AsEP: Benchmarking Deep Learning Methods for Antibody-specific Epitope Prediction

Anonymous Authors[1]

Abstract

Epitope identification is vital for antibody design yet challenging due to the inherent variability in antibody. Additionally, the challenge is heightened by the lack of a consistent evaluation pipeline, limited dataset size and epitope diversity. Our contributions are two-fold. First, we provide the largest specialized epitope prediction dataset – AsEP, consisting of 1723 filtered antibody-antigen complexes. AsEP addressed the dataset diversity issue with clustered epitope groups. Second, most current methods for epitope prediction focus solely on antigen while few consider *both* antibody and antigen. Instead, we conceptualize the antibody-antigen interaction as bipartite graphs and formulate epitope prediction as link prediction tasks. Such formulation allows attributing model prediction to interaction types, providing more interpretability. Our method, WALLE, leverages protein language models for capturing sequence-level information and graph networks for incorporating structure information. WALLE outperforms existing models, achieving an MCC of 0.210 and roughly six times better than MaSIF-site. The curated dataset AsEP and our method WALLE are available to the research community, fostering open-source collaboration and advancement of the field.

1. Introduction

Antibodies are specialized proteins produced by our immune system to combat foreign substances called antigens. Their unique ability to bind with high affinity and specificity sets them apart from regular proteins and small-molecule drugs, making them increasingly popular in therapeutic engineering.

While the community is shifting toward computational antibody design (Jin et al., 2022; Zhou et al., 2024; Bennett et al., 2024) given a pre-determined epitope, accurate prediction of epitopes remains underexplored. Accurate identification of epitopes is beneficial for understanding antibody-antigen interactions, antibody function, and streamlining antibody engineering. This task remains challenging due to multiple factors (Akbar et al., 2022; Hummer et al., 2022), such as, the lack of comprehensive datasets, limited interpretability and generalizability. Available datasets are limited in size, up to 582 complexes from Bepipred-3.0 (Clifford et al., 2022), and feature disproportionate representation among various epitopes. Existing methods perform poorly on epitope prediction task (Cia et al., 2023), with a ceiling MCC (Matthew's Correlation Coefficient) of 0.06. However, recent advancements in graph-based learning, coupled with an increase in available antibody structures in the Protein Data Bank (PDB) (Berman et al., 2003), highlight the need to reevaluate current methods and establish a benchmark dataset for predicting antibody-antigen interactions.

We approach the problem as a bipartite graph link prediction task. Traditionally, graph link prediction focuses on identifying connections within the same graph, such as in protein-protein interaction networks. Our research extends this concept to bipartite graphs at the molecular level and proposes our own model, WALLE. Because existing methods generally predict protein binding sites or epitopes in antibody-antigen complexes rather than residue-residue interactions, we focus on benchmarking the node classification task while providing WALLE's performance on the bipartite link prediction task as a baseline.

2. Problem Formulation

Antibody-antigen interaction is important for analyzing protein structures. The problem can be formulated as a bipartite graph link prediction task. The inputs are two disjoint graphs, an antibody graph $G_A = (V_A, E_A)$ and an antigen graph $G_B = (V_B, E_B)$, where V_x is the vertice set for graph x and E_x is the edge set for graph x. Since the neural networks only take continuous values as input, we

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML) 2024 AI for Science workshop. Do not distribute.

also have a function h to encode each vertex into a vector $h : V \to \mathbb{R}^D$. The design of the encoding function depends on the methods. For example, h can be a one-hot encoding layer or pretrained embeddings given by a protein language model. We use different encoding functions for antibodies and antigens: $h_A : V_A \to \mathbb{R}^{D_A}$, and $h_B : V_B \to \mathbb{R}^{D_B}$.

In addition, $E_A \in \{0,1\}^{|V_A|\times|V_A|}$ and $E_B \in \{0,1\}^{|V_B|\times|V_B|}$ denote the adjacency matrices for the antibody and antigen graphs, respectively. In this work, the adjacency matrices are calculated based on the distance matrix of the residues. Each entry e_{ij} denotes the proximity between residue i and residue j; $e_{ij} = 1$ if the Euclidean distance between any non-hydrogen atoms of residue i and residue j is less than 4.5Å, and $e_{ij} = 0$ otherwise. The antibody graph G_A is constructed by combining the CDR residues from the heavy and light chains of the antibody, and the antigen graph G_B is constructed by combining the surface residues of the antigen. The antibody and antigen graphs are disjoint, i.e., $V_A \cap V_B = \emptyset$.

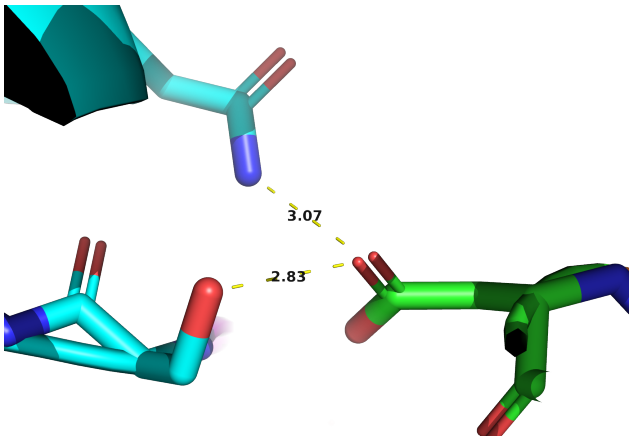


Figure 1. An example to illustrate interacting residues. The two dashed lines denote the distance between non-hydrogen atoms from different interacting residues from two different protein chains.

We consider two subtasks based on these inputs.

Epitope Prediction Epitopes are the regions on the antigen surface recognized by antibodies; in other words, they are a set of antigen residues in contact with the antibody and are determined from the complex structures using the same distance cutoff of 4.5Å as aforementioned. For a node in the antigen graph $v \in V_B$, if there exists a node in the antibody graph $u \in V_A$ such that the distance between them is less than 4.5Å, then v is an epitope node. Epitope nodes and the remaining nodes in G_B are assigned labels of 1 and 0, respectively. The first task is then a node classification within the antigen graph G_B given the antibody graph G_A.

This classification takes into account the structure of the antibody graph, G_A, mirroring the specificity of antibody-antigen binding interactions. Different antibodies can bind to various antigen locations, corresponding to varying subsets of epitope nodes in G_B. This question differs from conventional epitope prediction methods that do not consider the antibody structure and end up predicting the likelihood of the subset of antigen nodes serving as epitopes, such as ScanNet (Tubiana et al., 2022), MaSIF (Gainza et al., 2020).

The task is to develop a binary classifier $f : V_B \to \{0,1\}$ that takes both the antibody and antigen graphs as input and predicts the label for antigen nodes as formulated below:

$$f(v\,; G_B, G_A) = \begin{cases} 1 & \text{if } v \text{ is an epitope;} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Bipartite Link Prediction The second task takes it further by predicting concrete interactions between nodes in G_A and G_B, resulting in a bipartite graph that represents these antibody-antigen interactions. Moreover, this helps attribute the model performance to specific interactions at the molecular level and provide more interpretability. Accurately predicting these interactions is critical for understanding the binding mechanisms and for guiding antibody engineering.

We model the antibody-antigen interaction as a bipartite graph:

$$K_{m,n} = (V_A, V_B, E)$$

where $m = |V_A|$ and $n = |V_B|$ denote the number of nodes in the two graphs, respectively, and E denotes all possible inter-graph links. In this bipartite graph, a node from the antibody graph is connected to each node in the antigen graph via an edge $e \in E$. The task is then to predict the label of each bipartite edge. If the residues of a pair of nodes are located within 4.5Å of each other, referred to as *in contact*, the edge is labeled as 1; otherwise, 0. For any pair of nodes, denoted as (v_a, v_b) $\forall v_a \in V_A, v_b \in V_B$, the binary classifier $g : K_{m,n} \to \{0,1\}$ is formulated as below:

$$g(v_a, v_b; K_{m,n}) = \begin{cases} 1 & \text{if } v_a \text{ and } v_b \text{ are in contact} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

3. Benchmark Construction

Antibodies are composed of two heavy chains and two light chains, each of which contains a variable domain (areas of high sequence variability) composed of a variable heavy (VH) and a variable light (VL) domain responsible for antigen recognition and binding (Chothia & Lesk, 1987). These domains have complementarity-determining regions (CDR, Figure 2 top blue, yellow, and red regions), which are the primary parts of antibodies responsible for antigen recognition and binding.

3.1. Antibody-antigen complexes

We sourced our initial dataset from the Antibody Database (AbDb) (Ferdous & Martin, 2018), dated 2022/09/26, which contains 11,767 antibody files originally collected from the Protein Data Bank (PDB) (Berman et al., 2003). We extracted conventional antibody-antigen complexes that have a VH and a VL domain with a single-chain protein antigen, and there are no unresolved CDR residues due to experimental errors, yielding 4,081 antibody-antigen complexes. To ensure data balance, we removed identical complexes using an adapted version of the method described in Krawczyk et al. (2014). We clustered the complexes by antibody heavy and light chains followed by antigen sequences using MMseqs2 (Steinegger & Söding, 2017). We retained only one representative complex for each unique cluster, leading to a refined dataset of 1,725 unique complexes. Two additional complexes were manually removed; CDR residues in the complex `6jmr_1P` are unknown (labeled as 'UNK') and it is thus impossible to build graph representations upon this complex; `7sgm_0P` was also removed because of non-canonical residues in its CDR loops. The final dataset consists of 1,723 antibody-antigen complexes. For detailed setup and processing steps, please refer to Appendix A.2.

3.2. Convert antibody-antigen complexes to graphs

These 1,723 files were then converted into graph representations, which are used as input for WALLE. In these graphs, each protein residue is modeled as a vertex. Edges are drawn between pairs of residues if any of their non-hydrogen atoms are within 4.5Å of each other, adhering to the same distance criterion used in PECAN (Pittala & Bailey-Kellogg, 2020).

Exclude buried residues In order to utilize structural information effectively, we focused on surface residues, as only these can interact with another protein. Consequently, we excluded buried residues, those with a solvent-accessible surface area of zero, from the antigen graphs. The solvent-accessible surface areas were calculated using DSSP (Kabsch & Sander, 1983) via Graphein (Jamasb et al., 2021). It is important to note that the number of interface nodes are much smaller than the number of non-interface nodes in the antigen, making the classification task more challenging.

Exclude non-CDR residues We also excluded non-CDR residues from the antibody graph, as these are typically not involved in antigen recognition and binding. This is in line with the approach adopted by PECAN (Pittala & Bailey-Kellogg, 2020) and EPMP (Vecchio et al., 2021). Figure 2 provides a visualization of the processed graphs.

Node embeddings To leverage the state-of-the-art protein language models, we generated node embeddings for each residue in the antibody and antigen graphs using AntiBERTy (Ruffolo et al., 2021) (via IgFold (Ruffolo et al., 2023) package) and ESM2 (Lin et al., 2022) (`esm2_t12_35M_UR50D`) models, respectively. In our dataset interface package, we also provide a simple embedding method using one-hot encoding for amino acid residues. Other node embedding methods can be easily incorporated into our dataset interface. Please refer to Figure S1 for a schematic view of our processing pipeline.

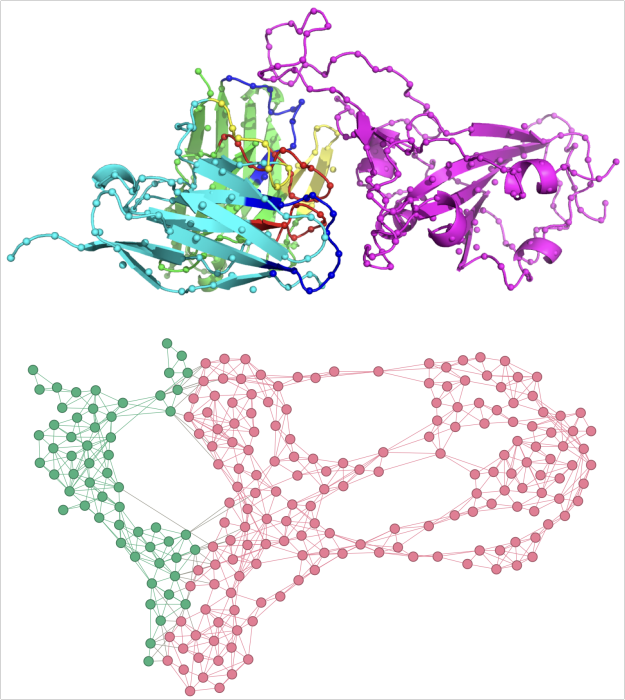


Figure 2. Graph visualization of an antibody-antigen complex. **Top**: the molecular structure of an antibody complexed with the receptor binding domain of SARS-Cov-2 virus (PDB code: 7KFW), the antigen. Spheres indicate the alpha carbon atoms of each amino acid. Color scheme: the antigen is colored in magenta, the framework region of the heavy and light chains is colored in green and cyan and CDR 1-3 loops are colored in blue, yellow, and red, respectively. **Bottom**: the corresponding graph. Green vertices are antibody CDR residues and pink vertices are antigen surface residues.

3.3. Dataset split

We propose two types of dataset split settings. The first is a random split based on the ratio of epitope to antigen surface residues, $\frac{\#\text{epitope nodes}}{\#\text{antigen nodes}}$; the second is a more challenging setting where we split the dataset by epitope groups. The first setting is straightforward and is used to evaluate the performance of WALLE along with the other four methods. The second setting is more challenging because it requires the model to generalize to unseen epitope groups. We describe the two settings in detail below.

Split by epitope to antigen surface ratio As aforementioned, the number of non-interface nodes in the antigen graph is much larger than the number of interface nodes. While epitopes usually have a limited number of residues, typically around 14.6 ± 4.9 amino acids (Reis et al., 2022), the antigen surface may extend to several hundred or more residues. The complexity of the classification task therefore increases with the antigen surface size. To ensure similar complexity among train, validation, and test sets, we stratified the dataset to include a similar distribution of epitope to non-epitope nodes in each set. Table 1 shows the distribution of epitope-to-antigen surface ratios in each set. This led to 1383 antibody-antigen complexes for the training set and 170 complexes each for the validation and test sets. The list of complexes in each set is provided in the Supplementary Table `SI-split-epitope-ratio.csv`.

Table 1. Distribution of epitope to antigen surface nodes in each set.

Epi/Surf	Training	Validation/Test
0, 5%	320 (23%)	40 (24%)
5% , 10%	483 (35%)	60 (35%)
10%, 15%	305 (22%)	38 (22%)
15%, 20%	193 (14%)	24 (14%)
20%, 25%	53 (4%)	6 (4%)
25%, 30%	19 (1%)	2 (1%)
30%, 35%	8 (0.6%)	0 (-)
35%, 40%	2 (0.1%)	0 (-)
sum	1383	170

Split by epitope groups This is motivated by the fact that antibodies are highly diverse in the CDR loops and by changing the CDR sequences it is possible to engineer novel antibodies to bind different sites on the same antigen. This was previously observed in the EpiPred dataset where Krawczyk et al. (2014) tested the specificity of their method on five antibodies associated with three epitopes on the same antigen, hen egg white lysozyme.

We inlcude 641 unique antigens and 973 epitope groups in our dataset. Figure 3 shows two examples of multi-epitope antigens in our dataset, hen egg white lysozyme (Figure 3a) and spike protein (Figure 3b). Specifically, there are 52 and 64 distinct antibodies in our dataset that bind to hen egg white lysozyme and spike protein, respectively. For visual clarity, we only show five and sixteen antibodies in Figure 3a and Figure 3b.

We can see that different antibodies bind to different locations on the same antigen. Details of all epitope groups are provided in the Supplementary Table `SI-AsEP-entries.csv` with annotation provided in Appendix A.5. We then split the dataset into train, validation, and test sets such that the epitopes in the test set are not found in either train or validation sets. We followed an 80%/10%/10% split for the number of complexes in each set. This resulted in 1383 complexes for the training set and 170 complexes for the validation and test sets. The list of complexes in each set is provided in the Supplementary Table `SI-split-epitope-group.csv`.

3.4. Evaluation

In this work, we focus on the epitope prediction task. We evaluate the performance of each method using consistent metrics. Matthew's Correlation Coefficient (MCC) is highly recommended for binary classification assessments (Matthews, 1975) and is especially advocated for its ability to provide equal weighting to all four values in the confusion matrix, making it a more informative metric about the classifier's performance at a given threshold than other metrics (Chicco & Jurman, 2020; 2023). We encourage the community to adopt MCC for the epitope prediction task as it takes into account true and false positives, as well as true and false negatives, offering a comprehensive measure of the performance. It is considered superior to the AUC-ROC, which is evaluated over all thresholds. For consistency, we also included Precision and Recall from prior studies EpiPred (Krawczyk et al., 2014) and PECAN (Pittala & Bailey-Kellogg, 2020), and we added Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and F1 score, both are typical binary classification metrics. For methods that predict antibody-antigen complex structures, we determine the epitopes using the same distance criterion as aforementioned.

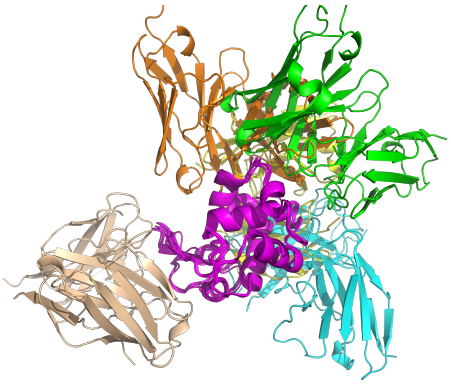
3.5. Dataset interface

We implemented a Python package interface for our dataset using PyTorch Geometric (Fey & Lenssen, 2019). Users can load the dataset as a PyTorch Geometric dataset object and use it with PyTorch Geometric's data loaders. We provide an option to load node embeddings derived from AntiBERTy and ESM2 or simply one-hot embeddings. Each data object in the dataset is a pair of antibody and antigen graphs; both node- and edge-level labels are provided, and the node-level labels are used for the epitope prediction task. The dataset will be made available upon acceptance.

4. WALLE: a graph-based method for epitope prediction

Alongside our dataset interface, we also provide a graph-based model named WALLE. It takes as input a pair of antibody and antigen graphs and makes node- and edge-level predictions.

Preprocessing Unlike previous studies, such as PECAN (Pittala & Bailey-Kellogg, 2020), that construct features



(a) Five different antibodies bound to hen egg white lysozyme. Complexes are superimposed on the antigen structure (magenta). AbDb IDs of the complexes and their color: 1g7i_0P (green), 2yss_0P (cyan), 1dzb_1P (yellow), 4tsb_0P (orange), 2iff_0P (wheat). Antigens are colored in magenta.

(b) Sixteen different antibodies bound to coronavirus spike protein. Complexes are superimposed on the antigen structure (magenta) and antibodies are in different colors. AbDb IDs of the complexes: 7k8s_0P, 7m7w_1P, 7d0b_0P, 7dzy_0P, 7ey5_1P, 7jv4_0P, 7k8v_1P, 7kn4_1P, 7lqw_0P, 7n8i_0P, 7q9i_0P, 7rq6_0P, 7s0e_0P, 7upl_1P, 7wk8_0P, 7wpd_0P.

Figure 3. Examples of different antibodies binding to the same antigen.

from scratch using methods like Position-Specific Scoring Matrix (PSSM) derived from BLAST (Altschul et al., 1990), our model leverages state-of-the-art protein language models. We utilize AntiBERTy (Ruffolo et al., 2021) to generate embeddings for antibody and ESM2 (Lin et al., 2022) for antigen sequences, referred to as *sequence embeddings*. These embeddings encapsulate rich, context-specific information, offering a more nuanced representation of the protein structures than traditional methods.

As aforementioned, we only keep the surface residues for the antigen graph and the CDR residues for the antibody graph. We then map these residues using the sequence models to generate the *node embeddings* for the antibody and antigen graphs. The edges are also calculated among these residues using the same distance cutoff of 4.5Å. This step is shown as the *Preprocessing* modules in Figure 4.

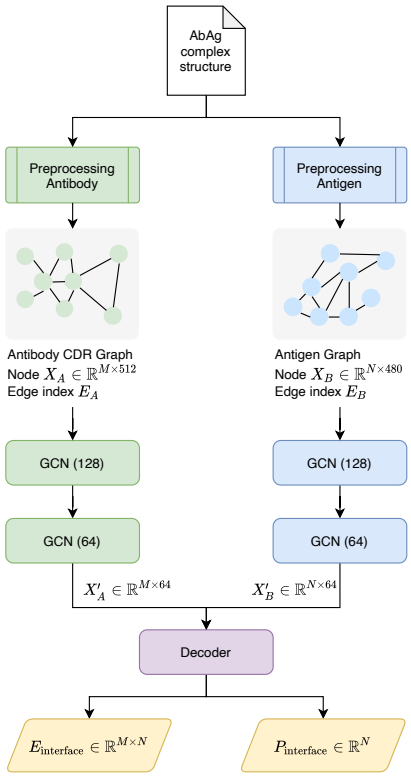


Figure 4. A schematic of the preprocessing step that turns an input antibody-antigen complex structure into a graph pair and the model architecture of WALLE.

Graph Modules The architecture of WALLE incorporates graph modules that process the input graphs of antibody and antigen structures, as depicted in Figure 3. Inspired by PECAN and EPMP, our model treats the antibody and antigen graphs separately, with distinct pathways for each. The antibody graph is represented by node embeddings X_A

with a shape of $(M, 512)$ and an adjacency matrix E_A, while the antigen graph is described by node embeddings X_B with a shape of $(N, 480)$ and its corresponding adjacency matrix E_B. The embedding sizes are consistent with AntiBERTy and ESM2 (`esm2_t12_35M_UR50D`).

Both antibody and antigen graph nodes are first projected into the dimensionality of 128 using fully connected layers. The resulting embeddings are then passed through two Graph Convolutional Network (GCN) modules consecutively to refine the features and yield updated node embeddings X'_A and X'_B with a reduced dimensionality of $(M, 64)$. The output from the first GCN layer is passed through a ReLU activate function. Outputs from the second GCN layer are directly fed into the Decoder module. These GCNs operate independently, each with its own parameters, ensuring that the learned representations are specific to the antibody or the antigen.

The use of separate GCN modules for the antibody and antigen allows for the capture of unique structural and functional characteristics pertinent to each molecule before any interaction analysis. This design choice aligns with the understanding that the antibody and antigen have distinct roles in their interactions and their molecular features should be processed separately.

Decoder

We used a simple decoder to predict the binary labels of edges between the antibody and antigen graphs. It takes a pair of node embeddings output by the graph modules as input; calculates the inner product or passes through a linear layer; and converts the logits through a sigmoid activation function to obtain the predicted probability of each edge. An edge is assigned a binary label of 1 if the predicted probability is greater than 0.5 or 0 otherwise. This is shown as the *Decoder* module in Figure 4. For the epitope prediction task, we convert edge-level predictions to node-level by summing the predicted probabilities of all edges connected to an antigen node; we assign the antigen node a label of 1 if the number of connected edges is greater than a threshold or 0 otherwise. The threshold is treated as a hyperparameter and is optimized in the experiments.

Implementation

We used PyTorch Geometric (Fey & Lenssen, 2019) framework to build our model. The graph modules are implemented using the *GCNConv* module from PyTorch Geometric. We trained the model to minimize a loss function consisting of two parts: a weighted binary cross-entropy loss for the bipartite graph link reconstruction and a regularizer for the number of positive edges in the reconstructed bipartite graph. We used the same set of hyperparameters and loss functions for both dataset settings. The loss function and hyperparameters are described in detail in Appendix A.4.

5. Prior State-of-the-art

We compared WALLE with the other four selected methods. These methods were selected because they are representative and use different features and approaches to predict epitopes. Features of these methods are summarized in Table 2. **EpiPred** (Krawczyk et al., 2014) is a systematic approach designed exclusively for epitope prediction. It takes a pair of antibody and antigen structures as input and performs local docking to generate multiple candidate complexes, which are then scored and ranked by a specific antibody-antigen scoring function. The decoy with the highest score is selected as the final prediction for benchmarking. We ran the method using the default settings and parameters provided by the authors. **ESMFold** (Lin et al., 2023) is a protein language model based on the protein language model, EMS2 (Lin et al., 2023). Its folding head was trained on over 325 thousand protein structures and reported to achieve similar performance as AlphaFold2 (Jumper et al., 2021) on both single-chain and multimeric proteins. Because it is much faster than AlphaFold2, we included it in our benchmarking. While both EpiPred and ESMFold take antibodies and antigens as input. We also include two methods that only consider antigens as input for comparison. These methods were designed for predicting binding sites on general protein sequences and structures. They are not specifically designed for epitope prediction but are representative of methods that only consider antigens as input. **ESMBind** (Schreiber, 2023) is a language model based on ESM2(Lin et al., 2023); we reproduced the fine-tuning procedures as documented using Low-Rank Adaptation (Hu et al., 2021) on a dataset of general protein sequences with binding sites annotated. It takes a single protein sequence as input and predicts protein binding sites. **MaSIF-site** (Gainza et al., 2020) is a geometric deep learning-based method that predicts protein binding sites on a given protein structure surface. It converts the antigen surface into a mesh graph, with each mesh vertex encoded with geometric and physicochemical descriptors, and uses a graph neural network to predict binding sites on the antigen surface. While its output is the probabilities of each mesh vertex being a binding site rather than the residues, we mapped mesh vertices with predicted probability $> 70\%$ to antigen residues that are within 1.2Å of the mesh vertex and considered them as predicted epitope residues.

6. Discussion and future work

Experiment results We evaluated each method for both dataset split settings on the test set using the metrics described in Section 3.4. Table 3a and Table 3b summarize the average performance metrics across the test set samples. WALLE generally shows better performance among all metrics except for recall for both dataset splits. We also provided the baseline performance for bipartite link

Table 2. Summary of Features Used in Benchmarking Methods.

	Antibody	Structure	PLM	Graph
WALLE	✓	✓	✓	✓
EpiPred	✓	✓	×	✓
ESMFold	✓	×	✓	×
MaSIF-site	×	✓	×	✓
ESMBind	×	×	✓	×

Antibody: Antibody is taken into consideration when predicting epitope nodes;
Structure: Topological information from protein structures;
PLM: Representation from Protein Language Models;
Graph: Graph representation of protein structures.

prediction in Table S5.

We also carried out ablation studies (Appendix C) to investigate the impact of different components of WALLE. When we replace the graph convolutional layers with fully connected layers, its performance degenerates considerably, suggesting that the graph convolutional layers contribute to the model's performance. This is related to the fact that the interaction between a pair of protein structures is dependent on the spatial arrangement of the residues as discussed by Reis et al. (2022) that the interface polar bonds, a major source of antibody specificity, tend to shield interface hydrophobic clusters. In addition, the language model embeddings also contribute to the model's performance, as performance drops when they are replaced by one-hot or BLOSUM62 embeddings. Finally, we also investigated whether the choice of language model affects the model's performance. We found that the model using AntiBERTy and ESM2 embeddings for antibodies and antigens performed slightly better than the model using ESM2 embeddings for both antibodies and antigens. This suggests that the choice of language model may impact the model's performance, but a model like ESM2, which is trained on general protein sequences, may contain sufficient information for the epitope prediction task.

While WALLE outperforms other methods in the second dataset split setting, its performance degenerated considerably from the first dataset split setting. This suggests that WALLE is likely biased toward the epitopes in the training set and does not generalize well to unseen epitopes. The performance of the other four methods is not ideal for this task. We believe this would require a more sophisticated model architecture and a more comprehensive dataset to improve the performance of epitope prediction. We leave this for future work.

Related work Two relevant works are PECAN (Pittala & Bailey-Kellogg, 2020) and EPMP (Vecchio et al., 2021). While the graph construction method in WALLE is inspired by both studies, we differentiate from them in the following aspects. Firstly, we used a different node embedding method. Both PECAN and EPMP generated node embeddings using a position-specific scoring matrix (PSSM) derived from sequence alignment. In contrast, we used pre-trained language models to generate node embeddings. Secondly, we simplified the graph decoder by using an inner product or a linear layer to predict the binary labels of edges between the antibody and antigen graphs. In contrast, they used graph attention networks to predict the binary labels of nodes in the antigen graph.

While there exist other related datasets with similar objectives, our dataset is the biggest and most diverse in terms of antibody-antigen complexes; there also exists work (Zhao et al., 2024) that benchmarked other methods complementary to our work, including docking methods and Alphafold-Multimer(Evans et al., 2022) on their dataset of antibody-antigen complexes. They came to the same conclusion that existing methods developed for general protein complexes need improvement to be effective on antibody-antigen complexes. For details, we refer readers to Appendix A.1.

Edge features In terms of structure representation, we only used a simple invariant edge feature, the distance matrix, to capture the neighborhood information of each residue. This topological descriptor already performs better than other methods that use sequence-based features. For future work, more edge features can be incorporated to enrich the graph representation, in addition to the invariant edge features used in this work, such as inter-residue distances and edge types used in GearNet (Zhang et al., 2023), and SE3 equivariant features, such as rotational and orientational relationships between residues as used in abdockgen (Jin et al., 2022).

Antibody types We also plan to extend our work to include other types of antibodies. Current work only looks at conventional antibodies, consisting of heavy- and light-chain variable domains. There are also an increasing number of novel antibodies, for example nanobodies, which are single-variable-domain antibodies derived from camelids. These will be included in future work.

7. Conclusion

In this work, we proposed a novel benchmarking dataset for the epitope prediction task and clustered the samples by epitopes. We also provided a model, WALLE, which combines protein language models and graph neural networks to leverage their abilities in capturing amino acid contextual and geometric information. We benchmarked WALLE and four other methods, showing that while WALLE outperforms existing methods on both tasks, there remains room for improvement. We also discussed possible future directions. This work can serve as a starting point for future research in this area.

Table 3. Performance on test set from dataset split by epitope to antigen surface ratio and epitope groups.

(a) Performance on dataset split by epitope to antigen surface ratio.

Algorithm	MCC	Precision	Recall	AUCROC	F1
WALLE	**0.210** (0.020)	**0.235** (0.018)	**0.422** (0.028)	**0.635** (0.013)	**0.258** (0.018)
EpiPred	0.029 (0.018)	0.122 (0.014)	0.180 (0.019)	—	0.142 (0.016)
ESMFold	0.028 (0.010)	0.137 (0.019)	0.043 (0.006)	—	0.060 (0.008)
ESMBind	0.016 (0.008)	0.106 (0.012)	0.121 (0.014)	0.506 (0.004)	0.090 (0.009)
MaSIF-site	0.037 (0.012)	0.125 (0.015)	0.183 (0.017)	—	0.114 (0.011)

(b) Performance on dataset split by epitope groups.

Algorithm	MCC	Precision	Recall	AUCROC	F1
WALLE	**0.077** (0.015)	0.143 (0.017)	**0.266** (0.025)	**0.544** (0.010)	**0.145** (0.014)
EpiPred	-0.006 (0.015)	0.089 (0.011)	0.158 (0.019)	—	0.112 (0.014)
ESMFold	0.018 (0.010)	0.113 (0.019)	0.034 (0.007)	—	0.046 (0.009)
ESMBind	0.002 (0.008)	0.082 (0.011)	0.076 (0.011)	0.500 (0.004)	0.064 (0.008)
MaSIF-site	0.046 (0.014)	**0.164** (0.020)	0.174 (0.015)	—	0.128 (0.012)

MCC: Matthews Correlation Coefficient; **AUCROC**: Area Under the Receiver Operating Characteristic Curve; **F1**: F1 score. Standard errors are included in the parentheses. We omitted the results of EpiPred, ESMFold and MaSIF-site for AUCROC. For EpiPred and ESMFold, the interface residues are determined from the predicted structures by these methods such that the predicted values are binary and not comparable to other methods; As for MaSIF-site, it outputs the probability of mesh vertices instead of node probabilities and epitopes are determined as residues close to mesh vertices with probability greater than 0.7.

8. Data and Code Availability

Our dataset, the code for our dataset interface and the baseline models are provided in the Supplementary zip file for review.

- The code of our dataset interface and our baseline model will be publically accessible after the manuscript is accepted.
- The Supplementary Table *SI-AsEP-entries.csv* groups the antibody-antigen complexes by epitope group as well as antibody CDR sequences is provided in the Supplementary Materials. An example is provided in Appendix A.5.
- The two dataset splits files are also provided in the Supplementary Materials, *SI-split-epitope-group.csv* and *SI-split-epitope-ratio.csv*

References

Akbar, R., Bashour, H., Rawat, P., Robert, P. A., Smorodina, E., Cotet, T.-S., Flem-Karlsen, K., Frank, R., Mehta, B. B., Vu, M. H., Zengin, T., Gutierrez-Marcos, J., Lund-Johansen, F., Andersen, J. T., and Greiff, V. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs*, 14(1):2008790, 2022. doi: 10.1080/19420862.2021.2008790. URL https://doi.org/10.1080/19420862.2021.2008790. PMID: 35293269.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(05)80360-2. URL https://www.sciencedirect.com/science/article/pii/S0022283605803602.

Basu, S. and Wallner, B. Dockq: A quality measure for protein-protein docking models. *PLOS ONE*, 11(8):1–9, 08 2016. doi: 10.1371/journal.pone.0161879. URL https://doi.org/10.1371/journal.pone.0161879.

Bennett, N. R., Watson, J. L., Ragotte, R. J., Borst, A. J., See, D. L., Weidle, C., Biswas, R., Shrock, E. L., Leung, P. J. Y., Huang, B., Goreshnik, I., Ault, R., Carr, K. D., Singer, B., Criswell, C., Vafeados, D., Garcia Sanchez, M., Kim, H. M., Vázquez Torres, S., Chan, S., and Baker, D. Atomically accurate de novo design of single-domain antibodies. *bioRxiv*, 2024. doi: 10.1101/2024.03.14.585103. URL https://www.biorxiv.org/content/early/2024/03/18/2024.03.14.585103.

Berman, H., Henrick, K., and Nakamura, H. Announcing

the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, Dec 2003. ISSN 1545-9985. doi: 10.1038/nsb1203-980. URL https://doi.org/10.1038/nsb1203-980.

Chicco, D. and Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, Jan 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7. URL https://doi.org/10.1186/s12864-019-6413-7.

Chicco, D. and Jurman, G. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4, Feb 2023. ISSN 1756-0381. doi: 10.1186/s13040-023-00322-4. URL https://doi.org/10.1186/s13040-023-00322-4.

Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196 (4):901–917, 1987. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(87)90412-8. URL https://www.sciencedirect.com/science/article/pii/0022283687904128.

Cia, G., Pucci, F., and Rooman, M. Critical review of conformational B-cell epitope prediction methods. *Briefings in Bioinformatics*, 24(1):bbac567, 01 2023. ISSN 1477-4054. doi: 10.1093/bib/bbac567. URL https://doi.org/10.1093/bib/bbac567.

Clifford, R. et al. Bepipred-3.0: Improved prediction of b-cell epitopes using protein sequence and structure data. *Protein Science*, 2022:4497, 2022. doi: 10.1002/pro.4497.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034. URL https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034.

Falkner, S., Klein, A., and Hutter, F. Bohb: Robust and efficient hyperparameter optimization at scale, 2018.

Ferdous, S. and Martin, A. C. R. AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database*, 2018:bay040, 04 2018. ISSN 1758-0463. doi: 10.1093/database/bay040. URL https://doi.org/10.1093/database/bay040.

Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric, 2019.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2): 184–192, Feb 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0666-6. URL https://doi.org/10.1038/s41592-019-0666-6.

Gao, M. et al. Deep learning-based method for predicting antibody-antigen binding residues with sequence information. *Computational and Biomedical Sciences*, 2022: 106064, 2022. doi: 10.1016/j.compbiomed.2022.106064.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992. doi: 10.1073/pnas.89.22.10915. URL https://www.pnas.org/doi/10.1073/pnas.89.22.10915.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Hummer, A. M., Abanades, B., and Deane, C. M. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, 2022. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2022.102379. URL https://www.sciencedirect.com/science/article/pii/S0959440X22000586.

Jamasb, A. R., Viñas, R., Ma, E. J., Harris, C., Huang, K., Hall, D., Lió, P., and Blundell, T. L. Graphein - a python library for geometric deep learning and network analysis on protein structures and interaction networks. *bioRxiv*, 2021. doi: 10.1101/2020.07.15.204701. URL https://www.biorxiv.org/content/early/2021/10/12/2020.07.15.204701.

Jin, W., Barzilay, R., and Jaakkola, T. Antibody-antigen docking and design via hierarchical equivariant refinement, 2022.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, Aug 2021. ISSN 1476-4687. doi:

10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: https://doi.org/10.1002/bip.360221211. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360221211.

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S. The cluspro web server for protein–protein docking. *Nature Protocols*, 12(2):255–278, Feb 2017. ISSN 1750-2799. doi: 10.1038/nprot.2016.169. URL https://doi.org/10.1038/nprot.2016.169.

Krawczyk, K., Liu, X., Baker, T., Shi, J., and Deane, C. M. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, 30(16):2288–2294, 04 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu190. URL https://doi.org/10.1093/bioinformatics/btu190.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Mahajan, S., Yan, Z., Jespersen, M. C., Jensen, K. K., Marcatili, P., Nielsen, M., Sette, A., and Peters, B. Benchmark datasets of immune receptor-epitope structural complexes. *BMC Bioinformatics*, 20(1):490, Oct 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3109-6. URL https://doi.org/10.1186/s12859-019-3109-6.

Martin, A. C., Cheetham, J. C., and Rees, A. R. Molecular modeling of antibody combining sites. *Methods Enzymology*, 203:121–153, 1991. ISSN 0076-6879. doi: https://doi.org/10.1016/0076-6879(91)03008-5. URL https://www.sciencedirect.com/science/article/pii/0076687991030085.

Matthews, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

Myung, J. et al. Csm-ab: A deep learning-based approach for predicting antibody-antigen binding sites. *Bioinformatics*, 2021:btab762, 2021. doi: 10.1093/bioinformatics/btab762.

Pierce, B. G., Hourai, Y., and Weng, Z. Accelerating protein docking in zdock using an advanced 3d convolution library. *PLOS ONE*, 6(9):1–6, 09 2011. doi: 10.1371/journal.pone.0024657. URL https://doi.org/10.1371/journal.pone.0024657.

Pittala, S. and Bailey-Kellogg, C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, 36(13):3996–4003, 04 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa263. URL https://doi.org/10.1093/bioinformatics/btaa263.

Raghavan, A. K. and Martin, A. C. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Molecular Immunology*, 45:3832–3839, 2008. doi: 10.1016/j.molimm.2008.05.022.

Reis, P. B. P. S., Barletta, G. P., Gagliardi, L., Fortuna, S., Soler, M. A., and Rocchia, W. Antibody-antigen binding interface analysis in the big data era. *Frontiers in Molecular Biosciences*, 9, 2022. ISSN 2296-889X. doi: 10.3389/fmolb.2022.945808. URL https://www.frontiersin.org/articles/10.3389/fmolb.2022.945808.

Ruffolo, J. A., Gray, J. J., and Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning, 2021.

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies, Apr 2023.

Schreiber, A. Esmbind and qbind: Lora, qlora, and esm-2 for predicting binding sites and post translational modification. *bioRxiv*, 2023. doi: 10.1101/2023.11.13.566930. URL https://www.biorxiv.org/content/early/2023/11/14/2023.11.13.566930.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1):539, 2011. doi: https://doi.org/10.1038/msb.2011.75. URL https://www.embopress.org/doi/abs/10.1038/msb.2011.75.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, Nov 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Sun, X. et al. Sagerank: A deep learning approach to predict antigenic epitopes using sequence and structure data. *Preprint at bioRxiv*, 2023. doi: 10.1101/2023.10.11.561985.

Tubiana, J., Schneidman-Duhovny, D., and Wolfson, H. J. Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, 19(6):730–739, Jun 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01490-7. URL https://doi.org/10.1038/s41592-022-01490-7.

Vecchio, A. D., Deac, A., Liò, P., and Veličković, P. Neural message passing for joint paratope-epitope prediction, 2021.

Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S.-Y. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Research*, 45(W1):W365–W373, 05 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx407. URL https://doi.org/10.1093/nar/gkx407.

Yeturu, K. and Chandra, N. Pocketmatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, 9(1):543, Dec 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-543. URL https://doi.org/10.1186/1471-2105-9-543.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining, 2023.

Zhao, N., Han, B., Zhao, C., Xu, J., and Gong, X. ABAG-docking benchmark: a non-redundant structure benchmark dataset for antibody–antigen computational docking. *Briefings in Bioinformatics*, 25(2):bbae048, 02 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae048. URL https://doi.org/10.1093/bib/bbae048.

Zhou, X., Xue, D., Chen, R., Zheng, Z., Wang, L., and Gu, Q. Antigen-specific antibody design via direct energy-based preference optimization, 2024.

A. Appendix-A

A.1. Related work

Comparison of Previous Datasets We would like to highlight our dataset, AsEP, is the largest curated AbAg benchmarking dataset to date. Existing ones either focus on general protein-protein complexes designed to develop general docking methods or are way smaller than AsEP if designed for AbAg interaction research. We summarized the sizes of existing datasets in the following table.

Table S1. Comparison of Dataset Sizes Across Different Methods

Method	Dataset Size
WALLE (AsEP)	1723 AbAg complexes
Gao et al. 2022 (Gao et al., 2022)	258 AbAg complexes
CSM-AB (Myung et al., 2021)	472 AbAg complexes
SAGERank (Sun et al., 2023)	287 AbAg complexes
Bepipred3.0 (Clifford et al., 2022)	582 AbAg complexes

SCEptRe (Mahajan et al., 2019) is a related dataset that keeps a weekly updated collection of 3D complexes of epitope and receptor pairs, for example, antibody-antigen, TCR-pMHC, and MHC-ligand complexes derived from the Immune Epitope Database (IEDB). Our approach for clustering antibody-antigen complexes regarding their epitopes is similar to theirs, with the difference in the clustering strategy. We cluster by antigen, then epitope group, and we allow mutated amino acids in the same epitope region because we turn the epitope sites into columns in the multiple sequence alignment. In contrast, SCEptRe clusters by antibody and then compares epitope similarity by epitope conformation using atom-pair distances via PocketMatch (Yeturu & Chandra, 2008), which is beneficial for comparing the function of various paratopes but is less suitable for our task of predicting epitope residues.

Complementary surveys We found two studies complementary to our work. Zhao et al. (2024) benchmarked a different type of methods, i.e., docking methods including ZDOCK (Pierce et al., 2011), ClusPro (Kozakov et al., 2017), and HDOCK (Yan et al., 2017) as well as Alphafold-Multimer (Evans et al., 2022) on their benchmark set of antibody-antigen complexes. Their benchmark includes 112 antibody-antigen complexes released after 1 June 2019, 98 conventional antibodies (the same antibody type as in our work), and 14 single-domain antibodies (not included in our work). They defined a prediction as successful if a method can predict a complex structure in top k decoys at an *Acceptable* or better defined by DockQ (Basu & Wallner, 2016). They showed that all docking methods gave a success rate at most 8.0% if using the top 5 decoys; Alphafold-Multimer showed a relatively better performance with a 15.3% success rate. While we focus on bipartite graph linkage prediction, different from their evaluation, both showed that there is room for improvement in the field of antibody-antigen complex prediction.

Another study by Cia et al. (2023) focuses on epitope prediction using a dataset of 268 antibody-antigen complexes. They differ from us in the definition of epitope residues in that they determined epitope residues using a change of at least 5% in relative solvent accessibility upon complex formation. They benchmarked sequence-based methods, structure-based methods, and an antibody-specific method, EpiPred, as we did in this work. Their findings agreed with ours that existing methods for the epitope prediction task are not sufficient for the task.

Sequence-based epitope predictor We also tested purely sequence-based epitope prediction tool, for example, Bepipred3.0 (Clifford et al., 2022) on our dataset. Bepipred3.0 uses ESM2 model, `esm2_t33_650M_UR50D` to generate sequence embeddings and was trained on a smaller dataset of 582 antibody-antigen structures and evaluated on 15 antibody-antigen complexes. The authors provided a relatively larger evaluation of linear B-cell epitopes derived from the Immune Epitope Database and reported an AUC-ROC of 0.693 on the top 10% of the predicted epitopes. We tested Bepipred3.0 on our dataset and found its performance degenerates significantly, as shown in the table below. This is not surprising because linear epitopes are consecutive positions in an antigen sequence, and this task fits better with language model design. Additionally, as pointed out by the authors, approximately 90% of epitopes (B-cell) fall into the conformational category (Clifford et al., 2022), which highlights the importance of the present benchmark dataset composed of conformational epitopes derived from filtered antibody-antigen structures. We believe these results underline the findings in our paper, showing that large language models alone, even if specialized for antibody-antigen interactions, do not encompass

all the relevant information needed for epitope prediction.

Confidence Threshold	Top 10%	Top 30%	Top 50%	Top 70%	Top 90%
AUC	0.693392	0.693392	0.693392	0.693392	0.693392
Balanced Accuracy	0.573132	0.636365	0.638755	0.604556	0.542274
MCC	0.109817	0.140183	0.134689	0.113372	0.071876
Precision-Recall AUC	0.176429	0.176429	0.176429	0.176429	0.176429
Accuracy	0.850178	0.701202	0.536947	0.362489	0.179051
Precision	0.169202	0.141361	0.120547	0.104286	0.090441
Recall	0.236760	0.553204	0.756607	0.892628	0.977723
F1-Score	0.173370	0.208366	0.197153	0.179294	0.160151

Table S2. Bepipred3.0 results for the presented AsEP dataset. The distance cutoff was changed to 4.0Å, as this is the threshold used by Bepipred3.0. Results are shown for five confidence thresholds as described in the BepiPred-3.0 paper. Across all stringency settings and metrics, Bepipred scored lower than Walle. Furthermore, it is possible that some of the structures within the dataset are contained within the Bepipred3.0 dataset, artificially increasing scores.

A.2. Steps to build antibody-antigen complex dataset

We sourced our initial dataset from AbDb (version dated September 26, 2022), containing 11,767 antibody files originally collected from the Protein Data Bank (PDB). We collected complexes numbered in Martin scheme (Raghavan & Martin, 2008) and used AbM CDR definition (Martin et al., 1991) to identify CDR residues from the heavy and light chains of antibodies.

We extracted antibody-antigen complexes that met the following criteria: (1) both VH and VL domains are present in the antibody; (2) the antigen is a single-chain protein consisting of at least 50 amino acids; and (3) there are no unresolved CDR residues, yielding 4,081 files.

To deduplicate complexes, we used MMseqs2 (Steinegger & Söding, 2017) to cluster the complexes by heavy and light chains in antibodies and antigen sequences. We used the *easy-linclust* mode with the *–cov-mode 0* option to cluster sequences; we used the default setting for coverage of aligned cutoff at 80%; we used different *–min-seq-id* cutoffs for antibodies and antigens because the antibody framework regions are more conserved than the CDR regions. We cluster heavy and light chains at *–min-seq-id* cutoff of 100% and 70%, respectively. We retained only one representative file for each unique set of identifiers, leading to a refined dataset of 1,725 files.

Two additional files were manually removed. File *6jmr_1P* was removed because its CDR residues are masked with 'UNK' labels and the residue identities are unknown; file *7sgm_0P* was removed because of a non-canonical residue 'DV7' in its CDR-L3 loop.

The final dataset consists of 1,723 antibody-antigen complexes.

A.3. Pipeline to build graph dataset from AbDb

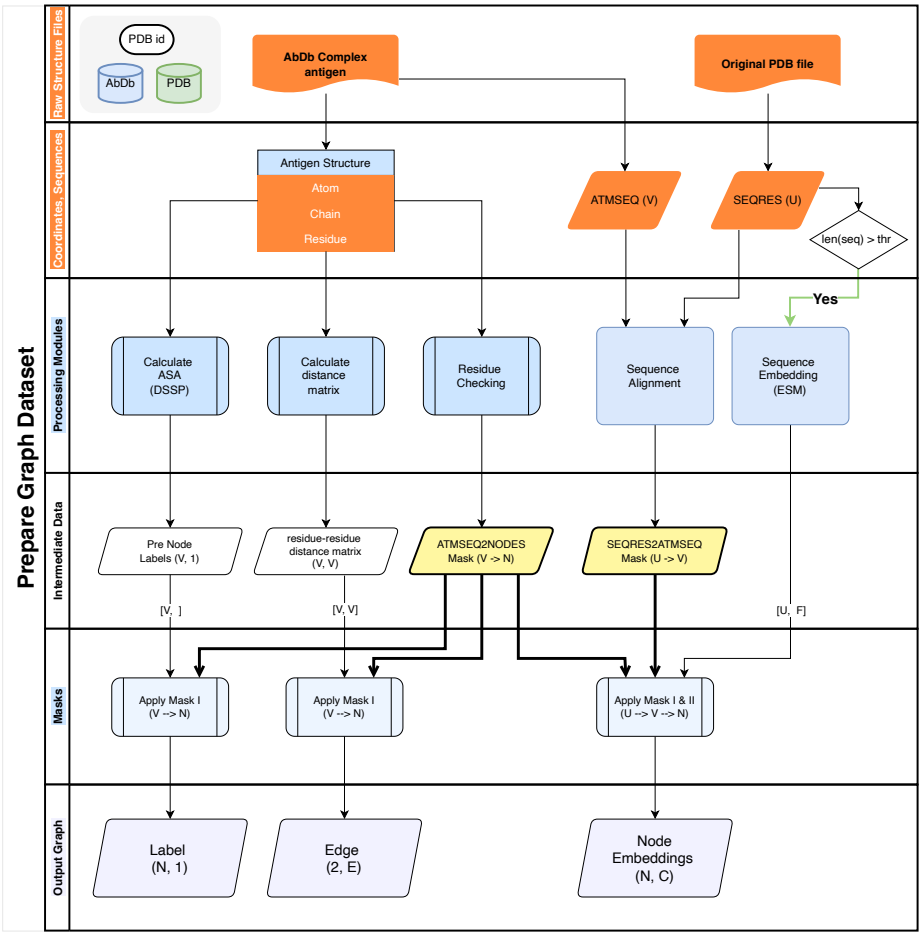


Figure S1. Pipeline to convert an antibody-antigen complex structure into a graph representation.

- **Row 1**: given an AbAg complex PDB ID, retrieve 'AbAg complex' from AbDb and 'raw structure' file from PDB as input, 'AbDb complex antigen' and 'Original PDB file' in the top lane.
- **Row 2**: They are then parsed as hierarchical coordinates (Antigen Structure), and extract ATMSEQ and SEQRES sequences.
- **Row 3**: these are then passed to a set of in-house modules for calculating solvent access surface area (ASA), distance matrix, and filtering out problematic residues, which generates an ATMSEQ2NODES mask. The sequence alignment module aligns ATMSEQ with SEQRES sequences to generate a mask mapping from SEQRES to ATMSEQ. The Sequence Embedding module passes SEQERS through the ESM module to generate embeddings. ESM requires input sequence length and therefore filters out sequences longer than 1021 amino acids.
- **Row 4**: holds intermediate data that we apply masks to generate graph data in **Row 5**.
- **Row 6**: Apply the masks to map SEQRES node embeddings to nodes in the graphs and calculate the edges between the graph nodes.

U, V and N denote the number of residues in the *SEQRES* sequence, *ATMSEQ* sequence and the graph, respectively. *thr* (at 50 residues) is the cutoff for antigen *SEQRES* length. We only include antigen sequences with lengths of at least 50 residues. *SEQRES* and *ATMSEQ* are two different sequence representations of a protein structure. *SEQRES* is the sequence of residues in the protein chain as defined in the header section of a PDB file, and it is the complete sequence of the protein

chain. *ATMSEQ* is the sequence of residues in the protein chain as defined in the ATOM section of a PDB file. In other words, it is read from the structure, and any residues in a PDB file are not resolved due to experimental issues that will be missing in the *ATMSEQ* sequence. Since we are building graph representations using structures, we used *ATMSEQ*. However, the input to the language models require a complete sequence, therefore we used *SEQRES* to generate node embeddings, and mapped the node embeddings to the graph nodes. We performed two pairwise sequence alignments to map such embeddings to graph vertices for a protein chain with Clustal Omega (Sievers et al., 2011). We first align the SEQRES sequence with the atom sequence (residues collected from the ATOM records in a PDB file) and assign residue embeddings to matched residues. Because we excluded buried residues from the graph, we aligned the sequence formed by the filtered graph vertices with the atom sequence to assign residue embeddings to vertices. *ASA* is the solvent-accessible surface area of a residue. If a residue has an ASA value of zero, it is considered buried and will be removed from the graph.

A.4. Implementation details

Exploratory Data Analysis We performed exploratory data analysis on the training dataset to understand the distribution of the number of residue-residue contacts in the antibody-antigen interface. We found that the number of contacts is approximately normally distributed with a mean of 43.42 and a standard deviation of 11.22 (Figure S2). We used this information to set the regularizer in the loss function to penalize the model for predicting too many or too few positive edges.

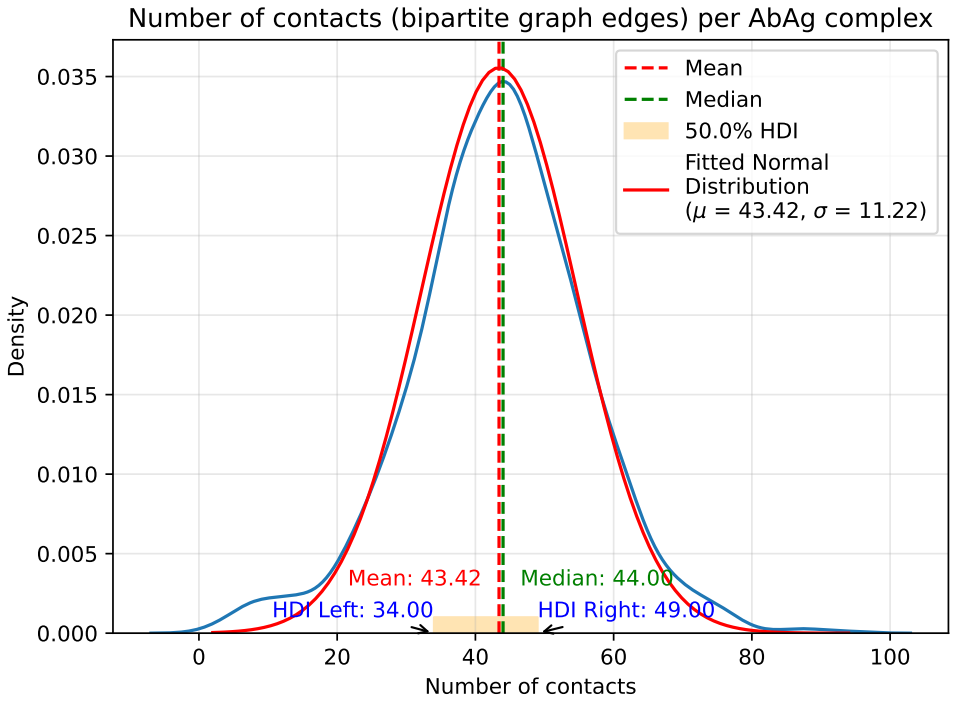


Figure S2. Blue line: distribution of the number of residue-residue contacts in antibody-antigen interface across the dataset with a mean and median of 43.27 and 43.00, respectively. Red line: fitted normal distribution with mean and standard deviation of 43.27 and 10.80, respectively.

Loss function Our loss function is a weighted sum of two parts: a binary cross-entropy loss for the bipartite graph linkage reconstruction and a regularizer for the number of positive edges in the reconstructed bipartite graph.

$$\text{Loss} = \mathcal{L}_r + \lambda \left| \sum^{N} \hat{e} - c \right| \tag{3}$$

$$\mathcal{L}_r = -\frac{1}{N} \sum_{i=1}^{N} \left(w_{\text{pos}} \cdot y_e \cdot \log(\hat{y_e}) + w_{\text{neg}} \cdot (1 - y_e) \cdot \log(1 - \hat{y_e}) \right) \tag{4}$$

The binary cross-entropy loss $\mathcal{L}_r$ is weighted by w_{pos} and w_{neg} for positive and negative edges, respectively. During

hyperparameter tuning, we kept w_{neg} fixed at 1.0 and tuned w_{pos}. N is the total number of edges in the bipartite graph, y_e denotes the true label of edge e, and $\hat{y}_e$ denotes the predicted probability of edge e. The regularizer $\left|\sum^N \hat{e} - c\right|$ is the L1 norm of the difference between the sum of the predicted probabilities of all edges of the reconstructed bipartite graph and the mean positive edges in the training set, i.e., c set to 43. This aims to prevent an overly high false positive rate, given the fact that the number of positive edges is far less than positive edges. The regularizer weight λ is tuned during hyperparameter tuning.

Hyperparameters We carried out hyperparameter search within a predefined space that included:

- **Weights for positive edges** in bipartite graph reconstruction loss, sampled uniformly between 50 and 150.
- **Weights for the sum of bipartite graph positive links**, where values were drawn from a log-uniform distribution spanning $1e-7$ to $1e-4$.
- **Edge cutoff (x)**, defining an epitope node as any antigen node with more than x edges, with x sampled following a normal distribution with a mean of 3 and a standard deviation of 1.
- **Number of graph convolutional layers** in the encoder, we tested using 2 and 3 layers.
- **Decoder type** was varied between two configurations:
  - A fully connected layer, equipped with a bias term and a dropout rate of 0.1.
  - An inner product decoder.

A.5. Antibody-antigen complex examples

To compare the epitopes of antibodies in AsEP, we first clustered these complexes by antigen sequences to group together antibodies targeting the same antigen via MMseqs2 (Steinegger & Söding, 2017). Specifically, we ran MMseqs2 on antigen SEQRES sequences and using the following setup:

- `easy-linclust` mode
- `cov-mode` set to `0` with the default coverage of 80%: this means a sequence is considered a cluster member if it aligns at least 80% of its length with the cluster representative;
- `min-seq-id` set to `0.7`: this means a sequence is considered a cluster member if it shares at least 70% sequence identity with the cluster representative.

We encourage the reader to refer to the MMseqs2 documentation `https://github.com/soedinglab/mmseqs2/wiki` for more details on the parameters used.

We then identify epitopes using a distance cut-off of 4.5 Å. An antigen residue is identified as epitope if any of its heavy atoms are located within 4.5 Å of any heavy atoms from the antibody.

To compare epitopes of antibodies sharing the same antigen cluster, we aligned the antigen SEQRES sequences using Clustal Omega (Sievers et al., 2011) (download from: `http://www.clustal.org/omega/`) to obtain a Multiple Sequence Alignment (MSA). Epitopes are mapped to and denoted as the MSA column indices. The epitope similarity between a pair of epitopes is then calculated as the fraction of identical columns. Two epitopes are identified as identical if they share over `0.7` of identical columns.

Table S3. Antibody-Antigen Complex Examples

(a) Antigen Group Information

abdbid	repr	size	epitope_group
7eam_1P	7sn2_0P	183	0
5kvf_0P	5kvd_0P	9	1
5kvg_0P	5kvd_0P	9	2

(b) CDR Sequences (Heavy Chain)

abdbid	H1	H2	H3
7eam_1P	GFNIKDTYIH	RIDPGDGDTE	FYDYVDYGMDY
5kvf_0P	GYTFTSSWMH	MIHPNSGSTN	YYYDYDGMDY
5kvg_0P	GYTFTSYGIS	VIYPRSGNTY	ENYGSVY

(c) CDR Sequences (Light Chain)

abdbid	L1	L2	L3
7zf4_1P	RASGNIHNYLA	NAKTLAD	QHFWSTPPWT
7zbu_0P	KSSQSLLYSSNQKNYLA	WASTRES	QQYYTYPYT
7xxl_0P	KASQNVGTAVA	SASNRYT	QQFSSYPYT

(d) Structure Titles

abdbid	resolution	title	repr_title
7zf4_1P	1.4	immune complex of SARS-CoV-2 RBD and cross-neutralizing antibody 7D6	SARS-CoV-2 Omicron variant spike protein in complex with Fab XGv265
7zbu_0P	1.4	Zika specific antibody, ZV-64, bound to ZIKA envelope DIII	Cryo-EM structure of zika virus complexed with Fab C10 at pH 8.0
7xxl_0P	1.4	Zika specific antibody, ZV-67, bound to ZIKA envelope DIII	Cryo-EM structure of zika virus complexed with Fab C10 at pH 8.0

Here we provide three example antibody-antigen complexes from the same antigen group, meaning the antigen sequences from each member complex share sequence identity of the aligned region at least 70%. Due to space limitation, we have broken the rows into four parts: Antigen group information, CDR sequences, and structure titles.

- **abdbid**: AbDb ID of the group member;
- **repr**: AbDb ID of the antigen representative;
- **size**: the number of complexes in the group;
- **epitope_group**: A categorical identifier of the epitope group the antibody-antigen complex belongs to;
- **H1**, **H2**, **H3**, **L1**, **L2**, **L3**: CDR sequences of the heavy and light chains;
- **resolution**: Structure resolution;
- **title**: Structure title of the member;
- **repr_title**: Structure title of the antigen representative.

A.6. Metrics definition

$$\text{MCC} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

A.7. Fine-tuning ESMBind on AsEP

The performance reported in the main text for ESMBind is derived by fine-tuning ESM2 on general protein binding sites. We performed a further fine-tuning experiment, fine-tuning it on the presented AsEP dataset and evaluating it on the AsEP test set to enable a more direct comparison of ESMBind to WALLE. Fine-tuning of ESMBind on AsEP was done using the Low-Rank Adaptation method (Hu et al., 2021).

Table S4. Performance Metrics

Metric	Value
MCC	**0.103923**
Accuracy	0.504478
AUC-ROC	0.584497
Precision	0.128934
Recall	0.707731
F1	0.213829

B. Appendix: Link prediction baseline

While the majority of existing studies focus on node-level prediction, i.e., predicting which residues are likely to be the epitope residues, we are interested in predicting the interactions between epitope and antigen residues. We argue that, on the one hand, this would provide a more comprehensive understanding of the interaction between epitopes and antigens, and on the other hand, it would be good in terms of model interpretability. Existing methods for predicting epitope residues are mostly based on sequence information, which is not directly interpretable in terms of the interaction between epitopes and antigens.

Our hyperparameter search was conducted within a predefined space as defined in Appendix A.4. We used the Bayesian optimization strategy implemented through Weights & Biases, targeting the maximization of the average bipartite graph link Matthew's Correlation Coefficient (MCC).

The optimization process was managed using the early termination functionality provided by the Weights & Biases' Hyperband method (Falkner et al., 2018), with a range of minimum to maximum iterations set from 3 to 27.

The best set of hyperparameters is 2 *GCNConv* layers, a batch size of 32, a weight for positive edges of 54.7, a weight for the sum of positive links at approximately $5.75e-7$, and an edge cutoff of 2.38. The resulting MCC evaluated on the test set was 0.072 (standard error: 0.009) for the bipartite graph link prediction task.

Table S5. Evaluation of WALLE on the bipartite graph link prediction task

Metric	Mean (Standard Error)
MCC	0.072 (0.009)
ROC-AUC	0.582 (0.011)
Precision	0.049 (0.008)
Recall	0.167 (0.023)
F1	0.053 (0.007)

C. Appendix: Ablation Studies

To investigate the impact of different components on WALLE's performance, we carried out ablation studies and described them in this section. For each model variant, we performed hyperparameter tuning and reported the evaluation performance using the model with the best performance on the validation set.

C.1. Ablation study: replace graph component with linear layers

To investigate whether the graph component within the WALLE framework is essential for its predictive performance, we conducted an ablation study in which the graph component was replaced with two linear layers. We refer to the model as 'WALLE-L'. The first linear layer was followed by a ReLu activation function. Logits output by the second linear layer were used as input to the decoder. The rest of the model architecture remained the same.

It differs from the original WALLE model in that the input to the first linear layer is simply the concatenation of the embeddings of the antibody and antigen nodes, and the linear layers do not consider the graph structure, i.e., the spatial arrangement of either antibody or antigen residues. The model was trained using the same hyperparameters as the original WALLE model. The performance of WALLE-L was evaluated on the test set using the same metrics as the original WALLE model.

C.2. Ablation study: WALLE with simple node encoding

The presented WALLE model utilizes embeddings from large language models, including ESM2 (Lin et al., 2022) or IgFold(Ruffolo et al., 2023) for representing amino acid sequences, as these models are able to capture the sequential and structural information inherent in protein sequences, providing a rich, context-aware representation of amino acids. To test the effectiveness of such embeddings in this downstream task, we conducted an ablation study where we replaced the embeddings from language models with simple node encodings. Specifically, we evaluated the performance of WALLE when using 'one-hot' encoding and 'BLOSUM62' encoding for amino acids in both antibody and antigen sequences.

C.3. Ablation study: WALLE with ESM2 embeddings for both antibodies and antigens

We also investigated whether the choice of language models can impact the predictive performance of WALLE; we conducted an ablation study to evaluate the performance of WALLE when both antibodies and antigens are represented using embeddings from the ESM2 language model (Lin et al., 2022) while the original model uses AntiBERTy (Ruffolo et al., 2023) for antibodies as it is trained exclusively on antibody sequences. This also tests whether a language model trained on general protein sequences can be used for a downstream task like antibody-antigen interaction prediction.

One-hot encoding

One-hot encoding is a method where each residue is represented as a binary vector. Each position in the vector corresponds to a possible residue type, and the position corresponding to the residue present is marked with a 1, while all other positions are set to 0. This encoding scheme is straightforward and does not incorporate any information about the physical or chemical properties of the residues. This method tests the model's capability to leverage structural and relational information from the graph component without any assumptions introduced by more complex encoding schemes.

BLOSUM62 encoding

BLOSUM62 (Henikoff & Henikoff, 1992) encoding involves using the BLOSUM62 matrix, which is a substitution matrix used for sequence alignment of proteins. In this encoding, each residue is represented by its corresponding row in the BLOSUM62 matrix. This method provides a more nuanced representation of residues, reflecting evolutionary relationships and substitution frequencies.

C.4. Hyperparameter tuning

We used the same hyperparameter search space defined in Appendix A.4 and performed a hyperparameter search as defined in Appendix B for each model variant in the ablation studies. We report the evaluation performance of the tuned model for each variant in Table S6.

We observed that WALLE's performance with simple node encodings ('one-hot' and 'BLOSUM62') is considerably lower than when using advanced embeddings from language models. This indicates that the embeddings derived from language

Table S6. Performance of WALLE without graph component and simple node encodings on test set from dataset split by epitope to antigen surface ratio.

Algorithm	Encoding	MCC	AUCROC	Precision	Recall	F1
WALLE	Both	**0.2097** (0.0195)	**0.6351** (0.0126)	**0.2346** (0.0183)	0.4217 (0.0279)	**0.2580** (0.0178)
WALLE-L	Both	0.1593 (0.0155)	0.6124 (0.0109)	0.1750 (0.0109)	0.4696 (0.0243)	0.2371 (0.0137)
WALLE	ESM2	0.1955 (0.0212)	0.6219 (0.0137)	0.2280 (0.0188)	0.4103 (0.0291)	0.2553 (0.0188)
WALLE-L	ESM2	0.1445 (0.0138)	0.6100 (0.0097)	0.1598 (0.0102)	0.5355 (0.0216)	0.2266 (0.0125)
WALLE	One-hot	0.0968 (0.0094)	0.5830 (0.0076)	0.1185 (0.0052)	**0.8923** (0.0118)	0.2026 (0.0081)
WALLE	BLOSUM62	0.0848 (0.010)	0.5739 (0.0081)	0.1182 (0.0055)	0.8401 (0.0151)	0.1993 (0.0083)

The values in parentheses represent the standard error of the mean; 'WALLE-L' refers to WALLE with the graph component replaced by two linear layers. 'ESM2' refers to the embeddings from the ESM2 language model `esm2_t12_35M_UR50D`. 'One-Hot' refers to one-hot encoding of amino acids. 'BLOSUM62' refers to the BLOSUM62 encoding of amino acids. 'Both' refers to embedding antibodies and antigens using the `esm2_t12_35M_UR50D` ESM2 model and AntiBERTy (via IgFold) language model, respectively. The best performing model is highlighted in bold.

models capture more nuanced information about the amino acids, enabling the model to better predict epitope-antigen interactions.

The degenerated performance of WALLE with simple encodings can be attributed to the lack of contextual information and structural features in these representations. The high recall but low precision values suggest that the model is unable to distinguish between true and false interactions, leading to a high number of false positives. This highlights the importance of using meaningful embeddings that capture the rich structural and sequential information present in protein sequences.

When comparing WALLE with WALLE-L (without the graph components), we observe that the model's performance drops considerably when the graph component is replaced with fully connected linear layers. This indicates that the topological information captured by the graph component also contributes to the model's predictive performance.

We also observed that WALLE with ESM2 embeddings for both antibodies and antigens achieved similar performance to WALLE with AntiBERTy and ESM2 embeddings for antibodies and antigens, respectively. This suggests that the ESM2 embeddings somehow provide effective information for both antibodies and antigens without training exclusively on antibody sequences.