the antibody-antigen pairs in protein data bank (PDB) format as illustratively presented in Figure 5. Both datasets comprise the IC50 values of the corresponding antibody-antigen pairs, which are either experimental or estimated.

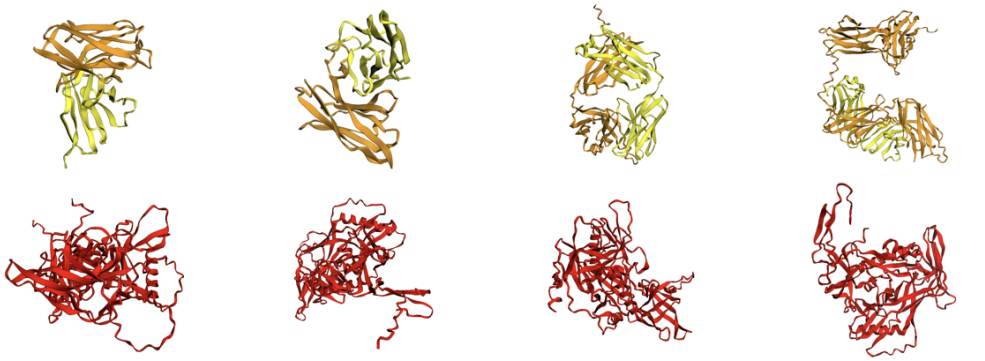
Dataset	Usable Datapoints*
Ab-Bind [26]	1 101
Ab-CoV [27]	1 420
CATNAP [28]	11 208
AlphaSeq [29]	87 808
P2PXML-Seq (Ours)**	**111 845**
P2PXML-PDB (Ours)**	**8 475**

Table 2: Dataset comparison. *Here, usable datapoints refer to the remaining datapoints after a defined set of preprocessing steps, including the duplicate removal and the exclusion of datapoints with no numerical value for binding affinity. **P2PXML-Seq and P2PXML-PDB are our curated protein sequence and structure datasets, respectively.

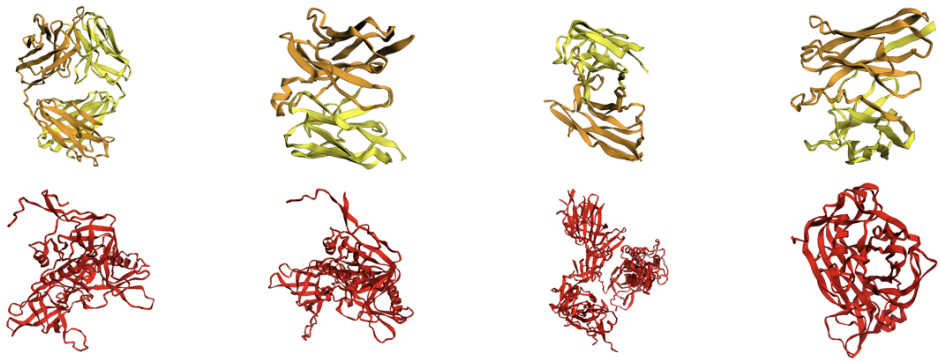
As per our knowledge, the datasets in Table 2 are the most extensive datasets curated explicitly for the antibody-antigen binding affinity prediction. Our datasets have a consistent format concerning the antibody-antigen



(a) 110.00 (119.54) (b) 0.01 (0.012) (c) 35.88 (46.80) (d) 1.86 (2.32)



(e) 19.00 (15.75) (f) 0.001 (0.002) (g) 8.24 (6.05) (h) 0.007 (0.024)



(i) 18.00 (22.10) (j) 50.00 (36.24) (k) 40.00 (32.07) (l) 52.40 (56.52)

Figure 5: py3Dmol [30]-based visualization of a randomly selected subset of our P2PXML-PDB dataset. Here, in each sub-figure, the top protein structure is the antibody while the below is the antigen. The caption for each antibody-antigen pair presents the corresponding binding affinity in IC50 value whereas the predictions from our model is mentioned within brackets.

pairs and the corresponding numerical values (i.e. IC50; half-maximal inhibitory concentration). It is to be noted that we assumed specific parameters or conditions including non-competitive inhibition, following [31, 32], to approximate IC50 whenever needed. As shown in Figure 6, IC50 distribution in our curated datasets densely covers the affinity variation in antibody-antigen pairs while closely mapping to normal distributions. Moreover, the curated datasets express sufficient generalizability since they contain numerous antigens such as SARS-CoV-2, HIV, MERS, and flu, as shown in Figure 7, and their related antibodies, in contrast to many datasets in the literature where only one antigen is considered [29, 27].

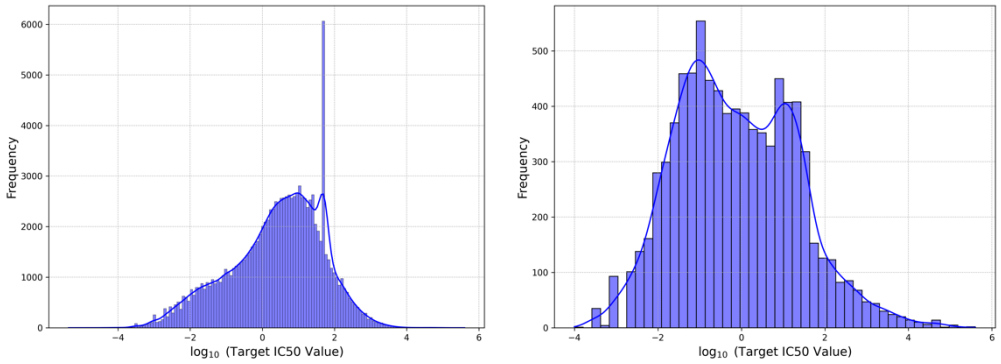


Figure 6: Distribution of binding affinity values (i.e., here, IC50 values) in the curated (a) P2PXML-Seq dataset (mean: 0.4070, standard deviation: 1.2916, min: −5.3665, max: 5.6021 in log scale) and (b) P2PXML-PDB dataset (mean: −0.1196, standard deviation: 1.4454, min: −4.0000, max: 5.6021 in log scale). Here, the IC50 values (in x-axis) are presented in log scale for better visualization.

3.2. Combined Models

The primary intuition behind the combined model is to incorporate the evolutionary and atomistic-level details of antigens and antibodies through the sequence-based model and the structure-based model, respectively, while sharing the information learned through each pipeline to imitate the chemical binding potential. Further, as depicted in Figure 8, it is evident that antibodies corresponding to a particular antigen variant do not exhibit distinctive clusters in low-dimensional space in both protein sequences and structures which further emphasizes the importance of modeling an information sharing

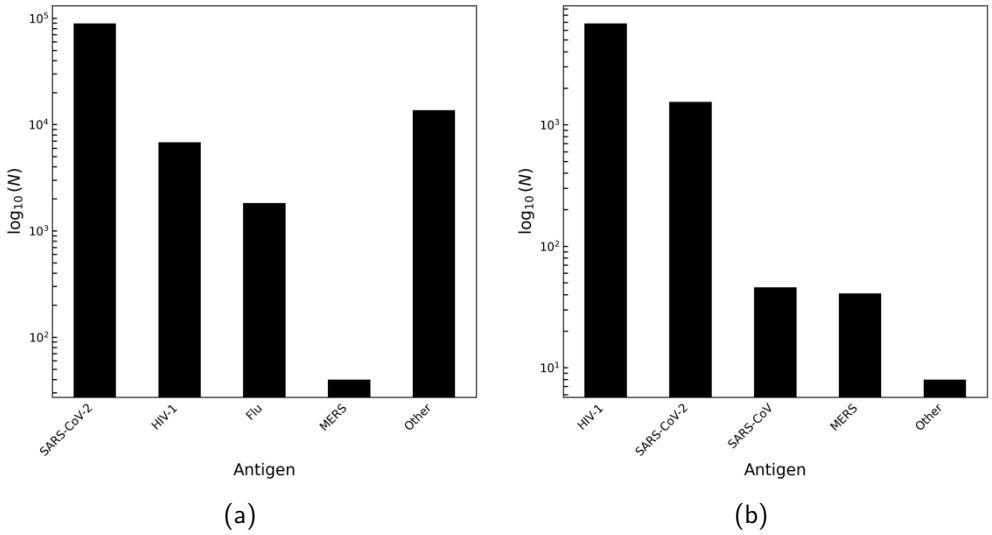
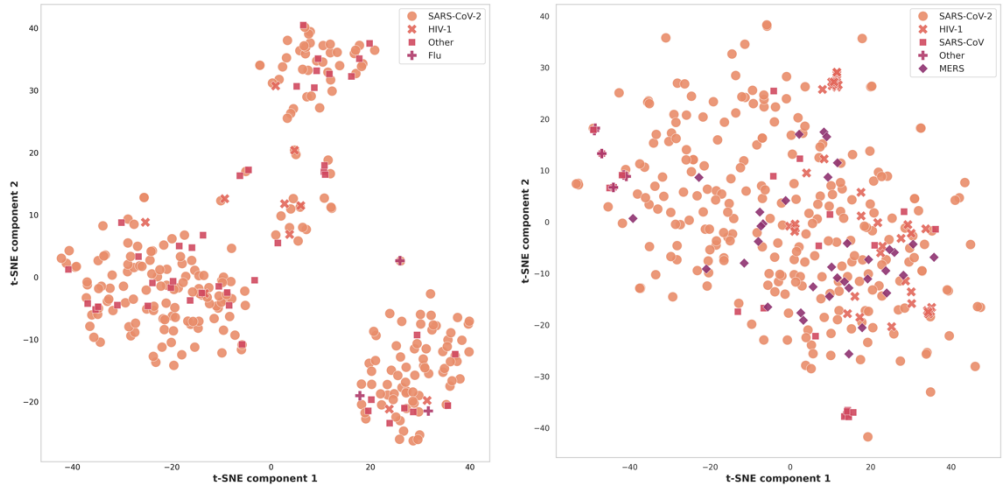


Figure 7: The number of antibody-antigen pairs (in log scale) vs the type of antigen strain in the curated (a) P2PXML-Seq dataset and (b) P2PXML-PDB dataset. Here, N refers to the number of antibody-antigen pairs and "Other" refers to antigen strains, which do not correspond to a specific bar in the plots, such as WIV1 [33] and SHC014 [34].

pipeline between antibodies and antigens while utilizing both their sequences and structures.

The curated P2PXML-PDB dataset was used to train, test and compare the performances of the models. As the proposed network comprises a sequence model in conjunction with a structure model, the results exhibited by the proposed network were compared with state-of-the-art sequence-based and structure-based models.

From Table 3, it is evident that our final Combined-V2 model outperforms all the considered state-of-the-art approaches at least by a margin of 10.6% while improving the performance of our individual sequence-based and structure-based models by 5.6% and 6.8% (in terms of MAE), respectively. The performance of the proposed combined model underscores the importance of incorporating both atomistic-level and evolutionary details in predicting the binding affinity. In addition, to ensure the statistical significance of the reported results, a paired t-test [39] and a Wilcoson signed-rank test [40] were conducted between the target and predicted IC50 values. The paired t-test resulted in a p-value of 0.000163, suggesting that the differ-



(a) (b)

Figure 8: Component 2 vs component 1 of the low-dimensional t-distributed stochastic neighbor embeddings [35] reduced from the (a) amino-acid sequence embeddings generated by following ESM-2 [36]. The sequences are from a randomly-sampled subset of the antibody amino-acid sequences in the P2PXML-Seq dataset. (b) whole-graph embeddings of the constructed graphs. The graph embeddings are generated by following graph2vec [37] and the graphs are corresponding to a randomly-sampled subset of the antibody protein structures in the P2PXML-PDB dataset. Here "Other" follows the same definition as in Figure 7.

ence between the predicted and target IC50 values is statistically significant (specifically in regard to mean difference) while the Wilcoxon signed-rank test yielded a p-value of 0.014451, re-confirming the significance of the differences observed, specifically in regard to the distribution of differences.

As depicted in Figure 9, the correlation between the predicted and target values is significantly strong: the linear correlation evaluated through Pearson correlation coefficient being 0.8703 and the rank-order correlation evaluated through Spearman's correlation coefficient being 0.9450 and concordance index being 0.8560, which further highlights the better performance of our approach. This is a significant improvement as prior works [7] and [38] only exhibit Pearson correlation coefficients of 0.81 and 0.75 along with the Spearman's correlation coefficients of 0.83 and 0.78 respectively. Furthermore, as illustrated in Figure 10a, the errors are randomly distributed around zero while as in Figure 10b, the error distribution is approximately

Model	MAE	MSE
1D CNN with attention [7]	1.1191	5.2266
GCN [38]	1.1348	5.2988
Parallel Transformer with cross-attention (Ours)	1.1083	5.1986
Parallel GCN + GATConv (Ours)	1.1338	5.3457
Combined-V2 + pre-trained weights (Ours)	**1.0005**	**4.6709**

Table 3: Overall results comparison for P2PXML-PDB dataset with respect to mean absolute error (MAE) and mean squared error (MSE). Here, '1D-CNN with attention' and 'GCN' refer to the respective state-of-the-art sequence and structure models whose performances are compared with our sequence-based pipeline (Parallel Transformer with cross-attention), our structure-based pipeline (Parallel GCN + GATConv) and the final combined model containing both pipelines (Combined-V2 + pre-trained weights). Only the best-performing structure and sequence models out of the considered work found in the literature are given. Other related research is mentioned under the subsections; Sequence-based Models and Structure-based Models.

normal, with a mean close to zero, which collectively imply that the proposed model does not exhibit significant bias in its predictions.

In our experiments, several variants of the combined model were tested along with two different amino acid encoding schemes. An overview of variations and performance comparison between the combined model variants are given below. See section 2.5 for more details.

- Combined-B: This refers to the combined base model where the outputs from the sequence-based model and the structure-based models are collectively considered for calculating the combined loss function.

- Combined-V1: The Combined-B model was modified such that the output latent vectors from sequence-based and structure-based models were concatenated. Then the concatenated vectors were propagated through parallel paths to obtain separate outputs which were eventually combined to calculate the combined loss function.

- Combined-V2: In contrast to the direct concatenation of latent vectors from two pathways as mentioned in the Combined-V1 model, here, the resulting vector from cross-attention between the two pathways was concatenated separately with each latent vector from two mod-

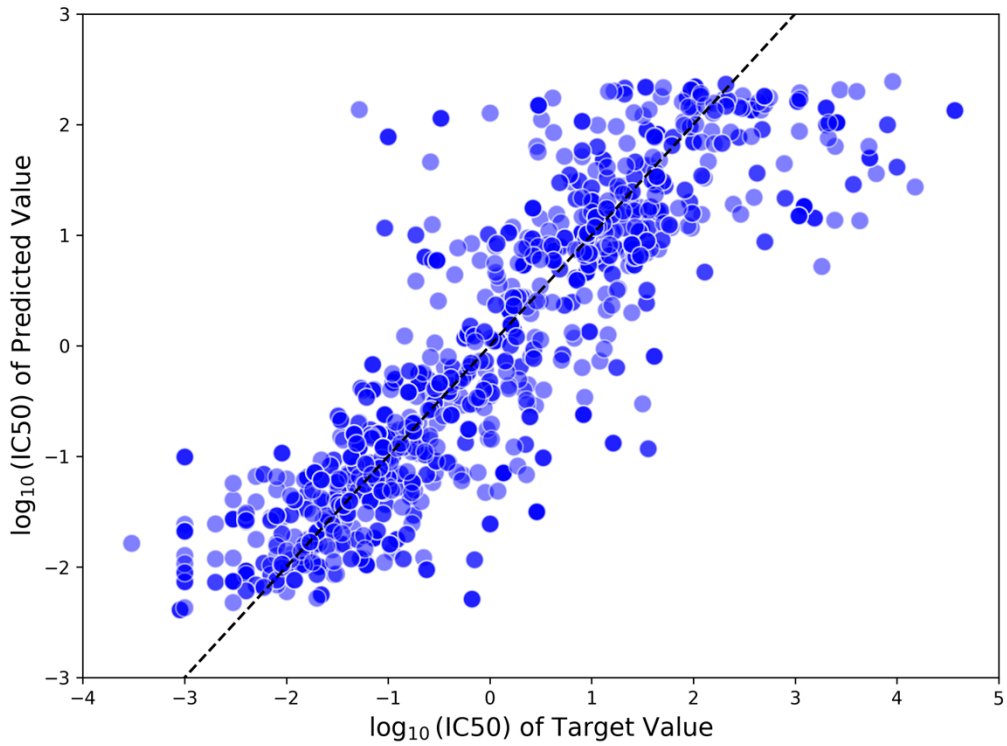


Figure 9: $\log_{10}(IC50)$ of the predicted values vs $\log_{10}(IC50)$ of the target values for the test set of P2PXML-PDB dataset using our best performing model. The Pearson correlation coefficient is 0.8703 and the Spearman's correlation coefficient is 0.9450 between the predicted and target values. The coefficient of determination is 0.7515.

els. The resulting concatenated vectors were processed similarly as in Combined-V1. See Figure 4.

- Combined-V2 + pre-trained weights: Instead of randomly initializing the weights of the sequence model in the Combined-V2 model, pre-trained weights obtained from training the sequence model using the P2PXML-Seq dataset were utilized. Note that the weights in the structure model of the Combined-V2 model were still initialized randomly. Since the Combined-V2 model was trained on the P2PXML-PDB dataset, pre-trained weights were not used for the structure-based counterpart in the Combined-V2 model as it would have otherwise pre-exposed the dataset to the model.

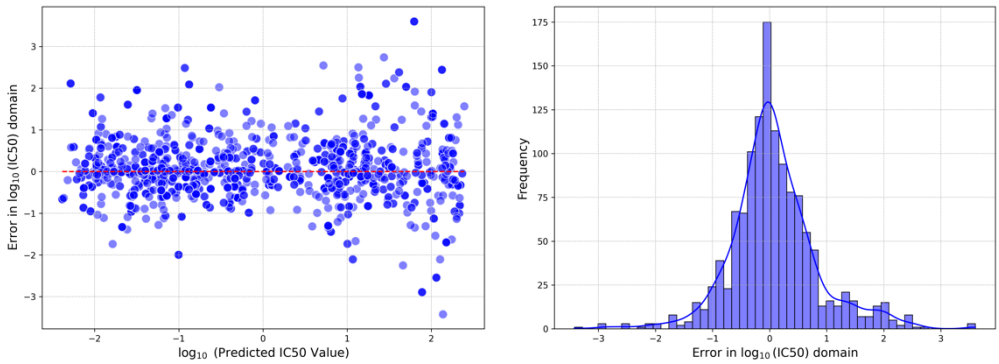


Figure 10: Error analysis for the test-set of P2PXML-PDB dataset (a) Error (i.e., $\log_{10} IC50_{target} - \log_{10} IC50_{predicted}$) vs $\log_{10} IC50_{predicted}$ (b) Histogram of the error (in log scale) distribution. The standard deviation of the errors is observed to be 0.77.

Model	MAE	
	VHSE8 encoding*	Protein language embeddings**
Combined-B	1.0567	1.1209
Combined-V1	1.0039	1.0988
Combined-V2	1.0007	1.0951
Combined-V2 + pre-trained weights	**1.0005**	**1.0924**

Table 4: Results comparison between the principal component score vector of hydrophobic, steric, and electronic properties (VHSE8) encoding and pre-trained protein language embeddings from ProtT5-XL-BFD [41] for protein amino-acid sequences. The dataset used is our P2PXML-PDB dataset and the performance parameter is MAE. *VHSE8 encoding of an amino acid is an 8D Vector of Hydrophobic, Steric, and Electronic properties. **Obtained from ProtT5-XL-BFD which is a pretrained model on protein sequences.

As given in Table 4, the Combined-V1 model outperformed the Combined-B model by a margin of 5% which highlights the impact of sharing evolutionary and atomistic details between sequence-based and structure-based pipelines. Applying cross-attention when sharing information between the sequence and structure models in Combined-V2 further improved the performance of Combined-V1, validating the importance of learning the extent to which information should be shared. At last, employing pre-trained weights,

as described earlier, slightly improved the results further to obtain the best MAE of 1.0005.

It can be deduced from Table 4 that the same pattern of performance improvement was perceived when utilizing the protein language embeddings from ProtT5-XL-BFD [41] instead of the principal component score vector of hydrophobic, steric, and electronic properties (VHSE8) encoding scheme to encode the protein amino-acid sequences. However, the superior performance of VHSE8 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 utilized the language embeddings from a pre-trained model rather than a model which was fine-tuned to the task in hand. We further believe that if we fine-tune the protein language model to our task, it might have a comparable or superior performance than the traditional encoding scheme since it could learn an enriched protein sequence representation tailored to the task rather than an ill-posed generalized representation which would not be sufficient to obtain the best performance.

3.3. Sequence-based Models

The sequence-based model was 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 [43, 7], which utilize protein sequences for the binding affinity prediction.

As per the results in Table 5, the sequence-based model with the best performance is the 'Parallel Transformer with cross-attention' model. The results indicate that the selected sequence-based model surpasses the state-of-the-art approaches under all three datasets by MAE margins of 4%, 44%, and 1% for VirusNet, P2PXML-Seq, and P2PXML-PDB datasets, respectively. Consequently, it was incorporated into the combined model to extract the information from protein amino acid sequences.

3.4. Structure-based Models

The structure-based model was developed to capture atomistic-level and residue-level structural information of the antibodies and antigens to facilitate the antibody-antigen binding affinity prediction. Based on this intuition,

Model	Parallel Pipeline	Cross attention	MAE		
			VirusNet [42]	P2PXML-Seq	P2PXML-PDB
SVR with RBF kernel [43]	✗	✗	102.1267	10.8236	2.4181
1D CNN with attention [7]	✓	✓	3.5124	0.0964	1.1190
LSTM	✓	✗	5.2853	0.2460	1.4448
LSTM	✓	✓	3.6288	0.0926	1.1275
Transformer	✓	✗	3.4995	0.0898	1.1175
Parallel Transformer with cross attention	✓	✓	**3.3707**	**0.0542**	**1.1083**
Transformer with multiple cross-attention*	✓	✓	3.4744	0.0553	1.1084
Transformer with distogram**	✓	✓	3.3882	0.0551	1.1108
Transformer with protein language embeddings***	✓	✓	3.4263	0.0623	1.1128

Table 5: Results comparison between the protein sequence-based models with MAE as the performance parameter. The datasets used are VirusNet [42], P2PXML-Seq and P2PXML-PDB. Apart from 'SVR with RBF kernel' and '1D CNN with attention' models, the other models were developed during our study to check the impact of different architectures on the MAE. *Here, 'Transformer with multiple cross-attention' refers to having cross-attention blocks in each stage, in addition to the two hierarchical cross-attention layers as in 'Parallel Transformer with cross-attention'. **In the 'Transformer with distogram' model, the calculated distograms were used as input instead of the encoded protein sequences. ***The pre-trained protein language embeddings generated through ProtT5-XL-BFD [41] (without fine-tuning the model to our task) were used as input to the transformer instead of encoded protein sequences.

several approaches were tested and compared with the state-of-the-art approaches [44, 38], which utilized protein 3D structures for the binding affinity prediction.

As per the results in Table 6, the final structure-based model is selected to be the 'Parallel Graph Convolution (GCN) + Graph Attention (GATConv)' model. The results depict that the proposed structure-based model has surpassed the state-of-the-art methods under our benchmark PDB dataset by a slight MAE margin of 0.08% and an MSE margin of 0.04%. Accordingly, it was integrated into the combined model to extract information from protein 3D structures.

4. Conclusions

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. Therefore, accurate

Model	MAE	MSE
Parallel GNN [44]	1.1633	5.5406
Parallel GCN [38]	1.1347	5.3490
Parallel GAT* (Ours)	1.1788	5.6719
Parallel GCN + cross-attention** (Ours)	1.1409	5.3818
Parallel GCN + GATConv (Ours)	**1.1338**	**5.3457**

Table 6: 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 was developed under the hypothesis that identifying the most essential nodes through attention would be sufficient for better binding affinity prediction. However, the performance of the model indicates that such information alone is inadequate to yield better performance. **The parallel GCN model with cross-attention was 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. Surprisingly, it was not as useful in the structure-based model as it was in the sequence-based model. We hypothesize that the local aggregation of the node features in the intermediate stages lacks or does not encapsulate sufficient global information, which is essential for the graph-level prediction task.

binding affinity prediction is critical in designing drugs with higher specificity towards the target. Our study focused on antibody-antigen binding, which is a subclass of protein-protein interactions. 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 Molecular Dynamics). Even though there is a trending interest in utilizing machine learning to predict binding affinity, the predictive performance of existing machine learning methods when calculating the binding affinity is highly dependent on the quality of the antibody-antigen 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 proposed a novel deep geometric network that comprises a structure 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 both atomistic-level infor-

mation and evolutionary details are appropriately incorporated into neighbor embeddings and/or between the pipelines for antibodies and antigens. The proposed model was trained on our curated dataset, which consists of sequences and structures of antigens and antibodies corresponding to a diverse set of common viruses such as Human Immuno-deficiency Virus (HIV), SARS-CoV-2, etc., to ensure sufficient generalizability within the dataset.

After extensive ablation studies which were performed to select the encoding schemes, node and edge features, and model hyperparameters, it was observed that the Combined-V2 model, which is our final model architecture including selected 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. Furthermore, the stand-alone sequence-based model was able to surpass the existing sequence-based state-of-the-art methods under all benchmark datasets, and our structure-based model marginally outperformed the existing works in the literature as well. Moreover, we developed a website as a community access tool to allow the interested community to obtain prediction results for their input proteins via our hosted models.

In summary, we believe that the work presented in this study would be beneficial in improving deep-learning-based binding affinity prediction of antibody-antigen pairs, especially in the domain of drug development.

Acknowledgements

We are grateful to Dr. Ranga Rodrigo, Former Head of the Department of Electronic and Telecommunication Engineering for allocating us GPUs which significantly accelerated the protein structure generation process and training the deep learning models. We would also like to extend our gratitude to other Senior Lecturers at the Department of Electronic and Telecommunication Engineering, University of Moratuwa, for the valuable feedback given during the project implementations. Further, we really appreciate the anonymous reviewers for their constructive feedback, which has significantly improved the quality of the manuscript.

References

[1] 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.*, 27(1):1, January 2020.

[2] Leonardo G. Ferreira, Ricardo N. Dos Santos, Glaucius Oliva, and Adriano D. Andricopulo. Molecular docking and structure-based drug design strategies, Jul 2015.

[3] Ankur Barsode and Aneesh A.M. Bypassing traditional molecular dynamics with artificial neural networks. *AIP Conference Proceedings*, 2023.

[4] Jessica Mustali, Ikki Yasuda, Yoshinori Hirano, Kenji Yasuoka, Alfonso Gautieri, and Noriyoshi Arai. Unsupervised deep learning for molecular dynamics simulations: A novel analysis of protein–ligand interactions in sars-cov-2 mpro. *RSC Advances*, 13(48):34249–34261, 2023.

[5] 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*, 2(2):116–123, feb 2020.

[6] 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*, 119(11), mar 2022.

[7] 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*, 38(7):1995–2002, jan 2022.

[8] Ye Yuan, Qushuo Chen, Jun Mao, Guipeng Li, and Xiaoyong Pan. Dg-affinity: predicting antigen–antibody affinity with language models from sequences. *BMC bioinformatics*, 24(1):430, 2023.

[9] Michael Nilges. Homology modeling. *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, page 814–817.

[10] 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*, 596(7873):583–589, Aug 2021.

[11] Thomas Harkey, Vivek Govind Kumar, Jeevapani Hettige, Seyed Hamid Tabari, Kalyan Immadisetty, and Mahmoud Moradi. The role of a crystallographically unresolved cytoplasmic loop in stabilizing the bacterial membrane insertase yidc2. *Scientific reports*, 9(1):14451, 2019.

[12] Wynne J Browne, ACT North, David C Phillips, Keith Brew, Thomas C Vanaman, and Robert L Hill. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *Journal of molecular biology*, 42(1):65–86, 1969.

[13] Richard Gowers, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, David Dotson, Sébastien Buchoux, Ian Kenney, and et al. Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations. *Proceedings of the Python in Science Conference*, 2016.

[14] Peter J. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and et al. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[15] Andrej Sali. Comparative protein modeling by satisfaction of spatial restraints. *Molecular Medicine Today*, 1(6):270–277, 1995.

[16] Per Larsson, Björn Wallner, Erik Lindahl, and Arne Elofsson. Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Science*, 17(6):990–1002, 2008.

[17] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, Jun 2022.

[18] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[19] James M Rini, Ursula Schulze-Gahmen, and Ian A Wilson. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science*, 255(5047):959–965, 1992.

[20] Marco Giulini, Constantin Schneider, Daniel Cutting, Nikita Desai, Charlotte M Deane, and Alexandre MJJ Bonvin. Towards the accurate modelling of antibody- antigen complexes from sequence using machine learning and information-driven docking. *Bioinformatics*, 40(10):btae583, 2024.

[21] Katherine Maia McCoy, Margaret E Ackerman, and Gevorg Grigoryan. A comparison of antibody–antigen complex sequence-to-structure prediction methods and their systematic biases. *Protein Science*, 33(9):e5127, 2024.

[22] Nathaniel L Miller, Thomas Clark, Rahul Raman, and Ram Sasisekharan. Learned features of antibody-antigen binding affinity. *Frontiers in Molecular Biosciences*, 10:1112738, 2023.

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

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

[25] David J. Lipman and William R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, Mar 1985.

[26] Sarah Sirin, James R Apgar, Eric M Bennett, and Amy E Keating. Ab-bind: antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[27] 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*, 38(16):4051–4052, 2022.

[28] Hyejin Yoon, Jennifer Macke, Anthony P West Jr, Brian Foley, Pamela J Bjorkman, Bette Korber, and Karina Yusim. Catnap: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research*, 43(W1):W213–W219, 2015.

[29] 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*, 9(1):653, 2022.

[30] Nicholas Rego and David Koes. 3dmol. js: molecular visualization with webgl. *Bioinformatics*, 31(8):1322–1324, 2015.

[31] Regina Z Cer, Uma Mudunuri, R Stephens, and Frank J Lebeda. Ic 50-to-k i: a web-based tool for converting ic 50 to k i values for inhibitors of enzyme activity and ligand binding. *Nucleic acids research*, 37(suppl_2):W441–W445, 2009.

[32] David C Swinney. Molecular mechanism of action (mmoa) in drug discovery. In *Annual Reports in Medicinal Chemistry*, volume 46, pages 301–317. Elsevier, 2011.

[33] Vineet D Menachery, Boyd L Yount Jr, Amy C Sims, Kari Debbink, Sudhakar S Agnihothram, Lisa E Gralinski, Rachel L Graham, Trevor Scobey, Jessica A Plante, Scott R Royal, et al. Sars-like wiv1-cov poised for human emergence. *Proceedings of the National Academy of Sciences*, 113(11):3048–3053, 2016.

[34] Vineet D Menachery, Boyd L Yount, Kari Debbink, Sudhakar Agnihothram, Lisa E Gralinski, Jessica A Plante, Rachel L Graham, Trevor Scobey, Xing-Yi Ge, Eric F Donaldson, et al. A sars-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nature medicine*, 21(12):1508–1513, 2015.

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[36] Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and

Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, pages 2022–12, 2022.

[37] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

[38] 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*, 12(1):8434, May 2022.

[39] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[40] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[41] 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.*, 44(10):7112–7127, October 2022.

[42] Rishikesh Magar, Prakarsh Yadav, and Amir Barati Farimani. Potential neutralizing antibodies discovered for novel corona virus using machine learning. *Scientific reports*, 11(1):5261, 2021.

[43] 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*, 13(1):20, Nov 2020.

[44] Kanchan Jha, Sriparna Saha, and Hiteshi Singh. Prediction of protein–protein interaction using graph neural networks. *Scientific Reports*, 12(1):8360, May 2022.

[45] Orhun Vural and Leon Jololian. Machine learning approaches for predicting protein-ligand binding sites from sequence data. *Frontiers in Bioinformatics*, 5:1520382, 2025.

[46] Yong Xiao Yang, Pan Wang, and Bao Ting Zhu. Binding affinity prediction for antibody–protein antigen complexes: A machine learning analysis based on interface and surface areas. *Journal of Molecular Graphics and Modelling*, 118:108364, 2023.

[47] Mark L Chiu, Dennis R Goulet, Alexey Teplyakov, and Gary L Gilliland. Antibody structure and function: the basis for engineering therapeutics. *Antibodies*, 8(4):55, 2019.

[48] Lediya Cheru, Yimin Wu, Ababacar Diouf, Samuel E Moretz, Olga V Muratova, Guanhong Song, Michael P Fay, Louis H Miller, Carole A Long, and Kazutoyo Miura. The ic50 of anti-pfs25 antibody in membrane-feeding assay varies among species. *Vaccine*, 28(27):4423–4429, 2010.

[49] 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*, 42(D1):D1140–D1146, 2014.

[50] Iain H Moal and Juan Fernández-Recio. Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012.

[51] 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*, 35(3):462–469, 2019.

[52] Tuomo Kalliokoski, Christian Kramer, Anna Vulpetti, and Peter Gedeck. Comparability of mixed ic50 data–a statistical analysis. *PloS one*, 8(4):e61007, 2013.

[53] Angela Andreella, Riccardo De Santis, Anna Vesely, and Livio Finos. Procrustes-based distances for exploring between-matrices similarity. *Statistical Methods & Applications*, 32(3):867–882, 2023.

[54] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[55] Marina Caskey, Till Schoofs, Henning Gruell, Allison Settler, Theodora Karagounis, Edward F Kreider, Ben Murrell, Nico Pfeifer, Lilian Nogueira, Thiago Y Oliveira, et al. Antibody 10-1074 suppresses viremia in hiv-1-infected individuals. *Nature medicine*, 23(2):185–191, 2017.

[56] Tongqing Zhou, Jiang Zhu, Xueling Wu, Stephanie Moquin, Baoshan Zhang, Priyamvada Acharya, Ivelin S Georgiev, Han R Altae-Tran, Gwo-Yu Chuang, M Gordon Joyce, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for hiv-1 neutralization by vrc01-class antibodies. *Immunity*, 39(2):245–258, 2013.

[57] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Appendix A. Formulating antibody-antigen binding affinity prediction as a regression problem

Given that the community is interested in both binding vs non-binding classification (i.e., binary interaction determination) [45] as well as precise binding affinity value prediction (i.e., quantifying the interaction strength) [8, 46], we intend to address the latter task in this work since a precise binding affinity prediction is beneficial, over a macro-level classification (which can be more effective for initial screening), for the downstream analyses in therapeutic antibody engineering such as enabling fine-tuning for efficacy [47], which is one of the core objectives in this work. In addition, the datasets mentioned in Tables 1 and 2 comprise numerical values to represent binding affinity in a continuous spectrum. Hence, to facilitate a macro-level binary classification, the numerical binding affinities must be thresholded. However, there is no universal threshold for IC50 value that can be used across all types of antibodies and antigens [48]. Therefore, such thresholding will lead to a sub-optimal problem formulation and downstream performance. Further, the baseline and recent works in the literature [8, 7, 38] treat the binding

affinity and closely affiliated metrics such as Kd and $IC50$ in a continuum spectrum and thereby, utilize regression metrics such as distance errors and correlation values to evaluate their performance. Given the above reasoning, we formulate the problem at hand in a regression fashion.

However, it is trivial that the proposed method can also be extended to obtain a binary output (i.e., as a *proxy* for screening "strong" vs "weak" binders) as well, by thresholding the predicted value at a clinically motivated cut-off η: if the predicted IC50 value from our method for the user input antibody-antigen pair is below η, then the users can safely assume that the pair has a strong binding, and vice versa.

Appendix B. Dataset Descriptions and Selection

Below we briefly summarize the details and contributions of each dataset.

- Ab-Bind [26]: This dataset presents 1,101 mutants from 32 antibody complexes with experimentally determined binding free energy changes. More specifically, AB-Bind focuses on 403 alanine single-point mutations and 242 non-alanine mutations and includes antibody-antigen, antibody-Fc receptor, and nanobody complexes. Optionally, the dataset also provides limited structural data from X-ray crystallography (1.50 to 3.79 Å) and homology models for computational structural benchmarking.

- Ab-CoV [27]: This dataset curates 1,780 coronavirus-neutralizing antibodies with more than 3,200 data points on IC50, EC50, and Kd values. Further, it includes epitope and paratope annotations, predicted mutation impacts on stability and affinity, and Collier de Perles plots for optional structural analysis.

- CATNAP [28]: CATNAP contains HIV-focused resource integrating neutralization data for 172 antibodies against 722 HIV-1 viruses (529 sequenced) from 49 studies. The datset links antibody potency, with IC50 and IC80, with viral Envelope protein sequences and epidemiological data and features Fisher's exact test for detecting antibody resistance and sensitivity signatures in viral sequences.

- SAbDab [49]: This is a structural antibody database containing annotated antibody-antigen complexes where annotations include exper-

imental details such as resolution and method, affinity data, and sequence features such as CDR loops, and species origin. At the time of our download, there were 1327 complex structures along with the corresponding $\Delta\Delta G$ or affinity value.

- SKEMPI [50, 51]: The dataset covers general protein-protein interactions and their binding relationships including energy, kinetics, and thermodynamics. In summary, this dataset of binding free energy changes upon mutation contains 7085 binding data for structurally resolved protein-protein interactions. Further, the dataset reports kinetics changes for 1844 mutations and entropy changes for 440 mutations.

- AlphaSeq [29]: The dataset presents quantitative binding scores of antibodies against a SARS-CoV-2 target peptide collected via an AlphaSeq assay. Raw dataset contains over a million datapoints whereas around 100K designs are attributed to the dataset as per the paper [29]. Antibody affinity values cover a wide range from $37pM$ to $22mM$.

As described in section 2.1, we consider the median value of the provided IC50 values for repeated entries of antibody-antigen pairs to mitigate the impact of outliers. As shown in prior works such as [52], this approach is effective in accounting for extreme outliers (Fig. B.11), which can otherwise disproportionately affect the mean (i.e., skewing the central trend and thereby leading to an inaccurate representation), specifically given the nature of IC50 values: potentially spanning several orders of magnitude. Further, this approach is particularly reasonable in the task of binding affinity prediction since extreme outliers in experimental IC50 measurements can predominantly result from experimental variability (or rare, highly/lower potent binding cases), but here a more representative value for typical antibody-antigen interactions is preferred.

Appendix C. Ablations on Sequence Encoding Schemes

As discussed in the section 2, it is required to convert the letter sequences of protein amino acids into a numerical format before feeding into the sequence-based models. Accordingly, two traditional encoding schemes, namely, one-hot encoding and VHSE8 encoding, and one deep learning-based protein language embedding were utilized for this task. Our inspiration behind these selections is as follows:

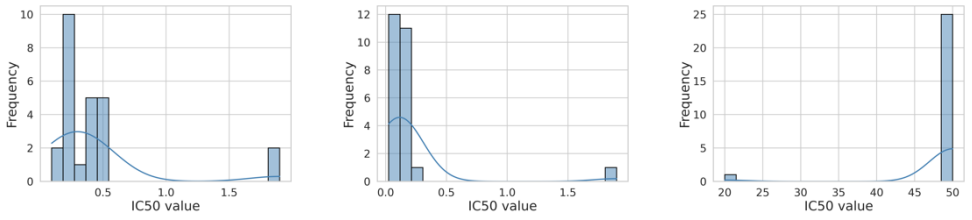


Figure B.11: Exemplar affinity distributions for repeated antibody-antigen pairs. In these cases, we consider the median value of the provided IC50 values to mitigate the impact of extreme outliers.

- In a high level, since the sequence encoding method is model-agnostic (with respect to our learning approach) and is not a contribution of this work, our objective of these encoding ablations was to quantitatively compare the performance sensitivity of our approach in response to the different representative encoding schemes in the literature.

- To this end, we selected the "one-hot" scheme as a representative baseline for conventional non-biased methods since it presents a simple method where each amino-acid is represented independently of others. Further, it is data-driven, without embedding any prior assumptions or knowledge about amino-acid sequences.

- The VHSE8 encoding scheme is a representative of the methods which are equipped with expert-based domain-specific knowledge. In this regard, VHSE8 encapsulates key biochemical properties, such as hydrophobicity and steric effects, which are crucial and highly relevant for protein interactions.

- The ProtT5-XL-BFD protein language embedding [41] is the representative for model learning-driven sequence encoding methods at which the encoding is learned through a representation learning process via a significantly larger amino-acid sequence dataset(s). Therefore, it is reasonable to identify this approach as a hybrid of the above two classes where domain-specific knowledge is learned through a data-driven pipeline.

As indicated in Table C.7, the VHSE8 encoding scheme was utilized in the model implementations (including sequence-based models and the combined models) due to its superior performance over the other two encoding schemes. We believe that the chemical and physical properties of the amino acids embedded within the VHSE8 encoding scheme are the reason for the better performance compared to one-hot encoding, which is a sparse encoding scheme, and BLOSUM encoding, which is a domain-specific method specifically targeting evolutionary relationships between amino acids based on substitution frequencies.

Encoding scheme	MAE
One-hot	0.9757
VHSE8	**0.9682**
BLOSUM	0.9774

Table C.7: Results comparison between different encoding schemes. Here, the P2PXML-Seq dataset is utilized and the model is a parallel multi-layer perception model.

As discussed previously under the Results section, the protein language embeddings from ProtT5-XL-BFD [41] were also utilized to numerically encode the protein amino acid sequences. However, it was shown in Table 4 that the performance of such an embedding representation is inferior to the VHSE8 encoding scheme in this context.

Appendix D. Ablations on Model and Training Hyperparameter Selections

In all our deep learning experiments, each utilized dataset is partitioned into training, validation, and test sets using a ratio of $17:1:2$, respectively. Further, 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 Table D.8, it is evident that the best MAE is given by a learning rate of 0.0001, along with the ADAM optimizer and 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 the best MSE. However, under the hyperparameters that produced 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 to use the set of hyperparameters that produced the best MAE value.

LR	Optimizer	D/O	EV	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
	ADA-DELTA	0.05	✗	0.089726	0.036105
		0.1	✗	0.091582	0.036193
		0.2	✗	0.094721	0.037684

Table D.8: Optimal hyperparameter determination through ablations by exposing the proposed sequence model to the AlphaSeq dataset. *LR, D/O, EV refer to Learning Rate, Dropout and Embedding Vector, respectively.

Appendix E. Model Optimization

We intend to incorporate both evolutionary and atomistic details into our final antibody-antigen binding affinity prediction model through the employed sequence-based and structure-based models respectively. To this end, we hypothesize that both aspects are equally critical for the task at hand and therefore, we set the initial weights for sequence-based and structure-based learning guidance and in the final model to be approximately 0.5. However, according to our reported results (as in Tables 3, 5 and 6), our sequence model slightly outperforms the structure model in general, even though both models are better than the existing corresponding works in the literature, respectively. However, to avoid a potential learning imbalance, we (1) heuristically determine the hyperparameters in the objective function (λ, μ and ν) (by adjusting within the local search space through empirical evidence as observed in ablation studies as in Fig. E.12a) that resulted in a respective combination of $\{0.425, 0.525, 0.05\}$ validating our previous observation on the slight edge in importance of sequences for the task. Note that

	Feature	Experiments						
Node	x-coordinate	✓	✓	✓	✓	✓	✓	✓
	y-coordinate	✓	✓	✓	✓	✓	✓	✓
	z-coordinate	✓	✓	✓	✓	✓	✓	✓
	Atomic Number	✓	✓	✓	✓	✓	✓	✓
	Chirality	✓	✓	✓	✗	✗	✗	✗
	Implicit Valence	✓	✓	✗	✓	✗	✗	✗
	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 D.9: Evaluation of node and edge features using the protein structures generated (using AlphaFold-V2 multimer model) for the protein sequences in Ab-Cov and AB-Bind datasets

the auxiliary third term ($MSE(\hat{y_{str}}, \hat{y_{seq}})$) is to minimize the potential disparity between two supervision paths and thus, it is a complementary weight of which the value is set to this auxiliary loss term after the assignment of first two weights. Subsequently, we (2) carefully (and heuristically) select the weights (δ_1 and δ_2) of final binding affinity prediction (Eq. 7) to be $\{0.45, 0.55\}$ which also aligns with above observations (Fig. E.12b). Therefore, we believe that these measures are reasonable and sufficient to avoid (1) a potential learning imbalance while (2) respecting the superiority of the sequence model and (3) adequately acquiring the impact of both models (by balancing around 0.5).

To address the trade-off between the sufficiency of the model complexity and overfitting risk, we implement the following steps accordingly:

- We equip the model with the potential to capture input data complexity through:
  - Sequence model is comprised of (1) a set of multi-head attention layers to learn a representative latent of the internal data complexity of antibody and antigen separately whereas (2) hierarchical cross-attention layers mimic the multi-scale information sharing between antibody and antigen (Fig. 2)
  - Structure model consists of (1) a set of graph convolution lay-

ers to aggregate the local atomistic complexity through the $N-$ hop neighbourhood whereas (2) graph attention is employed to determine the node relevance in a learnable fashion (Fig. 3).

  - Further in combined model (Fig. 4), the domain and modality-specific knowledge learned from sequence and structure-based model is shared through a set of cross-attention layers.

- We minimize the risk of overfitting through:

  - By following the standard machine learning practices, we add dropout layers and regularization terms in both sequence and structure-based model paths to combat the overfitting risk within the model architecture.

  - When splitting the samples for training and testing, we ensure that there is a limited possibility for data leakage through (1) eliminating sequences which have higher sequence similarity than a set threshold and (2) encouraging the samples relating to same antigen variant into same data split, apart from standard data preparation procedures such as the removing the duplicates.

  - Optionally, given that (1) amino-acid sequences are considerably long and higher dimensional (as an example, in our P2PXML-Seq dataset, the sequences can be as large as having $3.1K$ amino-acid letters, which are reduced to lower latent representations as in Fig. 8a in the paper), whereas (2) protein structures are significantly structurally-complex due to the larger number and variety of atoms (which are reduced to lower dimensional representations as in Fig. 8b in the paper), it is evident that there is a *lesser risk* of overfitting to protein sequence and structure hyper-space.

Appendix F. Ablations on Graph Node and Edge Features

One of the most crucial steps in geometric deep learning, specifically graph-based deep learning, is the initial graph representation, as it would directly impact the subsequent node and edge level aggregations and information flow.
Therefore, an extensive ablation study was performed to find an optimal set of node and edge features for the graph representation of proteins and thereby

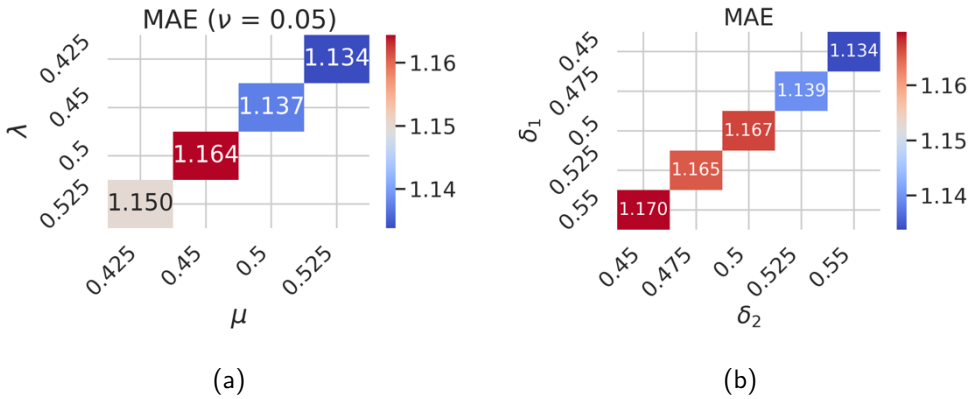


Figure E.12: (a) Ablation study on the weight hyperparameters of the learning objective (Eq. 8), and (b) Ablation study on the weight hyperparameter of the final prediction (Eq. 7)

strengthen the chemistry-based intuition of the final model. The results under two datasets, namely, Ab-CoV and Ab-Bind were used to compare different combinations. As per the results presented in Table D.9, 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 were selected to be used in the remaining experiments associated with the structure-based model and the combined model.

Appendix G. Synergies between Sequence and Structure Learning

Given that our intuition behind the embedding combination in the combined model, via the cross-attention and concatenation, is to share the information learned from both sequence and structure models in a learnable fashion, we further extend our analysis on the patterns of learning and integration between two modalities, potentially to discover any learned cross-sample synergies between the learning of two modalities. Here, our analysis is two-fold:

- As the trivial approach, we first compare the predicted binding affinities from both sequence and structure models before they are utilized to

generate the final binding affinity. As shown in Fig. G.13a, it is evident that the learning from both sequence and structure models is highly synergetic and aligned, where both the Pearson correlation coefficient and Spearman's correlation coefficient are 0.998, and the coefficient of determination is 0.995, denoting a higher level of (global alias high-level) agreement between the sequence and structure encoders.

- Even though comparing the internal weight spaces of both sequence and structure models is theoretically more rigorous and sound, such a layer-wise comparison is impractical and impossible in our case due to different model architectures. Therefore, as a workaround, we compare the similarity between the most relevant latent space, i.e., the regression layer weights immediately before the corresponding predictions, from both sequence and structure models as a *proxy* towards revealing their learned synergies or conflicts. As per our analysis, as shown in Fig. G.13b, the overall matrix similarity, determined through a range of metrics including cosine mean cosine similarity, Frobenius ratio, between two regression layers is weak. This denotes that despite the mutual information sharing, each model learns different and unique representations that are primarily specific to the input types and interaction spaces, as expected.

Further, as shown in Table 4, our combined model with cross-attention outperforms simple concatenation and late fusion methods, while consistently outperforming sequence-only and structure-only models (Table 3), demonstrating that the modalities are complementary rather than redundant.

Appendix H. Weight Space-based Interpretability Analysis

Appendix H.1. High-level Empirical Weight Space Analysis

To empirically investigate the internal learning patterns of the model, we first objectively analyse the model weight spaces through visualizing the attention layers of numerous dimensionality and nature. As depicted in the first row of Fig. H.14, the evidences hint at a non-redundant and homogeneous learning pattern where most of the linear biases are not null, but balanced around zero. In addition, as shown in the second row of the same figure, the learning weights are diverse, in both initial and deeper (from Fig. H.14 (f) to (j) respectively) layers, suggesting both the corresponding importance of

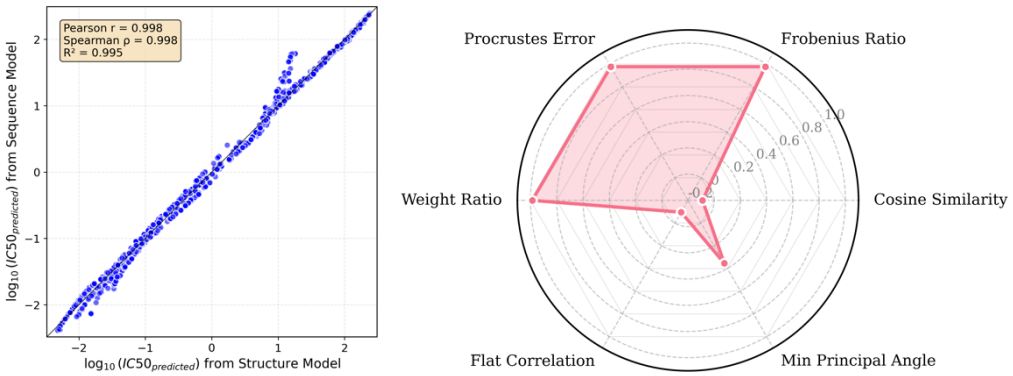


Figure G.13: (a) $\log_{10}(IC50)$ of the predicted values from sequence model vs $\log_{10}(IC50)$ of the predicted values from the structure model for the test set of P2PXML-PDB dataset using our best performing models, and (b) Weight space similarity metrics analysed on the last regression layers of both sequence and structure models: here, we utilize a range of diverse metrics such as mean cosine similarity, Frobenius norm ratio, minimum principal angle, Procrustes error, mean weight ratio, and mean Pearson correlation. More details on these metrics can be found at [53].

input features and an efficient learning process. Further, the first few layers (such as Fig. H.14(f)) hint that some features are more sensitive (i.e., a higher variation in low and high weights) towards the output, whereas some features seem to be complementary (i.e., visibly similar patterns or stripes). The third row shows that the sequence transformers mostly learned a Gaussian or mixture of Gaussian distributions across the feature space, hinting at a similar normal distribution in the input data space, at least with respect to the task at hand.

Appendix H.2. Per-Sample Case Study

To further widen our analysis towards realizing the aspect of biological interpretability of the proposed method, we conduct an input feature importance analysis for a selected antibody-antigen pair targeting the identification of highly influential atoms and residues in the input graphs and sequences, respectively. To this end, as our method combines different model architectures with multiple non-linear learnable layers, it is non-trivial and ill-posed to functionally map between the input features and the weight space. Therefore, to circumvent this difficulty, we utilize a wide array of well-established techniques, that are tailored to each model architecture (and by extension,

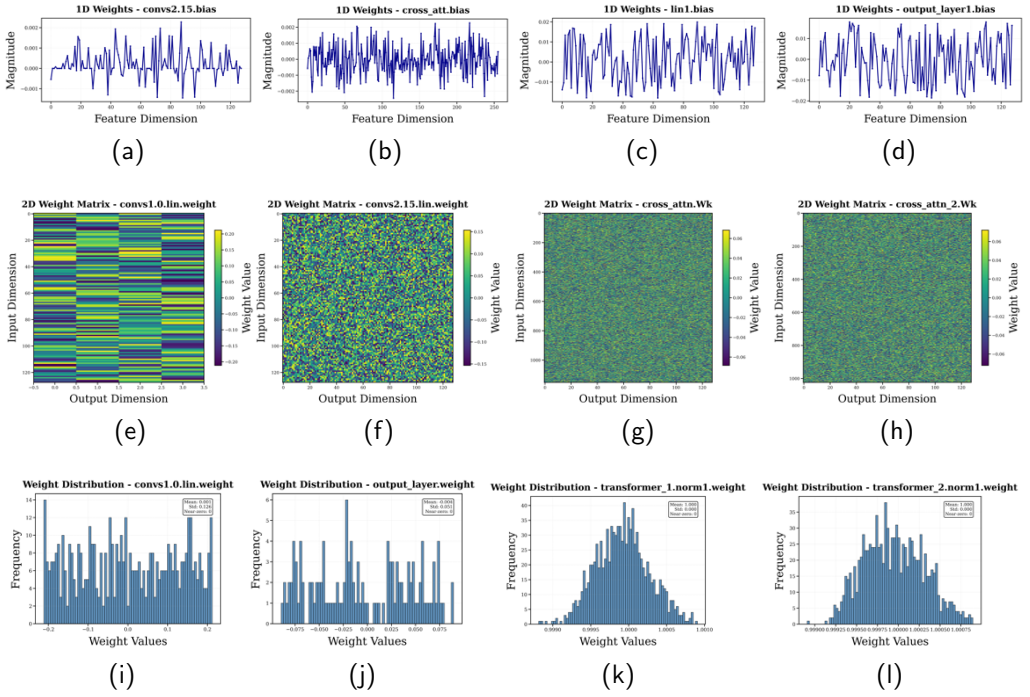


Figure H.14: Weight space analysis: The first row depicts a representative sequence of 1D weight bias propagation (i.e., weight magnitude vs the feature dimension) of GCN layers (a), cross attention (b), dense (c) and output layers (d). The second layer shows a representative sequence of 2D weights propagation (i.e., input feature dimension vs output feature dimension of the layer in consideration) of graph attention (e and f) and cross-attention layers (g and h). The colours refer to the corresponding weight value, as in the colour bar on the left of each plot. The third layer is a representative sequence of FASTA sequence transformer 1D weights (i.e., weight histogram of each layer) of linear (i), output (j), antibody (k) and antigen MHA (l), and the legend in each plot presents the mean weight, the standard deviation of the weight distribution, and the count of near-zero weights in the weight distribution. The complete set of weight analysis figures can be found at the project GitHub repository.

to each protein representation type), to derive insights on the corresponding feature importance. In summary, we implement GNNExplainer [54] to evaluate the structural importance whereas sequence-aware attention, gradient-based saliency, and alanine scanning are utilized to assess the importance of residues in the amino-acid sequences. We describe these techniques in detail below.

Appendix H.2.1. Interpretability Methods

- *GNNExplainer* [54]: This is a model-agnostic (i.e., applicable to any *GNN* model) approach for providing interpretable explanations for predictions on any graph-based machine learning task. When operating, for a given input, this technique identifies a compact subgraph structure and a small subset of node features that have a dominant role in GNN's prediction by maximizing the mutual information between the prediction and distribution of possible subgraph structures. In our implementation, we run the GNNExplainer for 100 epochs, under the settings of masked node features and masked edges. In addition, the specific values for the configuration parameters: *mode*, *task-level*, and *return-type*, are set to *regression*, *graph-level prediction*, and *raw values* respectively, considering the nature of our experiment.

- *Sequence-aware attention analysis*: Noting that we preprocess the amino-acid sequences (i.e.,first encode and subsequently flatten the encoded sequences) as an architectural requirement, we utilize a gradient-based approach to attribute the attention importance scores to the residue level. This gradient-based approach is more reliable than directly obtaining attention scores learned from flattened sequences, as it is able to capture the cumulative effect across the flattened encodings of the input amino-acid sequences. To this end, here, we analyse how each attention parameter is weighted (i.e., thereby their importance) at the inference in predicting binding affinity for an antibody-antigen pair, as shown in Algorithm 1. As below, we denote our (trained) combined binding affinity prediction model as M, which is with the parameter space: $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, and $L = \sum \hat{y}$ where L and $\hat{y}$ are the loss function and the final prediction respectively. Consequently, we assign an importance score for each attention parameter θ_i as:

$$I(\theta_i) = \left\| \frac{\partial L}{\partial \theta_i} \right\|_2 \text{ where } \frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial \theta_i} \text{ through backpropagation} \tag{H.1}$$

  We believe that this gradient norm serves as a measure of parameter sensitivity, with larger values indicating that small changes in the parameter significantly affect the binding affinity prediction for the input antibody-antigen pair.

- *Gradient-based saliency analysis*: Here, our objective is to explore which specific residues in the antibody and antigen sequences would most affect the predicted binding affinity if they were slightly changed. To this end, our intuition can be presented threefold:

  - Sensitivity: residues with high gradient magnitudes indicate that small changes at those positions would significantly affect the predicted binding affinity. These are likely critical residues for antibody-antigen interaction.
  - Feature importance: the saliency map identifies which sequence features (i.e., residues) the model relies on most heavily when making its prediction, and therefore, high saliency values hint at important functional or structural residues.
  - More interpretable attribution: Unlike attention weights that show internal model computations, saliency maps directly show how input features influence the output, providing a more direct interpretation of feature importance.

  Based on this reasoning, here (Algorithm 2) we define the saliency map for a sequence tensor S as the absolute value of the gradient of the model output with respect to the input sequence: $G(S) = \left|\frac{\partial y}{\partial S}\right|$. Therefore, for each residue position i in the sequence, the saliency score is: $G(S)_i = \left|\frac{\partial y}{\partial S_i}\right|$.

- *In-silico alanine scanning*: To further strengthen our analysis on exploring more biologically important residues for antibody-antigen binding, we utilize the direct computational equivalent of experimental alanine scanning mutagenesis: *in-silico* alanine scanning. The core principle behind this approach is to systematically measure the functional importance of each residue by simulating the effect of mutating it to alanine, the simplest amino acid. We specifically select Alanine since it (1) contains minimal side chain (i.e., potentially removing functional groups while maintaining the backbone structure), (2) leads to conservative mutations (i.e., minimal structural disruption), and (3) follows a standard benchmark with well-established experimental backing. Therefore, in summary, this approach is presented as a more biological grounded, interpretable (i.e., direct reflect on functional impact),

model-agnostic and quantitative (i.e., present exact changes in predicted affinity) approach to evaluate the residue importance. As structured in Algorithm 3, the alanine scanning effect for residue i in sequence S is defined as: $\Delta(i) = M(S) - M(S_{\text{mutation}(i \to A)})$ where $M(S)$ is the predicted binding affinity for wild-type sequence, $M(S_{\text{mutation}(i \to A)})$ is the predicted binding affinity after mutating residue i to alanine, and $\Delta(i)$ is the change in binding affinity due to alanine mutation. Therefore, for a sequence of length L, the change vector would be: $\Delta = [\Delta(1), \Delta(2), \ldots, \Delta(L)]$.

Algorithm 1 Gradient-based Attention Importance Analysis

Require: Combined model M, which includes self-attention and cross-attention modules; Antibody and antigen data: $Ab_{\{str,seq\}}$, $Ag_{\{str,seq\}}$ respectively

Ensure: Attention importances A: dictionary mapping each parameter to its cumulative gradient norm

Set model to evaluation mode: M.eval()
Forward pass: $\hat{y} \leftarrow M(Ab_{\{str,seq\}}, Ag_{\{str,seq\}})$ ▷ Get final prediction
Compute gradient: $\nabla \hat{y}_{\text{sum}} \leftarrow \nabla(\sum \hat{y})$ ▷ Backward pass
for each module $m \in M$ **do**
    if m is SELF-ATTENTION or CROSS-ATTENTION **then**
        for each parameter θ in m **do**
            if $\nabla\theta$ exists **then**
                Compute gradient norm: $g \leftarrow \|\nabla\theta\|_2$
                Store: $A[m : \theta] \leftarrow g$
            end if
        end for
    end if
end for
return A

Appendix H.2.2. Antibody-antigen pair

To conduct our per-sample case study, we select the following antibody-antigen pair: antibody *10-1074*, which is a well-studied potent, broadly neutralizing monoclonal antibody that typically targets a specific site (i.e., the V3 glycan supersite) on the HIV-1 envelope protein to prevent or treat HIV-1

Algorithm 2 Gradient-based Saliency Map Analysis

Require: Combined model M; Antibody and antigen data: $Ab_{\{str,seq\}}$, $Ag_{\{str,seq\}}$ with sequence tensors Ab_{seq}, Ag_{seq} respectively

Ensure: Saliency maps G_{Ab}, G_{Ag}: gradient magnitudes for each residue

Set model to evaluation mode: M.eval()

Enable gradients: Ab_{seq}.requires_grad $\leftarrow$ True, Ag_{seq}.requires_grad $\leftarrow$ True

Forward pass: $\hat{y} \leftarrow M(Ab_{\{str,seq\}}, Ag_{\{str,seq\}})$ ▷ Get final prediction

Compute gradient: $\nabla\hat{y} \leftarrow \frac{\partial \hat{y}}{\partial \{Ab,Ag\}_{seq}}$ ▷ Backward pass

Compute saliency maps:

$G_{Ab} \leftarrow \left|\frac{\partial \hat{y}}{\partial Ab_{seq}}\right|$ ▷ Absolute gradients for antibody

$G_{Ag} \leftarrow \left|\frac{\partial \hat{y}}{\partial Ag_{seq}}\right|$ ▷ Absolute gradients for antigen

Disable gradients: Ab_{seq}.requires_grad $\leftarrow$ False, Ag_{seq}.requires_grad $\leftarrow$ False

return G_{Ab}, G_{Ag}

Algorithm 3 In-Silico Alanine Scanning

Require: Combined model M; Antibody and antigen data: $Ab_{\{str,seq\}}$, $Ag_{\{str,seq\}}$ with sequence tensors Ab_{seq}, Ag_{seq} respectively; Amino acid set $\mathcal{A} = \{A, C, D, \ldots, Y\}$ (A=alanine at index 0)

Ensure: Effect arrays Δ_{Ab}, Δ_{Ag}: binding affinity changes for each mutation

Get baseline prediction: $\hat{y}_{\text{orig}} \leftarrow M(Ab_{\{str,seq\}}, Ag_{\{str,seq\}})$

for each residue position $i = 1$ to L_{Ab} in antibody sequence Ab_{seq} **do**

    Create mutant: $Ab^{\text{mut}}_{seq} \leftarrow \text{clone}(Ab_{seq})$

    Set residue i to alanine: $Ab^{\text{mut}}_{seq}[i] \leftarrow \text{encode}(A)$

    Compute mutant prediction: $\hat{y}_{\text{mut}} \leftarrow M(Ag^{\text{mut}}_{\{str,seq\}}, Ag_{\{str,seq\}})$

    Record effect: $\Delta_{Ab}[i] \leftarrow \hat{y}_{\text{orig}} - \hat{y}_{\text{mut}}$

end for

for each residue position $j = 1$ to L_{Ag} in antigen sequence Ag_{seq} **do**

    Create mutant: $Ag^{\text{mut}}_{seq} \leftarrow \text{clone}(Ag_{seq})$

    Set residue j to alanine: $Ag^{\text{mut}}_{seq}[j] \leftarrow \text{encode}(A)$

    Compute mutant prediction: $\hat{y}_{\text{mut}} \leftarrow M(Ab_{\{str,seq\}}, Ag^{\text{mut}}_{\{str,seq\}})$

    Record effect: $\Delta_{Ag}[j] \leftarrow \hat{y}_{\text{orig}} - \hat{y}_{\text{mut}}$

end for

return Δ_{Ab}, Δ_{Ag}

infection [55], and the antigen *0013095_2_11* [56], which is a clade C HIV-1 Env isolate whose gp120 loop D and V5 sequences are studied both as a neutralization test virus and as donor sequences for loop-swap chimeras in coreE gp120 constructs—thereby helping to map how specific gp120 loop elements shape VRC01-class antibody binding and neutralization.

Appendix H.2.3. Evaluation Method

To independently evaluate the closeness of the identified highly influential residues against the structural biophysics, we employ the following steps and metrics.

- We first generate the antibody-antigen complex using AlphaFold3 [57] as it is one of the state-of-the-art methods for bio-macromolecule complex structure prediction.
- Subsequently, we estimate the interface residues between the antibody and antigen in the complex by utilizing a fixed distance threshold between any atom pairs from the antibody and antigen (i.e., $5\mathring{A}$), thereby indirectly inferring the paratope through the interface residues in the antibody, and the epitope through the interface residues in the antigen.
- We then numerically evaluate the similarity and overlap between the inferred interface residues (in the complex) and the previously identified residues (in their unbound forms) using the metrics below. Here, our analysis is twofold: (1) the pairwise overlap between the inferred interface residues and unbound residues in a strict sense, using metrics such as precision, recall, F_1 score, and the enrichment $p-$value as shown in Equations H.4, H.5, and H.6 respectively, (2) difference in the solvent-accessible surface area (SASA) between the unbound structures and complex structures considering both per-residue-level (Eq. H.7) and collective residue-level (Eq. H.8 and H.9).
- In our residue overlap analysis, we denote the set of residues deemed highly influential by model weight analysis as $\text{Pred}_{unbound}$, whereas the set of inferred interface residues is denoted by $\text{Pred}_{complex}$. Further, following the typical practise in the field, TP, FP, and FN denote true positives, false positives (Eq. H.2), and false negatives (Eq. H.3), respectively.

$$TP = |\text{Pred}_{complex} \cap \text{Pred}_{unbound}| \; ; \; FP = |\text{Pred}_{unbound} \setminus \text{Pred}_{complex}| \tag{H.2}$$

$$FN = |\text{Pred}_{complex} \setminus \text{Pred}_{unbound}| \tag{H.3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \text{ and Recall} = \frac{TP}{TP + FN} \tag{H.4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{H.5}$$

$$p = P(X \geq k) = \sum_{i=k}^{\min(n,K)} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{H.6}$$

- In our SASA analysis, we denote the SASA of residue r in the unbound structure as $SASA_{unbound}(r)$, whereas $SASA_{complex}(r)$ denotes the SASA of residue r in the complex structure (Eq. H.7). Further, S denotes a set of residues in consideration.

$$\Delta SASA(r) = SASA_{\text{unbound}}(r) - SASA_{\text{complex}}(r) \tag{H.7}$$

$$\overline{\Delta SASA}(S) = \frac{1}{|S|} \sum_{r \in S} \Delta SASA(r) \tag{H.8}$$

$$\overline{\Delta SASA}_S = \overline{\Delta SASA}(S) \tag{H.9}$$

where,

$$S \in \{\text{all interface residues}, \text{predicted residues}, \text{non-predicted residues}\} \tag{H.10}$$

Appendix H.2.4. Results & Discussion

First, we discuss and deduce insights from the results of each interpretability method applied to the selected antibody-antigen pair in detail. Subsequently, we present and discuss the evaluation results based on the metrics we present in Appendix H.2.3.

- *GNNExplainer* [54]: As depicted in Fig. H.15, GNNExplainer highlights 12 (out of 3495 nodes) and 11 (out of 3828 nodes) nodes in antibody and antigen graphs respectively to hold dominant mean features for the task at hand (i.e., beyond the statistical threshold of $(\mu + 3\sigma)$ where μ is the mean importance score and σ is the standard deviation of importance scores), and thereby it is plausible to state that those atoms are deemed to be crucial for binding affinity prediction (and

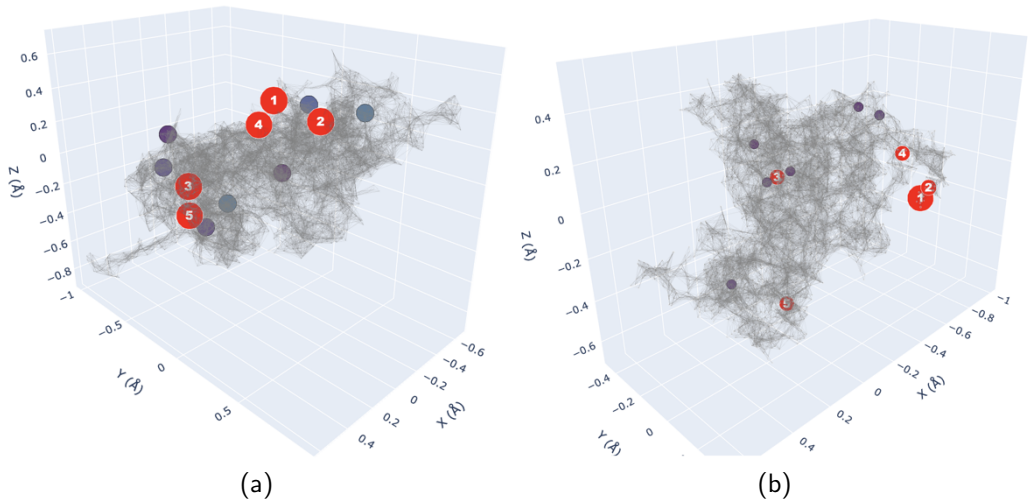


Figure H.15: GNNExplainer [54]-based structure importance analysis for the selected antibody-antigen pair in a post-hoc fashion. Here, the threshold for importance is defined in a strict sense: three standard deviations away from the mean (i.e., $\mu + 3\sigma$). The top 5 important nodes (i.e., ranked as 1 being the most important) are highlighted in red, whereas other important nodes are colored. The background nodes (i.e., with lesser importance) and edges are coloured in gray. (a) Top nodes in antibody graph (rank, atom type, importance): {(1, N, 0.306), (2, C, 0.305), (3, N, 0.304), (4, O, 0.303), (5, C, 0.303)}, (b) Top nodes in antigen graph (rank, atom type, importance): {(1, N, 0.367), (2, C, 0.311), (3, C, 0.310), (4, C, 0.308), (5, O, 0.307)}

potentially lie within the binding interface). For antibody graph, the mean importance is 0.275 and the top node, which corresponds to a N atom, carries an importance score of 0.306. In contrast, for the antigen graph, the mean importance is 0.274 and the top node, which also corresponds to a N atom, carries an importance of 0.367.

- *Sequence-aware attention analysis*: As depicted in Fig. H.16, the importance scores, which are deduced through the cumulative gradient norms of each attention parameter, hint that the sequence pipeline within our combined model focuses heavily on the individual amino-acid sequences of both antigen and antibody (i.e., indicated through the top 7 self-attention or transformer layers with the highest importance), while also adequately imitating the interaction space between the antibody and antigen (i.e., by making 3 cross-attention layers into the top 19 layers out of all 41 related layers). These post-hoc observations,

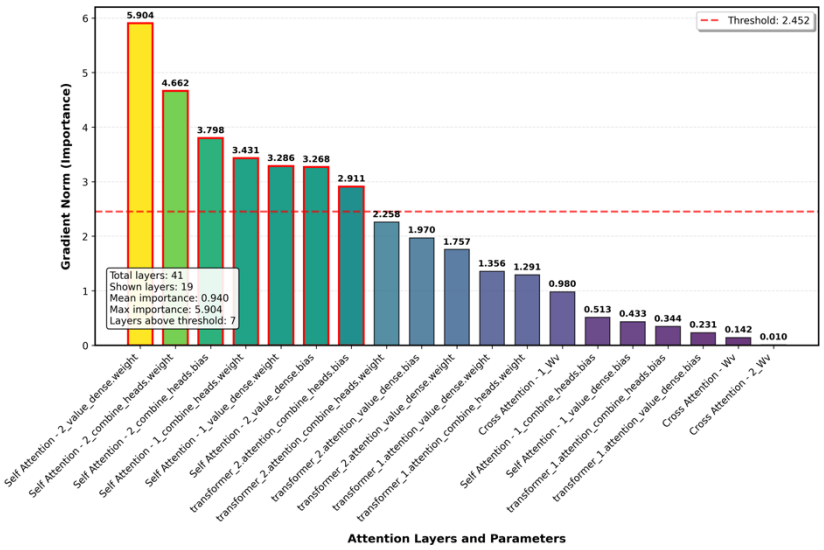


Figure H.16: Gradient-based importance of each attention-parameterized layer for the selected antibody-antigen pair at inference. Here, the threshold for importance is defined as one standard deviation away from the mean (i.e., $\mu+\sigma$). The complete naming convention of the layers can be found at the project GitHub repository.

at least for the selected antibody-antigen pair, align with our initial hypothesis, where we initially intend to extract self-contained evolutionary insights from unbound sequences for binding affinity prediction while also computationally replicating the sequence-based interactions through a non-redundant learning mechanism.

- *Gradient-based saliency analysis*: Our residue importance results deduced through gradient saliency (Fig. H.17) point 20 residues (out of 458 residues) in the antibody sequence, and 16 residues (out of 486 residues) in the antigen sequence, as significant (i.e., beyond the statistical threshold of $(\mu + 2\sigma)$ where μ is the mean importance score and σ is the standard deviation of importance scores) for binding affinity prediction. This observation suggests that the changes to these residues will significantly alter the binding affinity prediction, revealing them as important functional or structural residues. Notably, all the important residues highlighted in the antibody sequence lie within its heavy chain, hinting at a close alignment with the established biophysical knowledge (and potentially revealing the paratope region of

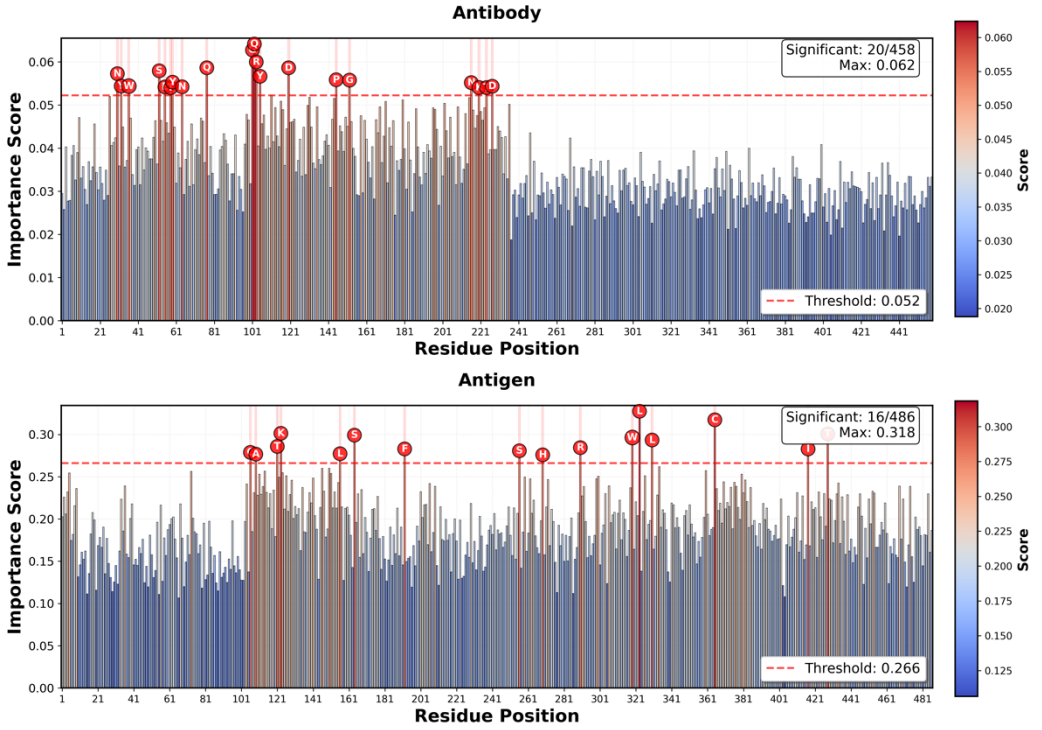


Figure H.17: Residue importance deduced through the cumulative gradient saliency. Here, the thresholds for importance are defined as two standard deviations away from the mean (i.e., $\mu + 2\sigma$) separately for antibody and antigen. Top figure includes complete important residue list of antibody (residue position, residue): {(29,N), (31,Y), (35,W), (51,S), (54,E), (57,T), (58,Y), (63,N), (76,Q), (100,G), (101,Q), (102,R), (104,Y), (119,D), (144,P), (151,G), (215,N), (219,K), (223,T), (226,D)}. Bottom figure includes complete important residue list of antigen (residue position, residue): {(105,C), (108,A), (120,T), (122,K), (155,L), (163,S), (191,F), (255,S), (268,H), (289,R), (318,W), (322,L), (329,L), (364,C), (416,I), (427,T)}

the antibody when binding).

- *In-silico alanine scanning*: This approach is specifically aimed at highlighting the functionally important (i.e., indirectly quantifying each residue's contribution to the total binding affinity) residues while minimally disrupting the structure (i.e., since Alanine preserves structure while removing side-chain functionality). Through our experiment, while using the same statistical threshold of $(\mu + 2\sigma)$, 17 residues are found to be functionally important: either as essential or detrimen-

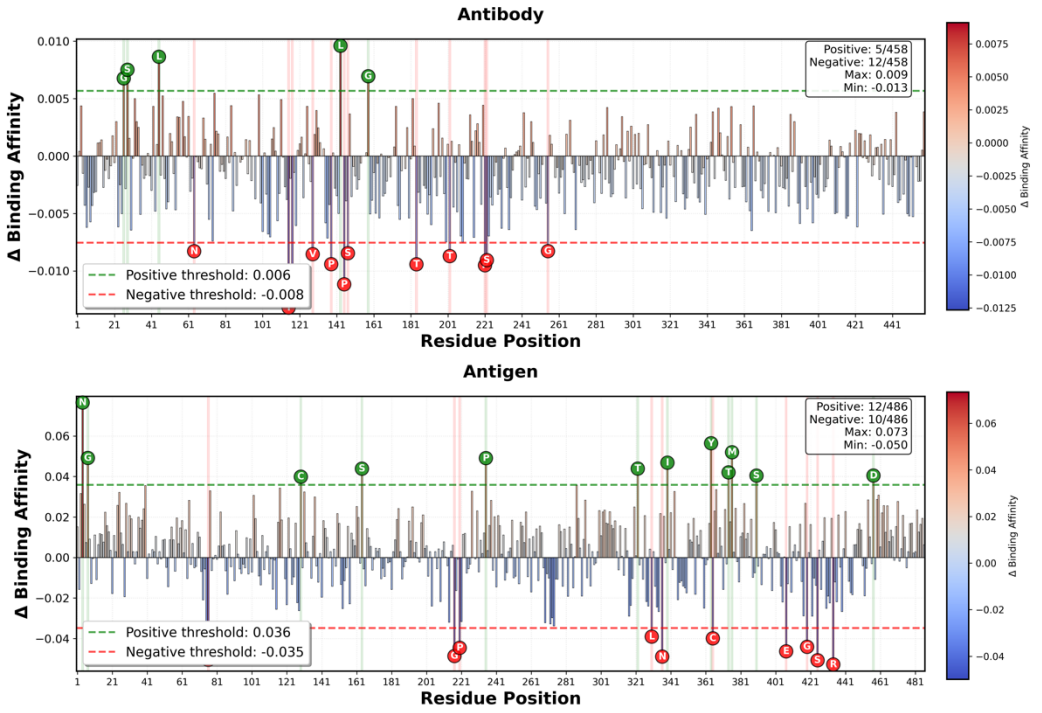


Figure H.18: Residue functional importance deduced through the *in-silico* alanine scanning. Here, the thresholds for importance are defined as two standard deviations away from the mean (i.e., $\mu + 2\sigma$) separately for positive and negative Δ sets in both antibody and antigen sequences. Top figure includes complete important residue list of antibody: {Positive: (25,G), (27,S), (44,L), (142,L), (157,G); Negative: (63,N), (114,Y), (116,Y), (127,V), (137,P), (144,P), (146,S), (183,T), (201,T), (220,P), (221,S), (254,G)}. Bottom figure includes complete important residue list of antigen: {Positive: (3,N), (6,G), (128,C), (163,S), (234,P), (321,T), (338,I), (363,Y), (373,T), (375,M), (389,S), (456,D); Negative: (75,V), (216,G), (219,P), (329,L), (335,N), (364,C), (406,E), (418,G), (424,S), (433,R)}

tal, in the antibody sequence, whereas 22 residues are highlighted to be functionally important in the antigen sequence (Fig. H.18). Here, we denote the residues that, when removed, hurt the binding affinity as *essential residues*, whereas the ones that, when removed, help the binding affinity as *detrimental residues*. For the selected antibody-antigen pair, we observe that 5 residues, highlighted in green, in the antibody sequence result in lower $IC50$ values (i.e., higher binding affinity) than the corresponding wildtype $IC50$ values, thereby marking them as detrimental residues. Similarly, we find 12 such residues in

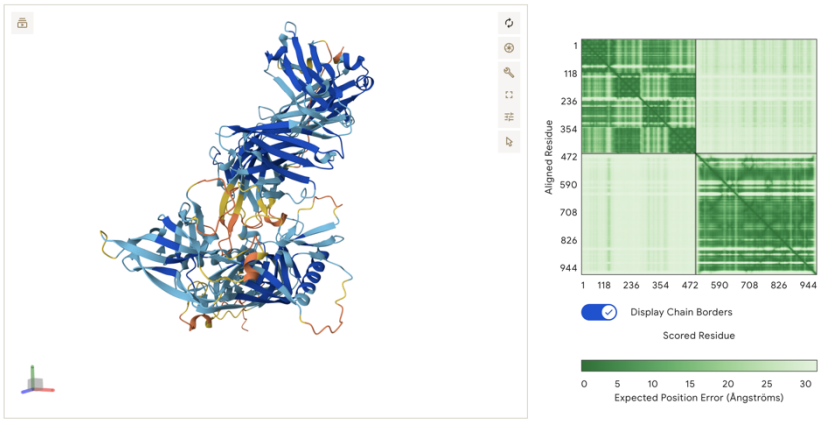


Figure H.19: Predicted antibody-antigen complex structure by AlphaFold3 for the selected antibody-antigen pair. As shown in the position error matrix, even though AlphaFold3 is significantly confident in predicting the individual antibody and antigen structures, it has a higher predicted error for the interface, which is also reflected in lower predicted interface TM-score: 0.19, which in turn, causes to reduce the overall predicted TM-score for the complex: 0.49.

the antigen sequence. In contrast, the residues, highlighted in red, in both the antibody (12 residues in antibody) and antigen (10 residues in antigen) sequences result in higher $IC50$ values (i.e., lower binding affinity) than the corresponding wildtype $IC50$ values, and thereby marking them as essential residues. Regardless of being essential or detrimental, these residues, in general, contribute significantly to the binding energy and likely form binding interfaces. In addition, surprisingly, the residue positions marked by both gradient-based saliency and alanine scanning exhibit a similar global positioning pattern, further reinforcing the corresponding findings.

As presented in Appendix H.2.3, we first generate the antibody-antigen complex structure using AlphaFold3 [57] as shown in Fig. H.19. Here, it is to be noted that even though AlphaFold3 is significantly confident in predicting the individual antibody and antigen structures, it has higher predicted error for the interface, which is also reflected in lower predicted interface TM-score: 0.19, which in turn, causes to reduce the overall predicted TM-score for the complex: 0.49. This is expected as numerous studies show that even the state-of-the-art bio-macromolecule complex structure predictors fail at confidently predicting the antibody-antigen complex structures [20].

As presented in Table H.10, the residue overlap between the inferred interface residues and the highly influential residues (from both methods) is not strong since the average F_1 score is 0.4 and the enrichment $p-$value is higher than the typical statistical threshold (except for residues in antigen deduced through *in-silico* alanine scanning which results in a considerably lower enrichment $p-$value). This observation suggests that there is no sufficient , at least for the selected antibody-antigen pair, to claim that the highly influential residues in unbound forms are enriched for inferred interface residues. Further, for both methods, SASA analysis reveals a similar pattern: (1) When antibody residues appear slightly more exposed in the complex compared to the unbound structure when binding (suggesting a conformational change), highly influential residues show a slight positive burial (whereas non-predicted residues show much more exposure), (2) When antigen residues tend to lose exposure (i.e., become buried upon binding as expected), the highly influential antigen residues are more buried than average inferred interface residues (suggesting the model tends to prioritize the deeply buried core epitope residues).

While acknowledging that these results exhibit a promising biophysical signal (i.e., non-random) towards the interpretability aspect of our method, we prefer to explicitly mention a set of limitations which *may* have caused to introduce noise into the analysis as well: (1) While simple, gradient-based saliency analysis is more prone to noise and saturation effects, (2) alanine scanning assume additivity (i.e., does not capture epistatic effects between mutations), and (3) the poorly predicted complex structure from AlphaFold3 (with lower pTM and ipTM) may hinder the true binding interfaces and thus, limits the efficacy of the downstream analysis. We expect to explore more on this direction in our future works.

Appendix I. Timing Analysis

One of the most critical challenges in traditional wet-lab-based or molecular dynamics-based experiments for antibody-antigen binding affinity estimation is the time complexity. However, through deep learning-based approaches, it is possible to obtain an accurate prediction for the binding affinity within minutes or even seconds. Accordingly, our Combined-V2 model is able to predict the binding affinity of a given antibody-antigen pair within 1 minute, provided a graphic processing unit (GPU) and within 3 minutes on a CPU, as shown in Table I.11.

Metric	Gradient-based saliency analysis	In-silico alanine scanning
Residue overlap analysis (antibody/antigen)		
Precision↑	0.39/0.38	0.27/0.44
Recall↑	0.39/0.31	0.34/0.45
$F_1 \uparrow$	0.39/0.39	0.36/0.45
$p \downarrow$	0.5/0.45	0.5/0.24
SASA analysis (antibody/antigen)		
$\overline{\Delta SASA}_{all-interface}$	-6.65/17.34	-6.65/17.34
$\overline{\Delta SASA}_{predicted}$	2.51/44.15	3.91/27.16
$\overline{\Delta SASA}_{non-predicted}$	-8.55/11.40	-8.30/15.65

Table H.10: Results of similarity and overlap analysis between inferred interface residues (in the antibody-antigen complex) and the highly influential residues (in the unbound forms) identified by either gradient saliency method or alanine scanning.

Machine	Elapsed Time (s)	Avg. Inference Time (s)*
GPU (NVIDIA T4)	2694	22
CPU	18621	149

Table I.11: Timing analysis for the Combined-V2 model. *Here, the elapsed time refers to the time taken to run inference on a random sample of 125 antibody-antigen pairs from the P2PXML-PDB dataset. The times are expressed to the nearest second.

Appendix J. Limitations

Even though the proposed method exhibits the state-of-the-art performance in antibody-antigen binding affinity prediction, we observe that there are cases where the predictions considerably differ from the true targets, such as instances where the absolute error in the log domain exceeds the limit: 3, indicating a catastrophic prediction failure (Fig. 10). Even though such instances account for less than 0.32% in the total test set, we believe that the deep insights, which are potentially retrievable through extensively exploring the weight space of the model or categorically analyzing the corresponding input data, from such failures may reveal existing bottlenecks about the method's generalizability, which in-turn may be useful in future developments, especially when handling open-set antibody-antigen pairs.

Even though we present a weight space analysis (in Appendix H in the

appendix) as a way to partially address the interpretability of the method, we acknowledge that a full white-box analysis on explainability is presumably impossible given the current state and nature of the deep learning paradigms, given that an end-to-end, across-sample interpretability analysis is also beyond the scope of this work. Even so, we believe that such a thorough interpretability analysis would be beneficial in empirically bridging the knowledge gap between established biophysical principles and the learned patterns from the trained model, by drawing potential connections between them.

Appendix K. Future Works

Our future works will explore the following: (1) techniques to enhance the interpretability of the proposed method, potentially by analytically exploring the weight space while bridging connections with the established biophysical principles, (2) a thorough quantitative study detailing the generalizability of the proposed method across different antigen and antibody variants, and (3) possibility of extending the proposed method into other closely related tasks such as *de novo* antibody design.

Appendix L. Website, Project Page and Code Availability

The web platform is developed as a community access tool where anyone interested can obtain 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 for each model on the website is as follows:

Model	Input format
Combined-V2	PDB files
Final sequence-based model	Text sequences
Final structure-based model	PDB files

Table L.12: The input format for each model hosted on the website. The units of the output are the predicted binding affinity in $\mu g/ml$. 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 multimedia materials, including the curated datasets, demonstration videos and the codes, will be available on the project page: `https://drug-discovery-entc.github.io/p2pxml/`.