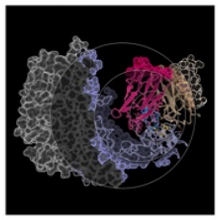
mAbs



ISSN: 1942-0862 (Print) 1942-0870 (Online) Journal homepage: www.tandfonline.com/journals/kmab20

Epi4Ab: a data-driven prediction model of conformational epitopes for specific antibody VH/VL families and CDRs sequences

Nhan Dinh Tran, Krithika Subramani & Chinh Tran-To Su

To cite this article: Nhan Dinh Tran, Krithika Subramani & Chinh Tran-To Su (2025) Epi4Ab: a data-driven prediction model of conformational epitopes for specific antibody VH/VL families and CDRs sequences, mAbs, 17:1, 2531227, DOI: 10.1080/19420862.2025.2531227

To link to this article: https://doi.org/10.1080/19420862.2025.2531227

MABS
2025, VOL. 17, NO. 1, 2531227
https://doi.org/10.1080/19420862.2025.2531227

Taylor & Francis
Taylor & Francis Group

REPORT

Epi4Ab: a data-driven prediction model of conformational epitopes for specific antibody VH/VL families and CDRs sequences

Nhan Dinh Tran, Krithika Subramani, and Chinh Tran-To Su

Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Matrix, Singapore

ABSTRACT

Antibodies recognize antigens via complementary and structurally dependent mechanisms. Therefore, inclusion of antibody inputs is crucial for accurate epitope prediction. Given the limited availability of antibody–antigen complex structures, any epitope prediction model will require minimal yet sufficient antibody inputs to ensure precise epitope identification. To address this need, we introduce Epi4Ab, an antibody-specific epitope prediction model that focuses on identifying unique in-contact antigen residues for a given antibody. Epi4Ab requires minimal antibody inputs, specifically VH/VL families and complementarity-determining region sequences.

ARTICLE HISTORY
Received 14 March 2025
Revised 3 July 2025
Accepted 3 July 2025

KEYWORDS
Antibody-specific epitope prediction; CDR sequences; conformational epitope; graph-based model; machine learning; VH-VL families

Introduction

Therapeutic antibodies are crucial in disease detection and treatment, especially for cancer and infectious diseases.[1] While effective, monoclonal antibodies used in targeted therapies are a key factor contributing to the rising cost burden over the years.[2] Antibody discovery technologies against a target have recently been expanding via intensive experimental and/or computational-aided efforts.[3–5] The processes are, however, laborious, resource- and time-consuming. Moreover, the designed antibodies must effectively target their antigens while also ensuring safety to succeed as therapeutics.[6] However, resistance to existing drugs is not uncommon.[7] The requirements of effective therapeutics and the expense associated with developing new drugs from scratch can be balanced via drug repurposing, which enables approved therapeutics to be applied to mutated or newly emerging drug targets. This goal gives rise to our "reverse" approach in which we first identify an antibody-binding region (epitope) on a new antigen given a specific antibody, initiating a pipeline of antibody engineering for the therapeutics repurposing.

An epitope is a region on an antigen defined by a linear polypeptide chain or constituted of discontinuous residual segments forming conformational patches that are complementarily bound by an antibody. Epitope prediction is a computational approach to identify the antigen regions, either linear or conformational, that could be recognized by antibodies. In contrast to generic epitope prediction, which predicts epitopes targetable by all possible antibodies and requires only antigen as input, antibody-specific epitope prediction is more targeted and necessitates both antigen and antibody as starting points.

The generic epitope prediction approach is elicited by sequence-based and structure-based methods. The former exploits advantages of extensive protein sequence data available, using deep learning-based techniques facilitated with protein language models to identify sequence patterns and predict epitopes on new antigen sequences, as demonstrated in BepiPred 3.0,[8] SEMA 2.0,[9] EpiDope,[10] and EpiBERTope.[11] On the other hand, structure-based methods rely on physicochemical characteristics inherent in the antigen–antibody interactions embedded in both the bound conformations of the partners. The structural features are encoded and decoded in supervised manners to make prediction on new antigen structures, as exemplified in ElliPro,[12] epitope3D,[13] SEMA 2.0,[9] SEPPA 3.0,[14] and DiscoTope 2.0.[15] In either scenario, accurately predicting epitopes remain challenging[16] due to ambiguity of true epitopes, as well as limited availability of the antibody–antigen (Ab–ag) complex structures.

CONTACT Chinh tran-To Su chinhsutranto@bii.a-star.edu.sg Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, Matrix 138671, Singapore

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2025.2531227

Only a limited number of antibody-specific epitope prediction tools have been developed, requiring both antigen and antibody as inputs, such as EpiPred,[17] EpiScan,[18] and SEPPA-mAb.[19] These methods identify surface patches that could be antigenic and targeted by the given antibody. Within their own benchmarking, each method demonstrated reasonable performance and ability to distinguish epitope and non-epitope residues. Nonetheless, false positive rate remains high, leaving room for more development.

An antibody binds with high affinity to its antigen via networks of weak and non-covalent interactions. The Ab–ag interacting interface is formed by structural complementarity of the two partners' binding regions, and this recognition is specific to the antigen.[20,21] The antigen recognition specificity is determined by structural combinations of antibody elements such as complementarity-determining regions (CDRs) on both heavy and light chain variable domains (VH and VL; see Supplementary Figure S1). Previous studies[22–28] have shown manipulated combinations of VH-VL framework regions (FWRs) and CDRs could allosterically modulate distinguishingly the antigen bindings, suggesting that these specific antibody structural characteristics play a substantial role in the epitope identification on the antigen.

Like EpiScan, which leverages such heavy/light chain FWRs and CDRs information via the inputted antibody sequences for epitope prediction, we incorporate yet minimize the requirements of antibody details and concentrate on residual interaction networks embedded in the bound antigen, aiming for (1) less dependency on prior knowledge of the given antibody (which is occasionally limited) and (2) identification of potential unique in-contact residues on the targeted antigen with the specific antibody.

To do so, we introduce Epi4Ab, a data-driven graph-based model designed to predict conformational epitopes for specific antibody VH-VL and CDRs. Epi4Ab primarily focuses on identifying antigen residues that directly interact with the given antibody, requiring only minimal antibody inputs, specifically VH/VL families and CDRs sequences, for more efficient epitope prediction.

Methods

Epi4Ab model

The Epi4Ab model learns structural interaction patterns between antigen and antibody to predict potential antibody-interacting residues (epitope) on a new antigen, using minimal input of a certain specific antibody, such as heavy-light chain variable VH-VL families and CDRs sequences. This model is based on a hybrid Graph Neural Network and Residual Network[29] (GNNResNet) integrated with an *attention* mechanism to specifically capture distinct residual structural features embedded in the bound conformation of the antigen (Figure 1).

The Epi4Ab model leverages on two pretrained language models, ESM2[30] (t30_150M) and AntiBERTy,[31] to extract sequence patterns of the antigens and their antibody partners, respectively. For each Ab–ag pair, a full antigen sequence and six antibody CDRs (H1, H2, H3, L1, L2, and L3) were used. Subsequently, multi-head attention[32] (mHA) was used to estimate associations of these resulting pairs of tensors.

Simultaneously, structural features were derived using the antibody-bound conformations of the antigens. These bound antigen structures were then relaxed by reconstructing the side-chains to (1) partially mimic unbound conformation of the antigen and (2) augment data for subsequent model training. Antibody VH-VL families and CDR lengths were concurrently recorded. All the features were used for the graph construction. To enhance dynamics of residual interactions in the constructed graphs, we developed a meta-model to learn and optimize a potential scoring function, accommodating interaction potentials such as bond, charge, and Lennard-Jones potentials.

The residual graph G representing the antigens were fed to the multilayer GNNResNet facilitated with Graph Attention Network[33] (GAT) architecture to classify each antigen residue as background *non-epitope* (label 0), antibody-interacting residue, or *epitope* (label 1), and *potential epitope* (label 2). To investigate effectiveness of the attention mechanism, we also implemented a naïve Graph Convolutional Network (GCN) without the GAT for comparison. Details are shown in Figure 1.

Training and testing datasets

Ab–ag complexes were retrieved from the Structural Antibody Database[34] (SAbDab), accessed in November 2022. Only the high resolution (≤3Å) Ab–ag complexes containing both antibody heavy and

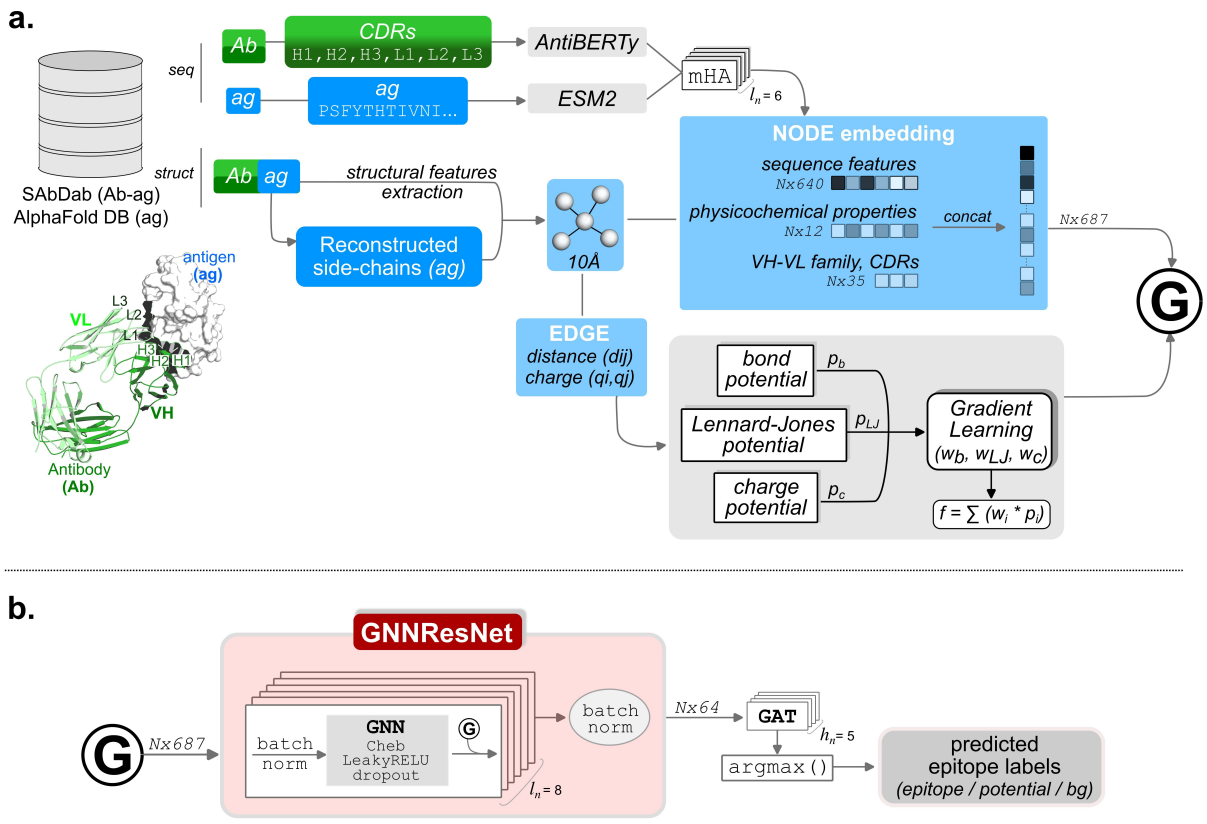


Figure 1. Epi4Ab workflow. (a). The model's highlights include (i) leveraging on two pretrained language models ESM2 and AntiBERTy and (ii) integrating a meta-model of learning molecular interaction potentials (bond, Lennard Jones, and charge) into the residual graph G of the antigen. (b) Demonstrates the Epi4Ab's core consisting of the residual graph Neural Network (GNNResNet) implemented with graph attention Network (GAT) architecture. Different numbers of layers (l_n) and GAT attention heads (h_n) were optimized, using tensors of "number of residues N × feature size" dimensions.

light chain variable domains (such as paired Fab or Fv) bound with a single-chain protein antigen (length ≤670 residues, given our limited computational resource) were selected, resulting in 1,379 Ab–ag complexes. To reduce redundancy, antigen sequences from these Ab–ag complexes were extracted and clustered based on their similarity using CD-HIT[35] with 70% similarity threshold. For each cluster, we further excluded members with >80% similarity to the cluster representative. All singletons (1-member clusters) were included in the training set. For the other clusters, we randomly selected 10% of the cluster members for the test set while the remaining 90% were included in the training set. As a result, we obtained 13 complexes for the test set and 303 Ab–ag complexes for the training set. Additionally, seven HER2 antigens were added into the final test set to further test the ability of predicting multiple epitopes on the same antigen.

All the 323 Ab–ag complexes were retrieved from Protein Data Bank (PDB) as mmCIF format, from which we searched for any information of missing residues, disordered atoms, and multiple residual configurations. When multiple configurations exist, we selected the first one (e.g., A) and excluded the others. Biopython v.1.83[36] were used to verify these gap/missing residues. MODELLER[37] was then used to add the missing residues, followed by minimization ("RepairPDB") using FoldX.[38] The pdb2pqr[39] was used to further address any remaining structural errors. Only the structure models that passed all these checkpoints were included in the final datasets. To note, we excluded "fully" disordered structures (by visualization) from the finetuned datasets.

Subsequently, a process of reconstructing sidechain, using FoldX,[38] was performed for each bound conformation of the antigens to partially mimic their unbound conformations. To further augment the data for model training, we retrieved counterparts of the static structures of several antigens, ~10% (30/303) of the selected data, from the AlphaFold Protein Structure Database[40] with pLDDT ≥70 (Supplementary Table S1). Overall, we finalized 636 (303×2 and 30) Ab–ag complexes for training and 20 complexes for testing.

Characterize structural embedded antibody–antigen interaction features

We characterized the Ab–ag interactions by examining physicochemical properties of their binding interfaces, particularly reflected on each of the antigen residues. For each Ab–ag complex, the bound conformation of the antigen was used to presume the structural dependency of the residual interaction network within the antigen structure (or allostery involved, if any) when bound by the corresponding antibody. The structural features included residue depth estimated using BioPython v.1.83, partial charges calculated using PDB2PQR,[39] and dihedrals including phi, psi, omega, and chi angles of each residue using MDAnalysis.[41] To further emphasize diverse properties of different amino acids, we estimated amino acid composition (*aac*) and attribute composition (*atc*) for each aa type: $aac_i = \frac{n_i}{\sum_i^{20} n_i}$ and $atc_i = \frac{n_i}{N_k}$ with n_i representing total number of aa of type i and N_k representing total number of aa in an attribute group k (k = *positive, negative, polar, hydrophobic*).

Antibody heavy and light variable domain (VH-VL) families and CDR features

The VH-VL families were recorded from the retrieved SAbDab dataset, including existing families such as heavy: [IGHV01, IGHV02, IGHV03, IGHV04, IGHV05, IGHV06, IGHV07, IGHV08, IGHV09, IGHV14], and light: [IGKV01, IGKV02, IGKV03, IGKV04, IGKV05, IGKV06, IGKV08, IGKV10, IGKV12, IGKV13, IGKV14, IGLV01, IGLV02, IGLV03, IGLV06]. For those families with few entries or not known, "others" and "unknown" were recorded, respectively.

Similarly, sequences of all six CDRs were extracted. For any missing CDR information, a manual CDR search was performed on the SAbDab webapp using corresponding PDB entry following Chothia CDR definition. The CDR lengths were calculated accordingly.

VH-VL-H3$_{len}$-L1$_{len}$ clustering

The Ab–ag complexes were first clustered accordingly to the antibody paired VH-VL families and CDR H3/L1 lengths. We selected H3/L1 as representative metrics because the H3 and L1 lengths are longest among heavy and light CDRs, respectively, in the retrieved dataset. Sequence global alignment using MAFFT[42] with 1000 iterates was performed for each H3 and L1 sequence within each cluster. An in-house python script was used to estimate the alignment score using BLOSUM62 and the defaulted MAFFT *gap_score* = −1.53. The H3 and L1 sequences with the highest score were extracted as templates, representing the corresponding VH-VL-H3$_{\text{len}}$-L1$_{\text{len}}$ cluster. The templates were used to estimate the *H3score* and *L1score* features for each of the CDR H3 and L1 sequences in the test set.

Ground truth residue labels for supervised learning

We classified the antigen residues such as background non-epitope (label "0"), direct antibody-interacting residues as *epitope* (label "1") and residues potentially targeted by other antibodies or non-interacting interfacial residues as *potential epitopes* (label "2").

The label "1" residues were defined as the Ab–ag interfacial direct contact residues within a cutoff distance of 5 Å and determined by CIPS.[43] The remaining residues were labeled as "0" (background). To lessen the resulting imbalance in the dataset labels, we introduced the label "2" as other potential epitope residues.

The generic epitope prediction tools Ellipro[12] (structure-based) and BepiPred 3.0[8] (sequence-based) were used to first identify potential antigenic surface residues. The two methods were selected due to their largest coverage among several available methods[16] to filter out background non-epitope residues. Consensus of the resulting antigenic residues (excluding those initiated with label "1") by the two methods were then defined as label "2" residues.

Residual graph construction

Each antigen was represented by a residual graph $G = \{V, E\}$ with nodes V representing the antigen residues and edges E representing interactions between the residues. A cut-off Cβ-Cβ (except for Gly using Cα) distance of 10 Å was used to determine neighboring nodes for each node v_i. Each node v_i was represented by a concatenated tensor including the characterized sequence and structural features together with the respective antibody records of the VH-VL families and CDRs lengths (Figure 1).

To enhance dynamics of the graph, we integrated several potentials estimating the edge attributes and hence mimicking the biophysical interactions within the networks. These include bond (p_b), Lennard-Jones (p_{LJ}), and charge (p_c) potentials incorporated in a scoring function as below:

$$f = \sum (w_i * p_i) = w_b * p_b + w_{LJ} * p_{LJ} + w_c * p_c$$

In which w_b, w_{LJ}, and w_c are the respective weighted coefficients and were optimized via an implemented gradient learning meta-model (Figure 1). This process was supervised and performed using 5-fold cross validations on the full training sets. The set of weights that resulted in the highest $F1_{epitope}$ score in the validation sets and least over-fitting was selected for further training.

The three potentials were defined as below:

(1) Bond mimicking potential: $p_b = \frac{1}{d_{ij}}$
(2) Lennard-Jones mimicking potential: $p_{LJ} = \begin{cases} V_{d_{ij}} - 2 \times V_{min}, if\ d_{ij} < d_{eq} \\ -V_{d_{ij}}, otherwise \end{cases}$

With $V_{d_{ij}} = \left[\left(\frac{\sigma}{d_{ij}}\right)^{12} - \left(\frac{\sigma}{d_{ij}}\right)^{6}\right]$ representing potential energy at distance d_{ij} between two residues *i,j* (Cα-Cα) and $\sigma = 3.4$ Å is the distance at which $V_{d_{ij}} = 0$.

$V_{min} \approx -0.25$, representing potential energy at the equilibrium $d_{eq} = 2^{\frac{1}{6}}\sigma \approx 3.8$ Å.

(3) Charge potential: $p_c = \frac{|q_i \times q_j|}{d_{ij}}$, in which $q_i, q_j$are partial charges of residues i, j.

The GNNResNet architecture

The hybrid Graph Neural Network and Residual Network (GNNResNet) was implemented with eight layers, each of which used a Chebyshev spectral graph convolutional operator[44] (Cheb) and activated using LeakyRELU. Various feature dropout ratios (0.2, 0.2, 0, 0.3, 0.4, 0.3 0, 0) and numbers of hidden channels were accordingly used for each of the layers, i.e., 128→64→96→128→96→96→32→64. Other parameters: *filter_size* = 5, *batch_size* = 44, and "BatchNorm" were used. The model was trained for 2000 epochs with *learning_rate* = 0.0001, using Adam optimizer and with 5 attention heads (GAT). These hyperparameters were optimized using Optuna[45] using a predefined objective function to maximize the "$F1_{epitope}$ score" in the validation set during 100 trials. The evaluation process was performed on 10-fold validation (with data splits of 90% in "train set" and 10% in "validate set"), from which 95% prediction interval was estimated on the test set.

Probability (*prob*) was estimated for each label (0, 1, 2), the largest of which determined the predicted label outcome for each antigen residue. The predicted probability was then converted into prediction score as following: $pred\_score = ln\left(\frac{prob}{1-prob}\right)$.

Benchmarking against several epitope prediction methods

The test set of 20 antigens was used for epitope prediction on several webservers with default settings, such as epitope3D,[13] SEMA 2.0,[9] DiscoTope 3.0,[46] SEPPA 3.0,[14] and SEPPA-mAb.[19]

To maintain competitiveness across all the methods, we reevaluated true positives using a predefined cavity sphere (green circle in Figure 2) that encompasses the Ab–ag interface in the ground truth. Therefore, a predicted epitope residue (by any of the tools) was considered as a true positive only if it was located

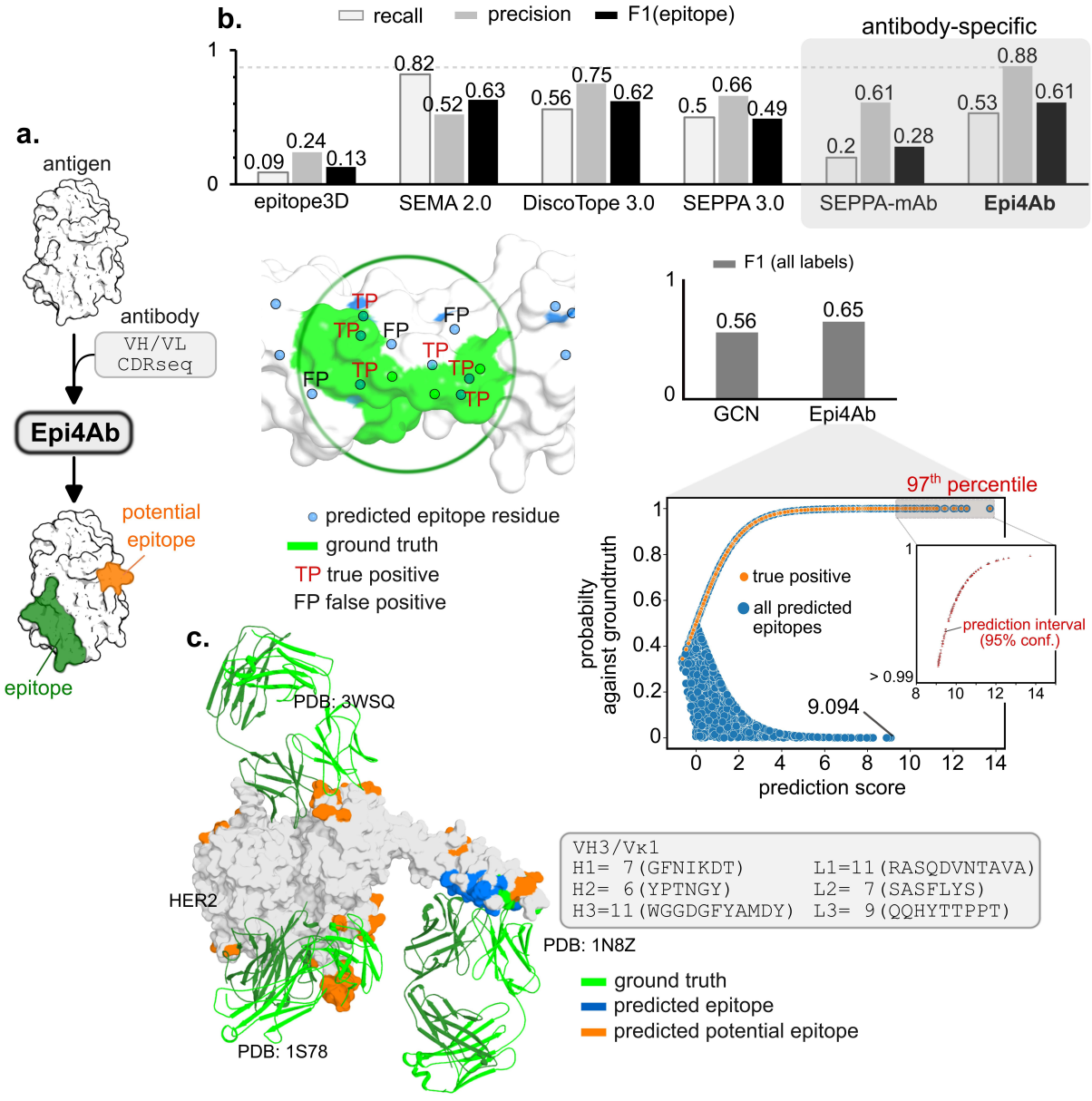


Figure 2. Performance of Epi4Ab. (a). Overview of the Epi4Ab model. (b) The "attention" mechanism enhanced the epitope identification, with highest prediction scores (>9.094) at 97th percentile with 95% confidence. Comparison of Epi4Ab model performance across various generic and antibody-specific epitope prediction methods. For benchmarking consistency, true/false positives were re-evaluated (see methods), by implementing a (green) cavity sphere encompassing the true antibody–antigen interaction interface. (c). Structural presentations of the Epi4Ab predictions using VH-VL and CDRs sequences extracted from an antibody targeting HER2 antigen (PDB:1N8Z) in the test set highlight the ability to predict potential epitopes recognizing by other antibodies, e.g., in 1S78 or 3WSQ. To simplify, HER2 is shown in gray surface and the antibodies in dark/light green cartoons.

within this sphere and/or directly interacted with the antibody (i.e., label 1 and 2 in the ground truth). Recall, Precision, and $F1_{epitope}$ score were used as metrics to compare across the methods.

Results

Epi4Ab predicts conformational epitopes given minimal antibody input

To address limitations and occasional unavailability of antibody structures, the Epi4Ab model was trained additionally with minimal antibody input, specifically VH-VL families and CDRs sequences, to identify

antibody-specific epitopes on antigens. We defined "epitope" as direct antibody-interacting residues on the antigen targeted by the specific antibody, distinguishing the predicted *epitopes* from the *potential epitopes* (regions potentially targeted by other antibodies) in our results.

Since interaction networks of epitope residues differed from those of neighboring non-epitope residues, an *attention* mechanism was incorporated into our GNNResNet model to focus on the distinct structural embedded features. Compared to our baseline model (a naïve GCN without the attention), adding the *attention* improved the prediction accuracy and increased true positives (reflected in F1 score) estimated on the 20 antigens in the test set (Figure 2). Repeated testing results (on 10-fold cross validation) affirmed the highest prediction score (threshold >9.094) ranked at 97th percentile, with 95% confidence.

We used this test set to benchmark our Epi4Ab model against various available generic epitope prediction tools (webservers), such as epitope3D,[13] SEMA2.0,[9] DiscoTope3.0,[46] and SEPPA3.0,[14] and a recent antibody-specific epitope prediction method SEPPA-mAb.[19] To maintain competitiveness and avoid biases among the tools, we reevaluated the *true positives* for each tool's predictions, i.e., only the predicted epitope residues that were located within the cavity sphere (green circle in Figure 2b) encompassing the antibody-interacting residues and/or directly interacted with the corresponding antibody. Results across the tools indicated that Epi4Ab performed modestly (ROC_AUC = 0.75 and PR_AUC = 0.66, shown in Table 1) and on par with the other tools, achieving the highest precision (0.88) and with $F1_{epitope}$ (0.61). Among these generic epitope prediction tools, SEMA 2.0 and DiscoTope 3.0 performed comparably well, with the precision of 0.52 and 0.75, $F1_{epitope}$ of 0.63 and 0.62, respectively (Figure 2b).

When comparing with the antibody-specific tool SEPPA-mAb, Epi4Ab could identify the epitope residues better ($F1_{epitope}$ = 0.61 vs 0.28). This suggests an advantage in identifying direct interacting residues over surface patches.

Does the minimal antibody input impact the Epi4Ab epitope prediction or just merely add supplementary features?

Previous studies[25,27,28,47] demonstrated antibody heavy/light chain variable FWRs and CDRs modulated antigen binding through allosteric effects. This elucidated the associated role of the variable (VH/VL) family, as determined by FWRs, and CDRs in identifying the interaction interface on antigens; hence, these antibody features are necessary to predict interfacial residues on the antigen using the Epi4Ab approach.

The 20 antibodies that bound to the tested antigens contain varying VH-VL family combinations such as the common VH1 to VH5 (heavy) paired with Vκ1, Vκ3, Vλ1, Vλ2, or Vλ3 (light) and diverse CDRs, particularly with different lengths of H3, L1, and L3 (Suplementary Table S2). To assess the model sensitivity to the antibody input, we performed the Epi4Ab inference on the tested antigens individually using 472 unique VH-VL-$H3_{len}$-$L1_{len}$ combinations available in the dataset (from which H3 and L1 were chosen since they are the longest H-CDR and L-CDR). To further increase the variances among the simulated inputs, the most differentiated H3 and L1 sequences from the cluster consensus were selected with respect to each of these VH-VL-$H3_{len}$-$L1_{len}$ combinations. The hypothesis stated that if the antibody input was irrelevant or having minute effect on the epitope prediction of the Epi4Ab, the changes in the model's ability to distinguish epitope from non-epitope residues would remain insignificant. We used F1

Table 1. Epi4Ab benchmarking against various generic and antibody-specific epitope prediction tools, estimating on ROC_AUC (area under the ROC curve), PR_AUC (area under precision–recall curve), and accuracy.

Epitope prediction model		ROC_AUC	PR_AUC	Accuracy
Generic	epitope3D	0.45 ± 0.05	0.39 ± 0.14	0.52 ± 0.09
	SEMA 2.0	0.67 ± 0.09	0.50 ± 0.13	0.65 ± 0.10
	DiscoTope 3.0	0.73 ± 0.10	0.60 ± 0.16	0.76 ± 0.07
	SEPPA 3.0	0.60 ± 0.09	0.49 ± 0.13	0.67 ± 0.10
Antibody-specific	SEPPA-mAb	0.59 ± 0.10	0.48 ± 0.16	0.67 ± 0.13
	Epi4Ab	**0.75 ± 0.16**	**0.66 ± 0.20**	**0.79 ± 0.13**

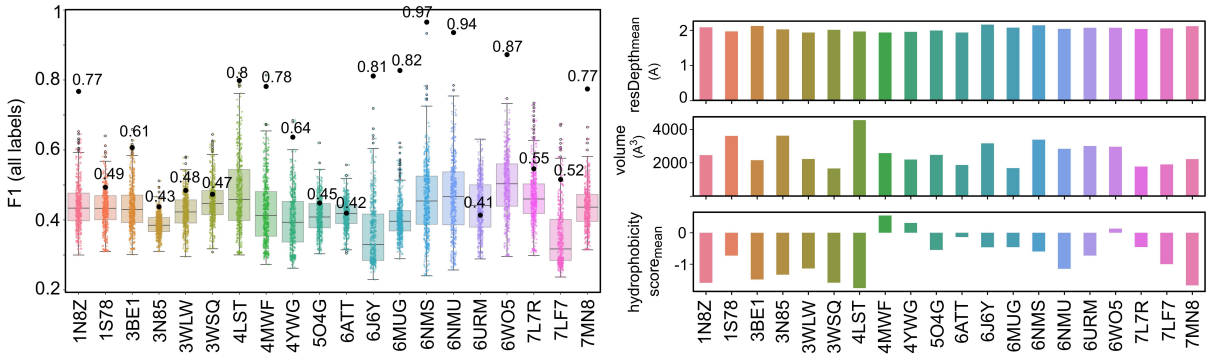


Figure 3. Effect of the VH-VL and CDRs input changes on the epitope prediction of the Epi4Ab model. The epitope inference results of each tested antigen given various VH-VL and CDRs inputs. The starting F1 scores corresponding to the original VH-VL and CDRs combination in each tested antigen are shown in black dots. The antibody-binding cavities in each of the original tested antigens are demonstrated using several main physicochemical features such as residue depth (mean), volume, and hydrophobicity score (the higher, the more hydrophobic).

score (all labels) to reflect this ability, also emphasizing the identification of the epitope residues (label 1 and 2) on each of the tested antigen.

Our results indicated that the calculated F1 scores varied noticeably given the VH-VL and CDRs input changes for each of the tested antigens, with substantial variations observed in all the cases (Figure 3), the majority of which contained large cavities (> 2000 Å^3) at the binding interfaces. Notably, antibodies in these cases contain Vκ3 paired with VH1/VH3/VH4 bearing long CDR-H3 (Supplementary Table S2), such as 4LST, 4MWF, 4YWG, 6NMS, and 6NMU. This suggests the antigen regions recognized by these VH*-Vκ3 pairs might be more susceptible to antibody manipulation. On the other hand, we observed that, in the test cases involving λ-light chain (except for 6J6Y), such as paired VH1-Vλ1, VH2-Vλ2 in 3WSQ and 3WLW, the scores deviated in modest ranges and less dramatically when challenged with various input VH-VL and CDRs combinations. Several test cases involving HER2 antigens exhibited such modest deviations.

Nevertheless, there was no correlation (data not shown) observed in the retrieved dataset between the antibody CDR lengths or sequences and other physicochemical features such as residue depth, volume, or hydrophobicity of the antibody-binding region on the antigens, indicating the independency of these features. The VH-VL and CDRs inputs were thus more than essential descriptors, as they also manipulated the epitope prediction outcomes. Since the necessary antibody inputs are minimal, this highlights the effectiveness of our Epi4Ab model in predicting the antibody-interacting residues on the given antigen.

Discussion

We developed the Epi4Ab model to predict potential antibody-interacting residues on a new antigen given certain specific antibody VH-VL families and CDRs sequences. The reasonable performance of Epi4Ab suggests feasibility for the epitope prediction to leverage minimal dependency on the antibody structural input, which otherwise is limited or occasionally unavailable.

We prioritized identification of unique in-contact residues rather than conventional antigenic surface patches to enhance manipulation of antigen targeting. This would consequently downplay the secondary role of non-interacting residues located at the Ab–ag interface. In exchange, we used bound conformations of the antigens to preserve both short- and long-ranged effects and hence maintain local interaction networks embedded in structural features within the bound conformation of the antigen. This, in fact, facilitated the Epi4Ab prediction of *potential epitope* (orange in Figure 2C), which depicted regions potentially recognized by any other antibodies. For instance, the HER2 antigen is recognized at different binding interfaces by multiple antibodies. In some cases, Epi4Ab accurately identified the corresponding interfacial patch with precise interacting residues, and, in fact, effectively distinguished the non-epitope residues. While accurately identifying epitopes on HER2 targeted by trastuzumab (1N8Z), Epi4Ab also predicted several residues potentially recognized by pertuzumab (1S78) or by another novel anti-HER2

antibody hHERmAb-F0178 (3WSQ) when provided with trastuzumab's inputs. Trastuzumab and pertuzumab colocalize and co-bind to HER2 at distinct epitopes, whereas the synergistic effects exhibited by pertuzumab and hHERmAb-F0178 have been studied previously,[48–51] demonstrating the possible allosteric mechanism within the HER2 structure when bound by either antibody.

Since antibody and antigen interactions involve structural dynamics of both partners, structural dependency plays a crucial role in their binding mechanism, suggesting that preserving a certain degree of the structural dependency during the model's supervised learning process might be essential to capture the underlying conformational changes, thereby facilitating the identification of the near-native epitopes. To accommodate the structural dependencies involving antibody interactions, current antibody-specific epitope prediction methods incorporate more comprehensive antibody inputs; for example, SEPPA-mAb requires additional bound conformation of the antibody while EpiScan uses antibody sequences. As illustrated in EpiScan's test cases,[18] incorporating antibody sequences alone might be insufficient for accurate prediction of conformational epitopes. Direct comparison with EpiScan was not performed in this study due to various inconsistencies, such as source code incompatibility and mismatched antigen sequences. Additionally, obtaining an accurate antibody structure, even those based on homology models, remains challenging.

Along with this context, Epi4Ab used the sidechain reconstructions to "relax" the bound conformation of the antigens, partially mimicking the unbound conformation, and augmented the training data with AlphaFold models (approximately 10% of the training set) to reduce the structural dependencies. However, the moderate performance of Epi4Ab, along with other tools benchmarked in this study (Table 1) highlight the ongoing difficulty of accurately predicting conformational epitopes.

Limited data is a crucial factor. Despite the abundance of available antigen/antibody sequences, high-quality structure data are scarce, presenting a substantial hurdle. The dominance of the conformational epitopes, which consist of distant residues forming a three-dimensional binding region on an antigen, affirms the involvement of structural dynamics underlying the interaction mechanism between antibodies and antigens. Static structural models of Ab–ag complexes may not fully capture or reflect these dynamics. Consequently, computational determination of "ground truth" epitopes can be ambiguous due to limited experimental annotations, leading to high false positive rates. It should be acknowledged that the static model of the antigen used as input for prediction might not accurately represent the configuration that the antigen assumes in solution.

Despite the clear need for advancements, various approaches and concepts in the current epitope prediction field are being developed with increasing capability and precision. Contributing to this, our epitope prediction model for specific antibody VH-VL families and CDRs sequences, Epi4Ab, provides a more holistic view of Ab–ag complexes. While there is considerable scope for further improvement, it is currently feasible to predict epitopes with minimal antibody input. Improved models could offer deeper insights into the role of diverse antibody VH-VL and CDRs combinations (particularly H3 and L1) in antigen targeting. The identified epitopes may facilitate the engineering of antibody CDR-H3 patterns for improved antigen binding, thereby contributing to therapeutics repurposing.

Acknowledgments

We thank Khin Bhone Pyae in assisting the model benchmarking. The authors used Copilot, a built-in in the Microsoft 365 (Word) to assist in language improvement.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Medical Research Council grant NMRC-OFYIRG (MOH-000661).

ORCID

Chinh Tran-To Su http://orcid.org/0000-0001-6465-5987

Data availability statement

The source codes of Epi4Ab can be found here https://github.com/AMPMgroup/Epi4Ab..

References

1. Lu R-M, Hwang Y-C, Liu I-J, Lee C-C, Tsai H-Z, Li H-J, Wu H-C. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020;27(1). doi: 10.1186/s12929-019-0592-z.
2. Hernandez I, Bott SW, Patel AS, Wolf CG, Hospodar AR, Sampathkumar S, Shrank WH. Pricing of monoclonal antibody therapies: higher if used for cancer? Am J Manag Care. 2018;24(2):109–112.
3. Vasquez M, Krauland E, Walker L, Wittrup D, Gerngross T. Connecting the sequence dots: shedding light on the genesis of antibodies reported to be designed in silico. Mabs-austin. 2019;11(5):803–808. doi: 10.1080/19420862.2019.1611172.
4. Kohler G, Milstein C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature. 1975;256(5517):495–497. doi: 10.1038/256495a0.
5. Feldhaus MJ, Siegel RW, Opresko LK, Coleman JR, Feldhaus JMW, Yeung YA, Cochran JR, Heinzelman P, Colby D, Swers J, et al. Flow-cytometric isolation of human antibodies from a nonimmune saccharomyces cerevisiae surface display library. Nat Biotechnol. 2003;21(2):163–170. doi: 10.1038/nbt785.
6. Finlay WJ, Lugovskoy AA. De Novo discovery of antibody drugs – great promise demands scrutiny. Mabs-austin. 2019;11(5):809–811. doi: 10.1080/19420862.2019.1622926.
7. Khan SU, Fatima K, Aisha S, Malik F. Unveiling the mechanisms and challenges of cancer drug resistance. Cell Commun Signal. 2024;22(1):109. doi: 10.1186/s12964-023-01302-1.
8. Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: improved B-cell epitope prediction using protein language models. Protein Sci: A Publ Of The Protein Soc. 2022;31(12):e4497. doi: 10.1002/pro.4497.
9. Ivanisenko NV, Shashkova TI, Shevtsov A, Sindeeva M, Umerenkov D, Kardymon O. SEMA 2.0: web-platform for B-cell conformational epitopes prediction using artificial intelligence. Nucleic Acids Res. 2024;52(W1):W533–W539. doi: 10.1093/nar/gkae386.
10. Collatz M, Mock F, Barth E, Hölzer M, Sachse K, Marz M. EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics. 2021;37(4):448–455. doi: 10.1093/bioinformatics/btaa773.
11. Park M, Seo S-W, Park E, Kim J. EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. bioRXiv. 2022; doi: 10.1101/2022.02.27.481241.
12. Ponomarenko J, Bui H-H, Li W, Fusseder N, Bourne PE, Sette A, Peters B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinf. 2008;9(1):514. doi: 10.1186/1471-2105-9-514.
13. Silva BMD, Myung Y, Ascher DB, Pires DEV. epitope3D: a machine learning method for conformational B-cell epitope prediction. Briefings Bioinf. 2022;23(1):1–8. doi: 10.1093/bib/bbab423.
14. Zhou C, Chen Z, Zhang L, Yan D, Mao T, Tang K, Qiu T, Cao Z. SEPPA 3.0–enhanced spatial epitope prediction enabling glycoprotein antigens. Nucleic Acids Res. 2019;47(W1):W388–W394. doi: 10.1093/nar/gkz413.
15. Kringelum JV, Lundegaard C, Lund O, Nielsen M, Peters B. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol. 2021;8(12):e1002829. doi: 10.1371/journal.pcbi.1002829.
16. Cia G, Pucci F, Rooman M. Critical review of conformational B-cell epitope prediction methods. Briefings Bioinf. 2023;24(1):1–9. doi: 10.1093/bib/bbac567.
17. Krawczyk K, Liu X, Baker T, Shi J, Deane CM. Improving B-cell epitope prediction and its application to global antibody-antigen docking. Bioinformatics. 2014;30(16):2288–2294. doi: 10.1093/bioinformatics/btu190.
18. Wang C, Wang J, Song W, Luo G, Jiang T. EpiScan: accurate high-throughput mapping of antibody-specific epitopes using sequence information. NPJ Syst Biol Appl. 2024;10(1):101. doi: 10.1038/s41540-024-00432-7.
19. Qiu T, Zhang L, Chen Z, Wang Y, Mao T, Wang C, Cun Y, Zheng G, Yan D, Zhou M, et al. SEPPA-mAb: spatial epitope prediction of protein antigens for mAbs. Nucleic Acids Res. 2023;51(W1):W528–W534. doi: 10.1093/nar/gkad427.
20. Janeway C, Travers P, Walport M, Shlomchik M. Immunobiology: the immune System in health and disease. 5th edn, ((NY): Garland Science; 2001.
21. Peng H-P, Lee KH, Jian J-W, Yang A-S. Origins of specificity and affinity in antibody-protein interactions. Proc Natl Acad Sci USA. 2014;111(26):E2656–E2665. doi: 10.1073/pnas.1401131111.

22. Hsu H-J, Lee K, Jian J-W, Chang H-J, Yu C-M, Lee Y-C, Chen I-C, Peng H-P, Wu C, Huang Y-F, et al. Antibody variable domain interface and framework sequence requirements for stability and function by high-throughput experiments. Structure. 2014;22(1):22–34. doi: 10.1016/j.str.2013.10.006.
23. Koenig P, Lee CV, Walters BT, Janakiraman V, Stinson J, Patapoff TW, Fuh G. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. PNAS. 2017;114(4):E486–E495. doi: 10.1073/pnas.1613231114.
24. Wang F, Sen S, Zhang Y, Ahmad I, Zhu X, Wilson IA, Smider VV, Magliery TJ, Schultz PG. Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. Proc Natl Acad Sci USA. 2013;110(11):4261–4266. doi: 10.1073/pnas.1301810110.
25. Su CTT, Ling WL, Lua WH, Poh JJ, Gan SKE. The role of antibody Vκ framework 3 region towards antigen binding: effects on recombinant production and protein L binding. Sci Rep. 2017;7(1):3766. doi: 10.1038/s41598-017-02756-3.
26. Ling WL, Lua W-H, Poh J-J, Yeo JY, Lane DP, Gan SKE. Effect of VH-VL families in pertuzumab and trastuzumab recombinant production, Her2 and FcγIIA binding. Front Immunol. 2018;9:469. doi: 10.3389/fimmu.2018.00469.
27. Ling W-L, Chinh T-TS, Lua W-H, et al. Variable-heavy (VH) families influencing IgA1&2 engagement to the antigen, FcαRI and superantigen proteins G, A, and L. Sci Rep. 2022;12(1):6510. doi: 10.1038/s41598-022-10388-5.
28. Zhao J, Nussinov R, Ma B. The allosteric effect in antibody-antigen recognition. Methods Mol Biol. 2021;2253:175–183. doi: 10.1007/978-1-0716-1154-8_11.
29. He K, Zhang X, Ren S, Sun J. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA; p. 770–778. https://ieeexplore.ieee.org/document/7780459
30. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–1130. doi: 10.1126/science.ade2574.
31. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. ArXiv. 2021; doi: 10.48550/arXiv.2112.07782.
32. Vaswani A, Shazeer N, Parmar N, et al. 31st Conference on Neural Information Processing Systems (NIPS); Long Beach, CA.
33. Velickovic P, Guillem C, Arantxa C, et al. Graph attention networks. ArXiv. 2018; doi: 10.48550/arXiv.1710.10903.
34. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):D1140–D1146. doi: 10.1093/nar/gkt1043.
35. Li W, Godzik A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–1659. doi: 10.1093/bioinformatics/btl158.
36. Pj C, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423. doi: 10.1093/bioinformatics/btp163.
37. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y, Pieper U, Sali A. Comparative protein structure modeling with MODELLER. Curr Protocol In Bioinformatics. 2006;15(1):.5.6.1–.5.6.30. doi: 10.1002/0471250953.bi0506s15.
38. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33(Web Server):W382–388. doi: 10.1093/nar/gki387.
39. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004;32(Web Server):W665–W667. doi: 10.1093/nar/gkh381.
40. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. AlphaFold protein structure Database in 2024: providing structure coverage for over 241 million protein sequences. Nucleic Acids Res. 2024;52(D1):D368–D375. doi: 10.1093/nar/gkad1011.
41. Gowers RJ. S Benthall Rostrup S. eds98–105 in Proceedings of the 15th Python in Science Conference.
42. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. doi: 10.1093/molbev/mst010.
43. Nadalin F, Carbone A, Valencia A. Protein-protein interaction specificity is captured by contact preferences and interface composition. Bioinformatics. 2018;34(3):459–468. doi: 10.1093/bioinformatics/btx584.
44. Defferrard M, Bresson X, Vandergheynst P. NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems; Barcelona, Spain; p. 3844–3852.
45. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, Alaska, USA; p. 2623–2631.
46. Hoie MH, Gade FS, Johansen J, Würtzen C, Winther O, Nielsen M, Marcatili P. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol. 2024;15:1322712. doi: 10.3389/fimmu.2024.1322712.

47. Lua WH, Su CTT, Yeo JY, Poh J-J, Ling W-L, Phua S-X, Gan SKE. Role of the IgE variable heavy chain in FcεRIα and superantigen binding in allergy and immunotherapy. J Allergy Clin Immunol Pract. 2019;144(2):514–523. e515. doi: 10.1016/j.jaci.2019.03.028.
48. Lua WH, Gan SKE, Lane DP, Verma CS. A search for synergy in the binding kinetics of trastuzumab and pertuzumab whole and F (ab) to Her2. NPJ Breast Cancer. 2015;1(1). doi: 10.1038/npjbcancer.2015.12.
49. Toth G, Szöőr Á, Simon L, Yarden Y, Szöllősi J, Vereb G. The combination of trastuzumab and pertuzumab administered at approved doses may delay development of trastuzumab resistance by additively enhancing antibody-dependent cell-mediated cytotoxicity. Mabs-austin. 2016;8(7):1361–1370. doi: 10.1080/19420862.2016.1204503.
50. Fu W, Wang Y, Zhang Y, Xiong L, Takeda H, Ding L, Xu Q, He L, Tan W, Bethune AN, et al. Insight into HER2 signaling from step-by-step optimization of anti-HER2 antibodies. Mabs-austin. 2014;6(4):978–990. doi: 10.4161/mabs.28786.
51. Fuentes G, Scaltriti M, Baselga J, Verma CS. Synergy between trastuzumab and pertuzumab for human epidermal growth factor 2 (Her2) from colocalization: an in silico based mechanism. Breast Cancer Research. 2011;13(3). doi: 10.1186/bcr2888.