BMC Bioinformatics

RESEARCH

Open Access

Prediction of antibody-antigen interaction based on backbone aware with invariant point attention

Miao Gu[1], Weiyang Yang[1] and Min Liu[1*]

*Correspondence:
lium@tsinghua.edu.cn

[1] Department of Automation, Tsinghua University, Beijing 100084, China

Abstract

Background: Antibodies play a crucial role in disease treatment, leveraging their ability to selectively interact with the specific antigen. However, screening antibody gene sequences for target antigens via biological experiments is extremely time-consuming and labor-intensive. Several computational methods have been developed to predict antibody-antigen interaction while suffering from the lack of characterizing the underlying structure of the antibody.

Results: Beneficial from the recent breakthroughs in deep learning for antibody structure prediction, we propose a novel neural network architecture to predict antibody-antigen interaction. We first introduce AbAgIPA: an antibody structure prediction network to obtain the antibody backbone structure, where the structural features of antibodies and antigens are encoded into representation vectors according to the amino acid physicochemical features and Invariant Point Attention (IPA) computation methods. Finally, the antibody-antigen interaction is predicted by global max pooling, feature concatenation, and a fully connected layer. We evaluated our method on antigen diversity and antigen-specific antibody-antigen interaction datasets. Additionally, our model exhibits a commendable level of interpretability, essential for understanding underlying interaction mechanisms.

Conclusions: Quantitative experimental results demonstrate that the new neural network architecture significantly outperforms the best sequence-based methods as well as the methods based on residue contact maps and graph convolution networks (GCNs). The source code is freely available on GitHub at https://github.com/gmthu66/AbAgIPA.

Keywords: Antibody-antigen interaction, Invariant point attention, Antibody structure prediction

Introduction

Identifying the gene or amino acid sequence of monoclonal antibodies with specific binding capabilities to disease-related antigen epitopes is a prerequisite for de novo antibody design and affinity optimization processes [1, 2]. From the perspective of protein function, individual antibody sequence data only provide a partial view of immunity. The

ability of an antibody to specifically bind to an antigen is directly related to the conformation of the antibody's paratope, which is formed by the folding of complementary determining region (CDR) amino acid residues that are not adjacent in sequence but are spatially adjacent [3]. The exploration of the interaction relationship between monoclonal antibody sequences and antigens mainly relies on conventional experiments such as phage display, pseudovirus assays, and enzyme-linked immunosorbent assays [4, 5]. Although these antibody-antigen (Ab-Ag) interaction experimental techniques have been widely used in the field of antibody discovery and optimization, conducting related experiments is time-consuming and laborious. With the development of artificial intelligence and structural biology, computational methods have received increasing attention [6, 7].

Antibody-antigen interactions are a very atypical case of protein-protein interactions because Ab and Ag do not share a similar co-evolutionary history [8, 9] like regular PPI partners. As the functionality of proteins is intricately linked to their three-dimensional conformations [10], deep learning methods leveraging known structures have made substantial strides in predicting Ab-Ag interaction interfaces and affinities. PECAN [11] employed Graph Neural Network (GCN) [12] to effectively predict epitope and paratope interfaces in the presence of high bias in positive and negative samples. CSM-AB [13] achieved affinity prediction for mutant antibodies based on amino acid contact maps. DLAB [14] accomplished binding prediction of Ab and Ag through a highly accurate 3D grid representation of known structures and Convolutional Neural Network (CNN). However, DLAB primarily focuses on scoring docking poses of known or predicted paratope-epitope structures, which limits its applicability to cases where the antibody structure is unknown. In the context of binding prediction tasks, the input of the antibody side is a large-scale antibody amino acid sequence [15], and it is impractical to obtain accurate crystal structures.

Due to the information gap between protein sequences and their functions, sequence-based methods for predicting Ab-Ag interactions often rely on extensive datasets of monospecific antibody sequences [16]. Mason [17] leveraged approximately ten thousand mutants of Trastuzumab binding data with the antigen human epidermal growth factor receptor 2 (HER2) and then proposed one-dimensional Convolutional Neural Network (1DCNN) to guide antibody evolution experiments. Furthermore, employing the sequences of 5879 anti-CTLA-4 antibodies and 6218 anti-PD-1 antibodies, [18] constructed separate antibody binding prediction models for each antigen. However, sequence-based methods rely heavily on extensive antibody sequencing data, and the lack of antigen information as input prohibits their adaptation on tasks of antigen-specific interaction prediction. Recently, AbAgIntPre [19] employed the composition of k-spaced amino acid pairs (CKSAAP) encoding, which is widely used in PPIs [20, 21], to establish feature matrix representations for the sequence of antibody/antigen amino acid. Building on a two-dimensional Convolutional Neural Network (2DCNN), they conducted Ab-Ag interaction prediction tasks, including antigenic diversity interactions and severe acute respiratory syndrome coronavirus (SARS-CoV) specific interactions, and achieved state-of-the-art performance in sequence-based approaches.

Despite the significant progress in the field of Ab-Ag interaction prediction, there are still challenges to address: (i) Structure-based methods heavily rely on high-precision

antibody and antigen structures, limiting their applicability; (ii) Sequence-based methods fail to effectively excavate implicit structural information within antibody sequences, resulting in limited predictive performance. Consequently, bridging the information gap between antibody sequences and binding specific antibody conformations poses a challenge in predicting Ab-Ag interactions when antibody structures remain unknown. Recently, in the broader domain of protein-protein interaction, SGPPI [22] utilized AlphaFold2 [23] to predict protein structures based on their sequences. Subsequently, geometric and structural features were extracted to establish a graph convolutional neural network-based PPI prediction model. The experimental results indicated that incorporating implicit structural information from protein sequences significantly improved the effectiveness of PPI prediction.

The protein structure data predicted by AlphaFold2 for naturally occurring proteins have been released and widely utilized [24]. However, due to the time-consuming nature of Multiple Sequence Alignment [25], AlphaFold2 is not suitable for large-scale prediction of non-natural antibody sequence structures [26, 27]. Encouragingly, there have been notable advancements in structure prediction for antibody. ABlooper [28] predicted CDR loop structures and provided confidence estimates, while it struggled to integrate CDR loop templates. DeepAb [29] utilized the residual dilated convolutional neural network and a criss-cross attention module [30] to predict geometric constraints among residues, which were fed into trRosetta [31] to obtain the complete prediction of the antibody variable fragments (Fv) structure while suffering from a time-consuming optimization process. To address this problem, IgFold [32] achieved a fast and direct prediction of antibody skeletal atomic coordinates based on the pre-trained antibody language model AntiBERTy [33].

The increasing array of protein structure prediction methods is being employed to address interaction-related challenges in the absence of known protein structures. Similar to SGPPI [22], methods utilizing GNN and binary residue contact maps derived from predicted protein structures have been applied to multiple tasks, including the prediction of protein function [34], protein interaction sites [35], and protein binding [36]. Although these studies have inspired efforts to bridge the gap between antibody sequences and structures, enhancing the potential for Ab-Ag binding predictions, they still struggle to effectively represent potential structural features and fail to capture geometric features such as relative positions and angles between residues due to limited available structural data for antibody-antigen interactions, particularly in the regions of antibody determinants characterized by flexible loop structures [10]. Therefore, it is desirable to explore novel approaches for representing potential structural features to improve the prediction of antibody-antigen interaction.

Inspired by recent strides in protein structure prediction, we propose a novel antibody-antigen interaction prediction neural network framework. We utilize IgFold [32] to obtain predicted antibody structures, using nonlinear epitope [37] candidate structures as inputs on the antigen side, but we discard the traditional approach of employing residue contact maps and GCN for structure feature extraction. Instead, for the first time in the field of Ab-Ag interaction prediction, we employ backbone framework description, which integrates rotation matrices and translation vectors to represent the positions of residue-heavy atoms relative to the origin, to establish antibody/antigen

structure description. And for the first time in the field of Ab-Ag interaction prediction, we use the Invariant Point Attention (IPA) [23, 38] mechanism for feature aggregation and updating of residue nodes on the backbone frame. We conduct experiments using different neural network architectures on an antigen-diverse interaction dataset and an antigen-specific interaction dataset. Our findings demonstrate the effectiveness of our approach in capturing the latent structural features of antibodies and antigens, overcoming the limitations of mainstream methods based on GCNs in describing the relative positional relation-ships between residues (beyond distance considerations). Quantitative experimental results indicate that our method excels in predicting interactions involving antigenic diversity interactions, as well as SARS-CoV specific interactions.

Materials and methods

We have proposed a novel deep learning framework, prediction of Antibody and Antigen interaction based on backbone aware with Invariant Point Attention (AbAgIPA), aimed at learning the potential structural features of antigen candidate epitopes and antibodies for predicting their interaction relationships.

The architecture is inspired by the conformational binding characteristics of antibodies and antigens. As shown in **Fig.** 1A, the regions determining whether an interaction occurs between the antigen and antibody are the antibody paratope and the antigen epitope. In this framework, we harness antibody structure predictions to bridge the gap

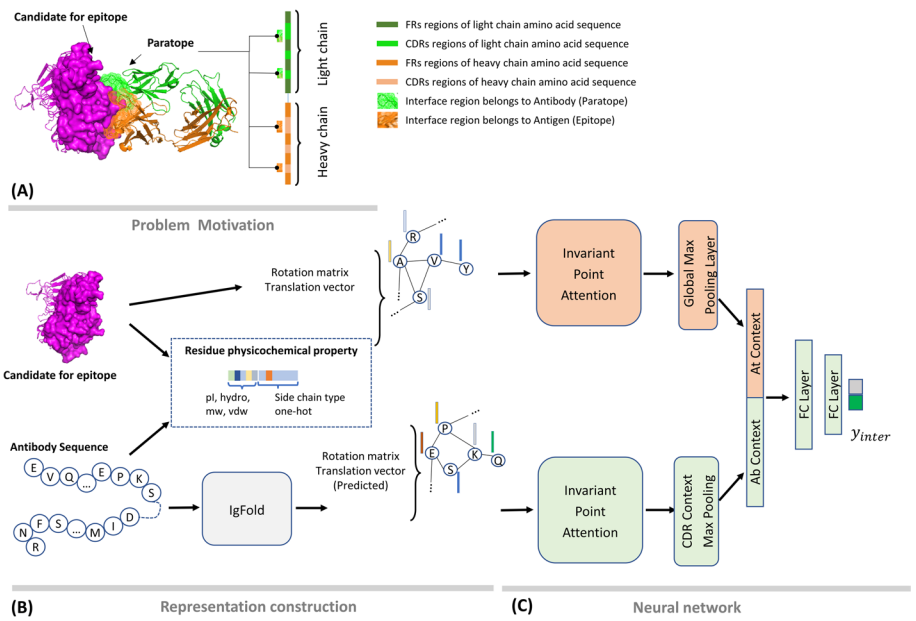

Fig. 1 Schematic overview of AbAgIPA. (A) Problem motivation. Antibody-antigen binding is directly determined by paratope conformation and candidate epitope conformation. (B) Representation construction. Inputs are antibody sequences and candidate epitope structures, establish backbone framework for the antibody structure obtained by IgFold and the known antigen structure, use amino acid physicochemical properties as residue node features. (C) Based on the residue node feature and backbone frame, Input the amino acid node features and backbone frame of antibody and antigen, and utilize invariant point attention layer and global max pooling to get the interaction pattern feature vector of antibody and antigen, in which CDR region mask is used in extracting the antibody interaction pattern feature vector. Further, the interaction labels were obtained using feature splicing and fully connected layer

between antibody sequences and potential structural features relevant to interactions. As shown in **Fig.** 1B, we establish a representation method incorporating a backbone framework describing the relative positions of residues and physicochemical features of amino acid nodes. Subsequently, utilizing modified invariant point attention within the framework, as shown in **Fig.** 1C, amino acid node features are propagated and updated, resulting in a feature representation of interaction patterns integrated with potential structures, enabling interaction prediction.

Problem statement: Given an antigen structure and an antibody sequence, to assign a label, either interact (positive class) or non-interact (negative class).

Representation construction

The interaction between antibody and antigen is primarily deter-mined by spatial complementarity and matching physicochemical properties. Thus, we have established the geometric representation of the backbone and the physicochemical representation of the residue node.

Geometric representation of backbone

Our framework does not rely on residue contact maps for geometric representation. Instead, similar to protein prediction method AlphaFold2, we utilize a rotation matrix and translation vector for the backbone geometry features of amino acids with heavy atoms relative to the origin. Specifically, with the coordinates of heavy atoms(alpha carbon, beta carbon, nitrogen, and oxygen atoms), we obtain the rotation matrix $R_i \in \mathbb{R}^{3\times3}$ and translation vector $t_i \in \mathbb{R}^{3\times1}$ for each residue relative to the ideal residue positions at the origin, the above computation is shown in the pseudo-code of the algorithm in the Supplementary Fig 1 and 2. Then, a framework representation for the i-th amino acid on the backbone is derived. Subsequently, this $T_i := \left(R_i, t_i\right)$ leads to the backbone representation $T^b = \{T_i^b\}_{i=1}^l$ of the antibody predicted structure and the backbone representation $T^g = \{T_i^g\}_{i=1}^L$ of the antigen candidate epitope structure. Here, i represents the i-th amino acid node, and p represents the heavy atom type; the superscript b denotes the notation for the antibody, and the superscript g denotes the notation for the antigen. l represents the number of antibody residues, and L represents the number of antigen residues.

Physicochemical representation of residue node

We established representation vectors for the residue node, denoted as $x_i \in \mathbb{R}^{10}$, based on the types of amino acid side chains and the physicochemical properties of amino acids. This hybrid vector encompasses both real-value vectors and one-hot vectors. The ranges and indices for various features are pre-sented in (Supplementary Table 1 and 2). (i) A vector of length four consisting of isoelectric point (pI), hydrophobicity (hydro), molecular weight (mw), and van der Waals radius (vdw). (ii) A one-hot vector of length six indicating amino acid side chain type, including aliphatic side chain, aromatic side chain, neutral side chain, positively charged side chain, negatively charged side chain, and specialized amino acid side chain. Based on the aforementioned amino acid node feature encoding method, we obtain the feature matrices $X^g = \{x_i^g\}_{i=1}^L$ for antigen candidate epitope amino acid nodes and $X^b = \{x_i^b\}_{i=1}^l$ for antibody amino acid nodes. Here, i represents the i-th amino acid node, the superscript b denotes the notation for the antibody, and the superscript g

denotes the notation for the antigen. l represents the number of antibody residues, and L represents the number of antigen residues.

Neural network

We construct a neural network consisting of an invariant point attention layer, a max-pooling layer, and a fully connected layer to characterize antibodies and antigens separately. This enables the extraction of interaction pattern feature vectors for classification prediction.

Invariant point attention

Invariant Point Attention (IPA) is an attentional mechanism that can introduce backbone information based on backbone representation T_i into the feature update process for residue nodes and edges, was first used for protein structure prediction tasks. IPA has been proved to be invariant under the global euclidean transformations of the above framework [38]. In order to efficiently incorporate the key structural information related to interactions under small samples, we modified invariant point attention layer to introduce backbone geometric representation for Ab-Ag interface pattern feature extraction. The implementation architecture of modified Invariant Point Attention module is shown in (Supplementary Fig 3 and 4).

Firstly, perform an unbiased linear transformation on the input residue node feature matrix as follows:

$$q_i^h, k_i^h, v_i^h = W_{h,d} x_i \tag{1}$$

$$q_i^{hp}, k_i^{hp}, v_i^{hp} = W_{h.p} x_i \tag{2}$$

where $W_{h,d} \in \mathbb{R}^{3d \times d}$ is the learnable parameters for the node feature vector x_i to produce $q_i^h, k_i^h, v_i^h \in \mathbb{R}^d$, which are query key and value features on the h-th head, $W_{h,p} \in \mathbb{R}^{(3\times3)\times d}$ is the learnable parameters that enable the implicit layer node vec-tor to produce $q_i^{hp}, k_i^{hp}, v_i^{hp} \in \mathbb{R}^3$ which are query key and value features of p-th heavy atom of residue i on the h-th head, they implying localized incremental translation information on the existing backbone.

We calculate attention scores between pairs of amino acid nodes i and j are based on residue features, backbone frame and point features. The calculation process is shown as follows:

$$a_{ij}^h = softmax_j \left(w_M \left(\frac{q_i^{h^T} k_j^h}{\sqrt{d}} - \frac{\gamma^h w_Q}{2} \sum_p \left|\left| T_i \circ q_i^{hp} - T_j \circ k_j^{hp} \right|\right|^2 \right) \right) \tag{3}$$

in this formulation, $q_i^{h^T} k_j^h / \sqrt{d}$ represents the attention score component derived from the inner product of node features for nodes i and j, normalized by the dimension d. The term $\frac{\gamma^h w_Q}{2} \sum_p \left|\left| T_i \circ q_i^{hp} - T_j \circ k_j^{hp} \right|\right|^2$ quantifies the distance affinity, as outlined in [23], which computes the sum of squared Euclidean distances between transformed coordinates of heavy atoms in the backbone structure. Here, $T_i \circ q_i^{hp}$ and $T_j \circ k_j^{hp}$ denote the transformed features of the p-th heavy atom from residues i and j, incorporating relative positional information post-transformation. The operator $\circ$ symbolizes Euclidean

transformations applied to frames defined by $T_i = (R_i, t_i)$, where $T_i \circ v = R_i v + t_i$, and v in $\mathbb{R}^3$ represents the atomic coordinates. The parameter γ^h is a tunable weight that adjusts the influence of different attention heads, enhancing the model's flexibility in learning from diverse structural data.

Furthermore, $w_Q = \sqrt{2/(9N_P)}$ is a scalar used to balance the contributions of the two components in calculating attention scores: the inner product based on node features and the distance affinity. Its value depends on the number of query heavy atoms, with $N_p = 4$. $w_M = \sqrt{1/3}$ is a scalar maintaining consistency with [23]. Subsequently, edge features and node features are aggregated using affinity and backbone as per the following equation:

$$o_i^h = \sum_j a_{ij}^h v_j^h \tag{4}$$

$$o_i^{hp} = T_i^{-1} \circ \sum_j a_{ij}^h (T_j \circ v_j^{hp}) \tag{5}$$

$$\widetilde{h}_\iota = W_o(concat_{h,p}(o_i^h, o_i^{hp}, \left|\left| o_i^{hp} \right|\right|)) \tag{6}$$

where o_i^h is the updated hidden layer residue node feature by attention scores, o_i^{hp} is the updated residue heavy atom point feature by attention scores and Euclidean transform by backbone frame, $W_o \in \mathbb{R}^{d\times(2d+1)}$ is the learnable parameters to obtain the updated node features $\tilde{h}_\iota \in \mathbb{R}^d$ by attention scores.

Based on the aforementioned invariant point attention calculation process, we obtain the antibody representation, we obtain the antibody representation $H^b = \{\tilde{h}_i^b\}_{i=1}^l$ and the antigen representation $H^g = \{\tilde{h}_i^g\}_{i=1}^L$ separately. In addition, based on the above IPA module, we have constructed both the parameter-sharing AbAgIPA-Twins framework and the parameter-non-sharing AbAgIPA-Paddle framework to evaluate the performance differences between neural networks with different structures.

Pooling and classification

We perform global max pooling on the antigen representation H^g and yields the antigen interaction pattern feature $h_{inter}^g \in \mathbb{R}^d$ as follows:

$$h_{inter}^g \in \max_{i\in\{0,\ldots L\}} \max_{j\in\{0,\ldots d\}} H_{ij}^g \tag{7}$$

where i represents the index of the i-th residue, and j represents the j-th dimension on the feature.

Similarly, we utilize the CDR indices obtained from abnumber [39] to perform max pooling on the antigen representation H^b to obtain the antibody interaction pattern feature $h_{inter}^g \in \mathbb{R}^d$ as follows:

$$h_{inter}^b = \max_{i\in S_{CDR}} \max_{j\in\{0,\ldots d\}} H_{ij}^b \tag{8}$$

where S_{CDR} represents the indices of residues belonging to the CDR area.

Then, the binding label between the antigen epitope and the antibody is finally obtained through the concatenation layer and the fully connected (FC) layer as follows:

$$h_{inter} = concat(h^{b}_{inter}, h^{g}_{inter}) \tag{9}$$

$$y_{inter} = W'_{h} ReLU(W_{h} h_{inter} + b) + b' \tag{10}$$

where $W_h \in \mathbb{R}^{2d \times 2d}$ and $W'_h \in \mathbb{R}^{2d \times 2}$ are learnable parameters of FC layers, b and b' are learnable bias of FC layers. y_inter is the label of interaction.

Dataset and evaluation metrics

Antigenic diversity Ab-Ag interaction dataset

(The Supplementary Fig 5) demonstrates the process of constructing the 5-fold cross-validation dataset from the Structural Antibody Database (SAbDab) [40]. Following the methodology outlined in [19] and utilizing CD-hit [41], antigens with sequence identity above 90% were considered identical. Correspondingly, their associated antibody-antigen interaction pairs were grouped into the same subset, resulting in 3800 pairs for positive samples. For each antigen represented in a positive sample, antibodies sampled from different subsets were employed to form the corresponding negative sample data, totaling 3800 pairs for negative samples. A phylogenetic tree for this subset collection was constructed using ClustalW [42], then divided into six clusters. To prevent data leakage, we ensured that the subsets to which antigens in the test set belonged did not overlap with those in the training set and the number of subsets belonging to the same cluster in the training and test sets was kept at approximately 4:1. Additionally, detailed information regarding the number of antigens and antibodies included in each subset for both training and test datasets has been provided in the Supplementary Table 3).Furthermore, we have included a statistical summary of antibody and antigen sequence lengths. The sequence length statistics from the 5-fold cross-validation experiments, including the mean, median, maximum, and minimum values, are detailed in Supplementary Table 4 and 5. We have also provided corresponding distribution histograms in Supplementary Figure 6. The antibody sequence lengths range from 208 to 250, while antigen sequence lengths range from 23 to 165. The distributions of antibody and antigen sequence lengths are fairly consistent across the training and test sets within each fold.

SARS-Cov specific interaction Ab-Ag interaction dataset

We collected positive and negative samples related to the interaction between SARS-CoV-1 and SARS-CoV-2 from the coronavirus antibody database (CoV-AbDab) [43]. Considering that there is a high degree of similarity between antibody sequences that bind the same antigen or similar antigens [44], we used CD-hit [41] for data redundancy reduction with a commonly used threshold parameter in antibody structure prediction [32], i.e., 98% antibody sequence identity. This resulted in a dataset comprising 7491 positive samples and 1672 negative samples, with 5561 positive samples specific to SARS-CoV-2 and 1930 positive samples specific

to SARS-CoV-1. Additionally, there were 708 negative samples for SARS-CoV-2 and 964 negative samples for SARS-CoV-1. To conduct experiments, we randomly and evenly partitioned the positive and negative samples from SARS-CoV-1 and SARS-CoV-2 into five folds each.

Evaluation metrics

All neural networks appearing in this paper were trained under the same conditions as (The Supplementary Table 6). We evaluate the predictive performance of the interaction model using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). AUROC and AUPRC are independent of the model's probability threshold setting. AUROC reflects the overall relationship between sensitivity and specificity, while AUPRC reflects the overall relationship between precision and recall. We also compute the specificity metric, representing the proportion of true negative samples among those predicted as negative. The precision, recall, F1 score and specificity can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

where TP, FP, TN and FN denote the number of true positive, true negative, false negative samples, respectively.

Results

We have implemented an antibody-antigen interaction prediction model for antigenic diversity in the presence of unknown antibody structures and a SARS-CoV specificity prediction model based on the proposed AbAgIPA framework as well as the dataset constructed in this paper.

To illustrate the advantages of the constructed framework in characterizing the latent structures of antibodies and antigens, we conducted a comparative analysis with the mainstream method based on residue contact maps and graphical neural networks, i.e., the SGPPI method, and the results show that our framework is more advantageous in characterizing the latent structures. In addition, we also compare the performance with the best available sequence-based antibody-antigen interaction method, AbAgInterPre. Ultimately, we show that the AbAgIPA framework provides a new and more effective approach to the problem of predicting antibody-antigen interactions with unknown antibody structures.

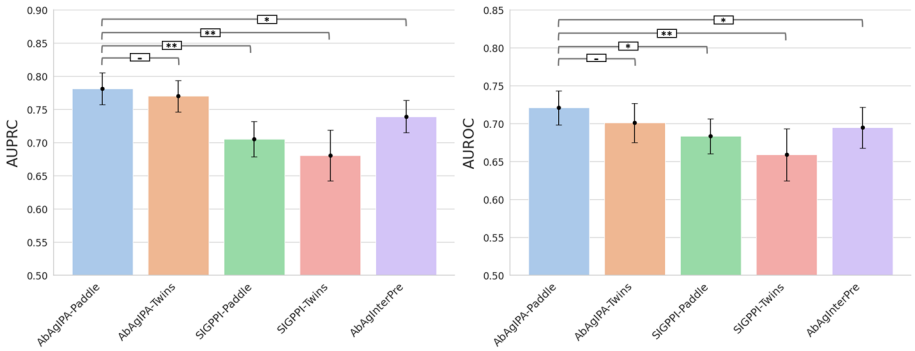

Fig. 2 Prediction performance through 5-fold cross-validation on AD-AbAg dataset. (A) AUPRC and (B) AUROC as the performance metric.(Wilcoxon-Mann–Whitney P-values: $^{**}P < 0.01$, $^{*}P < 0.05$, -not significant)

Table 1 Performance on the AD-AbAg database

Method	AUROC[a]	AUPRC[a]	Precision[a,b]	Recall[a,b]	Specificity[a,b]	F1 score[a,b]
AbAgIPA-Paddle	**0.721**	**0.781**	**0.810**	0.555	**0.865**	**0.654**
AbAgIPA-Twins	0.701	0.770	0.7560	0.561	0.821	0.643
SGPPI-Paddle[c]	0.683	0.705	0.653	0.564	0.700	0.603
SGPPI-Twins[c]	0.659	0.681	0.737	0.423	0.790	0.473
AbAgInterPre	0.694	0.739	0.666	**0.619**	0.686	0.640

[a]All metrics shown in the table are averages of 5-fold cross-validation;[b]We use a decision boundary of 0.5 to determine TP(True Positive), FP(False Positive), TN(True Negative) and FN(False Negative); [c]SGPPI-Twins are SGPPI's primitive structure, the GCN network layer corresponding to the two inputs is parameter shared, and SGPPI-paddle is the GCN network layer corresponding to the two inputs side by side, and the parameters are unshared

Antigenic diversity Ab-Ag interaction prediction

We provides a comprehensive summary of the predictive performance of different neural networks on an interaction dataset characterized by diverse antigen epitopes as Fig. 2. The analysis encompasses neural networks constructed based on various structures of the IPA module. Additionally, we include a comparison with the current top-performing methods: AbAgInterPre, which utilizes amino acid CKSAPP encoding and CNN, and SGPPI, a method for general protein interactions integrating predicted residue contact maps with graph convolutional networks (GCNs).

AbAgIPA-Paddle and AbAgIPA-Twins denote interaction pattern representations of antigenic epitopes and antibody backbones based on two parallel IPA modules and two parameter-sharing IPA modules. Similarly, SGPPI-Paddle and SGPPI-Twins represent interaction pattern features of antigenic epitopes and antibody backbones based on two parallel GCN modules and two parameter-sharing GCN modules. As shown in Fig. 2, AbAgIPA-Paddle and AbAgIPA-Twins were optimal and suboptimal in AUCPR and AUCROC, respectively.

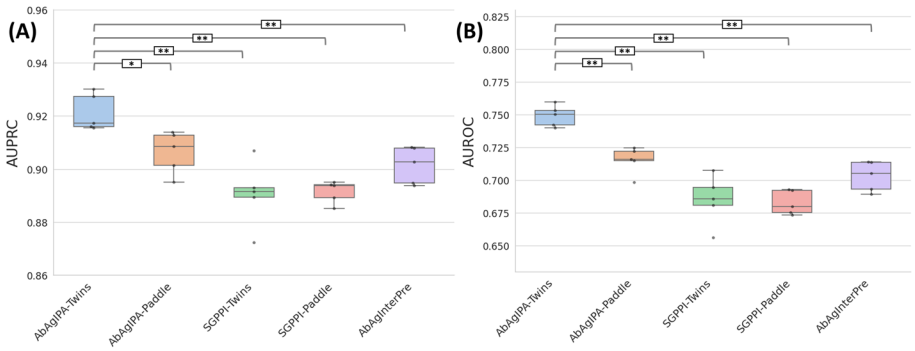

Fig. 3 Prediction performance through 5-fold cross-validation on SARS-Cov specific interaction dataset. (A) AUPRC and (B) AUROC as the prediction performance metric. (Wilcoxon-Mann–Whitney P-values: $^{**}P < 0.01$, $^{*}P < 0.05$, -not significant)

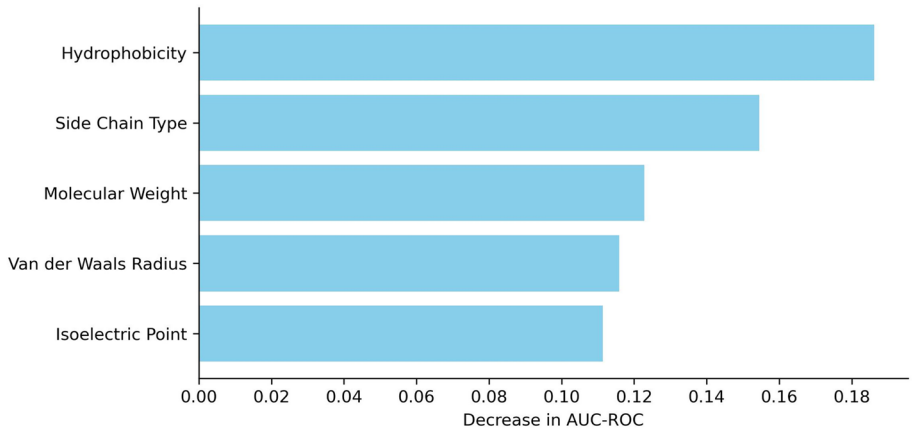

Fig. 4 Feature importance evaluated by decrease in AUC-ROC

Table 1 further quantifies the performance comparison between AbAgIPA and existing methods by introducing additional performance metrics, the results show that AbAgIPA achieves optimal performance in all metrics except Recall.

SARS-Cov specific interaction prediction

As shown in Fig. 3, compared with the current best-performing AbAgInterPre method based on amino acid CKSAPP coding and CNN, and the SGPPI method based on predicted structures with GCN in common protein interactions. In the scenario of imbalanced positive-to-negative sample ratio (1:5) within the dataset, AbAgIPA-Twins method achieves an AUPRC of 0.921 and an AUROC of 0.749 in a 5-fold cross-validation experiment. These results indicate that AbAgIPA-Twins outperforms other methods and exhibits relatively balanced performance. As shown in the (Supplementary Table 7), AbAgIPA-Twins achieved near-optimal metrics on Precision and Recall.

Feature importance and antibody spatial attention heatmap

As the method states, we derived a representation vector for each residue node based on amino acids' physicochemical properties and their side chains' characteristics. To determine which features predominantly influence the prediction of antibody-antigen

interactions, we conducted a feature permutation importance analysis [45] on the antigenic diversity Ab-Ag interaction prediction model, namely AbAgIPA-Paddle. By randomly shuffling each feature value and calculating the decrease in AUC-ROC before and after shuffling, we obtained a ranking of feature importance, as depicted in Fig. 4.

Feature importance analysis revealed that hydrophobicity is the most influential factor in predicting antibody-antigen interactions, consistent with the recognized role of hydrophobic residues in driving protein interactions [46]. Following hydrophobicity, the type of side chains significantly influences the predictions, underscoring their role in molecular recognition. In contrast, the isoelectric points of amino acids have minimal impact on prediction accuracy, suggesting their lesser relevance in the binding interfaces.

Antibody-antigen interactions are an atypical class of PPIs with different evolutionary patterns, where the immunological significance of antibody interchain features plays a pivotal role [47]. In line with the methodologies described in [47], we extracted attention values from the antigenic diversity AbAgIPA-Paddle model, focusing specifically on interactions between amino acids across the two antibody chains. These attention values were filtered to generate an interchain feature attention matrix, exclusively containing amino acid pairs that bridge the two chains. The resultant heatmap, displayed in Fig. 5A, visualizes these interchain interactions. To analyze these interactions further, we averaged the attention scores by sequence position and classified amino acid sites by their structural domains-CDR or FR. This classification allowed us to quantitatively evaluate the distribution of attention mean values, revealing a higher focus of interchain feature attention on the CDR regions than the FR regions, as illustrated in Fig. 5B. This analysis underscores the critical role of CDR regions in mediating Ab-Ag interactions. Additionally, the model proposed in this study demonstrates a certain level of interpretability, which is illustrated through the feature importance analysis and the spatial attention heatmap. These tools not only identify key amino acid residues influencing the antibody-antigen interactions but also provide visual insights into how these residues are spatially arranged to facilitate

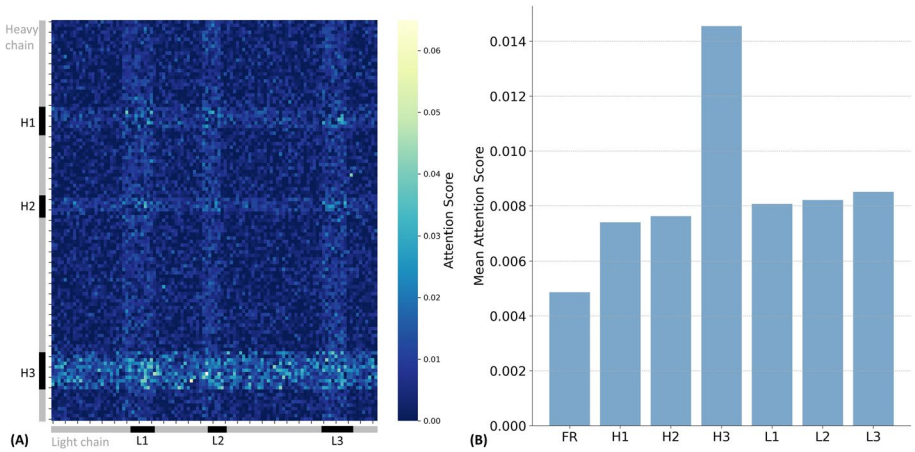

Fig. 5 Visualization of antibody position attention in AbAgIPA-Paddle model. H1/2/3: CDR1/2/3 of heavy chain. L1/2/3: CDR1/2/3 of light chain. FR: Framework Region

or inhibit interactions. This enhances our understanding of the underlying mechanisms of interaction and offers valuable insights for further antibody engineering and design efforts.

Performance discrepancy on actual structures and predicted structures

Keeping the network structure unchanged, we build residue contact graphs for the SGPPI method based on the predicted structure and the actual structure, and backbone framework representations for AbAgIPA, respectively, and compare the performance of both on the antigenic-diversity Ab-Ag interaction dataset, as shown in Fig. 6 and (Supplementary Table 8).

The results show that the performance of models using IgFold-predicted structures is significantly better than those using ESMFold-predicted structures, whether in the SGPPI or AbAgIPA framework. This indicates that IgFold outperforms ESMFold in predicting antibody-antigen interactions, consistent with IgFold's ability to leverage the pre-trained antibody language model AntiBERTy to predict skeletal atomic coordinates, resulting in more accurate and faster predictions for antibody-specific regions. Furthermore, there is no significant difference in performance between models using the actual structures and those using the predicted structures, regardless of whether SGPPI or AbAgIPA is used. This suggests that both SGPPI and AbAgIPA are robust to the accuracy of the predicted antibody structures. In addition, AbAgIPA consistently performs better than SGPPI, whether based on predicted or actual structures.

Discussion and conclusion

In this paper, we introduce AbAgIPA, a novel architecture based on a modified Invariant Point Attention (IPA) module. This method diverges from traditional approaches that rely on structural data to generate inter-residue contact maps and employ Graph Convolutional Networks (GCNs). Instead, AbAgIPA establishes a backbone framework that uses rotation matrices and translation vectors to describe the spatial relationships between amino acids. By incorporating not only distances but also relative angles between amino acids, this approach allows the model to extract richer geometric features from protein structures. The IPA mechanism, which is inherently rotation-invariant, further ensures that predictions remain consistent regardless of protein orientation.

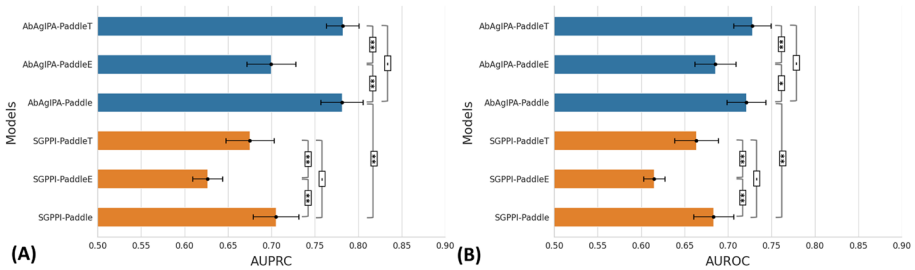

Fig. 6 Performance discrepancy on actual structures and predicted structures. (A) AUPRC and (B) AUROC as the prediction performance metric. The suffix "T" stands for the true structure, the suffix "E" stands for the ESMFold-predicted structure. (Wilcoxon-Mann–Whitney P-values: $^{**}P < 0.01$, $^{*}P < 0.05$, -not significant)

This combination of enhanced geometric representation and rotational invariance leads to a significant improvement in predicting antibody-antigen interactions.

Under the antibody-antigen interaction dataset with antigen diversity where the antigen similarity between those in the test set and those in the training set is strictly controlled, compared to antibody-antigen interaction methods based on predicted structures and Graph Convolutional Networks, the AbAgIPA method exhibits stronger robustness and higher accuracy under the same amino acid feature encoding. Notably, even when utilizing predicted structures, the performance of the AbAgIPA method surpasses the former. Compared to the best existing AbAgInterPre method based on CKSAPP coding and CNN, AbAgIPA shows a further improvement in AUCPR, AUROC, Precision and Specificity on antigenic diversity interaction prediction, and a remarkable improvement in AUCPR and AUROC on SARS-Cov specific interaction dataset.

A key strength of AbAgIPA is its interpretability. By visualizing attention heatmaps, we can see that the model effectively captures critical cross-chain interactions, particularly in the Complementarity Determining Regions (CDRs) of antibodies. This ability to identify key structural regions and interaction patterns not only enhances the model's performance but also provides deeper insights into the underlying mechanisms of antibody-antigen binding.

Our study could benefit clinical applications by improving the accuracy of predicting antibody-antigen interactions, particularly in therapeutic areas like oncology and immunology. This may speed up the discovery of antibodies with high affinity and specificity for disease-related antigens. Additionally, the model's ability to predict specific antigen interactions is valuable for identifying effective antibodies against evolving pathogens like SARS-CoV-2, aiding in rapidly developing treatments for emerging diseases.

In summary, this paper aims to integrate the antibody structure prediction with the challenge of predicting Ab-Ag interactions and to provide a new alternative to the traditional amino acid contact map representation and GCN architecture. Despite the competitive performance of AbAgIPA, several limitations must be acknowledged. First, the antigen-antibody interaction dataset constructed based on SAbDab has limited coverage of diverse antigen-antibody interaction modes. This restricts the generalizability of our model, especially in predicting interactions involving less common antigens or atypical antibodies. Additionally, while our model, developed on the SARS-CoV specific Ab-Ag interaction dataset, can predict the binding labels of antibody sequences to antigenic epitopes more accurately, its application is confined to SARS-Cov antibody-antigen pairs. This highlights a gap in the model's capabilities, underscoring the need for a more comprehensive and diverse dataset to improve the accuracy and broader applicability of antibody design predictions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05961-w.

Additional file

Supplementary file 1 (pdf 1024 KB)

Acknowledgements
Not applicable.

Author contributions
ML lead the investigation and the project, ML and MG designed the methodology, MG and WYY implemented the methods and analyzed the data, ML and MG wrote the original manuscript, ML, MG, WYY revised the manuscript, all authors approved the manuscript.

Funding
This research is supported by the National Natural Science Foundation of China (Grant No. 62173204).

Data availibility
The dataset and source code is freely available on Git Hub at https://github.com/gmthu66/AbAgIPA.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no Conflict of interest.

Received: 18 August 2024 Accepted: 16 October 2024
Published online: 06 November 2024

References

1. Akbar R, Bashour H, Rawat P, Robert PA, Smorodina E, Cotet T-S, Flem-Karlsen K, Frank R, Mehta BB, Vu MH. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. In: MAbs, vol. 14, p. 2008790. Taylor & Francis. https://doi.org/10.1080/19420862.2021.2008790
2. Makowski EK, Kinnunen PC, Huang J, Wu LN, Smith MD, Wang TX, Desai AA, Streu CN, Zhang YL, Zupancic JM, Schardt JS, Linderman JJ, Tessier PM. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. Nat Commun 2022;13(1) https://doi.org/10.1038/s41467-022-31457-3
3. Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner BD, Hu XZ, Adachi Y, Schief WR, Dunbrack RL. Rosettaantibodydesign (rabd): A general framework for computational antibody design. Plos Comput Biol 2018;14(4) https://doi.org/10.1093/bioinformatics/btac016
4. Zhang J, Du YS, Zhou PF, Ding JR, Xia S, Wang Q, Chen FY, Zhou M, Zhang XM, Wang WF, Wu HY, Lu L, Zhang ST. Predicting unseen antibodies' neutralizability via adaptive graph neural networks. Nat Mach Intell. 2022;4(11):964–76. https://doi.org/10.1038/s42256-022-00553-w.
5. Chen G, Zhang S, Ma X, Wilson G, Zong R, Fu Q. Antibody mimics for precise identification of proteins based on molecularly imprinted polymers: Developments and prospects. Chem Eng J 448, 148115 (2023) https://doi.org/10.1016/j.cej.2023.148115
6. Makowski EK, Chen HT, Tessier PM. Simplifying complex antibody engineering using machine learning. Cell Syst. 2023;14(8):667–75. https://doi.org/10.1016/j.cels.2023.04.009.
7. Wilman W, Wróbel S, Bielska W, Deszynski P, Dudzic P, Jaszczyszyn I, Kaniewski J, Mlokosiewicz J, Rouyan A, Satlawa T, Kumar S, Greiff V, Krawczyk K. Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. Briefings Bioinf 2022;23(4)
8. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel J-C, Deane CM. Progress and challenges in predicting protein interfaces. Brief Bioinform. 2016;17(1):117–31. https://doi.org/10.1093/bib/bbv027.
9. Hou Q, Stringer B, Waury K, Capel H, Haydarlou R, Xue F, Abeln S, Heringa J, Feenstra KA. Serendip-ce: sequence-based interface prediction for conformational epitopes. Bioinformatics. 2021;37(20):3421–7. https://doi.org/10.1093/bioinformatics/btab321.
10. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody structure and function: The basis for engineering therapeutics. Antibodies **8**(4) (2019) https://doi.org/10.1016/j.heliyon.2023.e15032
11. Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics. 2020;36(13):3996–4003. https://doi.org/10.1093/bioinformatics/btaa263.
12. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. Graph neural networks: A review of methods and applications. AI open 1, 2020;57–81. https://doi.org/10.1007/s11042-010-0645-5
13. Myung Y, Pires DEV, Ascher DB. Csm-ab: graph-based antibody-antigen binding affinity prediction and docking scoring function. Bioinformatics. 2022;38(4):1141–3. https://doi.org/10.1093/bioinformatics/btab762.
14. Schneider C, Buchanan A, Taddese B, Deane CM. Dlab: deep learning methods for structure-based virtual screening of antibodies. Bioinformatics. 2022;38(2):377–83. https://doi.org/10.1093/bioinformatics/btab660.
15. Fischman S, Ofran Y. Computational design of antibodies. Curr Op Struct Biol 2018;51:56–162. https://doi.org/10.1016/j.sbi.2018.04.007
16. Yuan Y, Chen Q, Mao J, Li G, Pan X. Dg-affinity: predicting antigen-antibody affinity with language models from sequences. BMC Bioinf. 2023;24(1):430. https://doi.org/10.1186/s12859-023-05562-z.

17. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, Ehling RA, Bonati L, Dahinden J, Gainza P, Correia BE, Reddy ST. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng. 2021;5(6):600. https://doi.org/10.1038/s41551-021-00699-9.
18. Lim YW, Adler AS, Johnson DS. Predicting antibody binders and generating synthetic antibodies using deep learning. In: MAbs, vol. 14, p. 2069075. Taylor & Francis. https://doi.org/10.1080/19420862.2022.2069075
19. Huang Y, Zhang ZD, Zhou Y. Abagintpre: A deep learning method for predicting antibody-antigen interactions based on sequence information. Front Immunol **13** (2022) https://doi.org/10.3389/fimmu.2022.1053617
20. Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Eng Des Select. 2009;22(11):707–12. https://doi.org/10.1093/protein/gzp055.
21. Wei LY, Liao MH, Gao X, Zou Q. An improved protein structural classes prediction method by incorporating both sequence and structure information. IEEE Trans Nanobiosci. 2015;14(4):339–49. https://doi.org/10.1109/TNB.2014.2352454.
22. Huang Y, Wuchty S, Zhou Y, Zhang ZD. Sgppi: structure-aware prediction of protein-protein interactions in rigorous conditions with graph convolutional network. Briefings Bioinf **24**(2) (2023) https://doi.org/10.1093/bib/bbad020
23. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, 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 AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583. https://doi.org/10.1038/s41586-021-03819-2.
24. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):439–44. https://doi.org/10.1093/nar/gkab1061.
25. Edgar RC, Batzoglou SJCoisb. Multiple sequence alignment 2006;16(3):368–373. https://doi.org/10.1186/1471-2105-5-113
26. Cohen T, Halfon M, Schneidman-Duhovny D. Nanonet: Rapid and accurate end-to-end nanobody modeling by deep learning at sub angstrom resolution. Front Immunol **13** (2022) https://doi.org/10.3389/fimmu.2022.958584
27. Jones DT, Thornton JM. The impact of alphafold2 one year on. Nat Methods. 2022;19(1):15–20. https://doi.org/10.1038/s41592-021-01365-3.
28. Abanades B, Georges G, Bujotzek A, Deane CM. Ablooper: fast accurate antibody cdr loop structure prediction with accuracy estimation. Bioinformatics. 2022;38(7):1877–80. https://doi.org/10.1093/bioinformatics/btac016.
29. Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns **3**(2) (2022) https://doi.org/10.1016/j.patter.2021.100406
30. Huang Z, Wang X, Wei Y, Huang L, Shi H, Liu W, Huang TS. Ccnet: Criss-cross attention for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2023;45(6):6896–908. https://doi.org/10.1109/TPAMI.2020.3007032.
31. Du ZY, Su H, Wang WK, Ye LS, Wei H, Peng ZL, Anishchenko I, Baker D, Yang JY. The trrosetta server for fast and accurate protein structure prediction. Nat Protoc. 2021;16(12):5634–51. https://doi.org/10.1038/S41596-021-00628-9.
32. Ruffolo JA, Chu L-S, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun. 2023;14(1):2389. https://doi.org/10.1038/s41467-023-38063-x.
33. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning (2021). arXiv preprint arXiv:2112.07782
34. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. Structure-based protein function prediction using graph convolutional networks. Nat Commun **12**(1) (2021) https://doi.org/10.1101/786236
35. Yuan QM, Chen JW, Zhao HY, Zhou YQ, Yang YD. Structure-aware protein-protein interaction site prediction using deep graph convolutional network. Bioinformatics. 2022;38(1):125–32. https://doi.org/10.1093/bioinformatics/btab643.
36. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks. Sci Rep **12**(1) (2022) https://doi.org/10.1038/s41598-022-12201-9
37. Demolombe V, Brevern AG, Felicori L, NGuyen C, Avila RA, Valera L, Jardin-Watelet B, Lavigne G, Lebreton A, Molina F et al. Pepop 2.0: new approaches to mimic non-continuous epitopes. BMC Bioinf 2019;20:1–14. https://doi.org/10.1186/s12859-019-2867-5
38. Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. Nat Commun **14**(1) (2023) https://doi.org/10.1038/s41467-023-43934-4
39. Dunbar J, Deane CM. Anarci: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. https://doi.org/10.1093/bioinformatics/btv552.
40. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi JY, Deane CM. Sabdab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):1140–6. https://doi.org/10.1093/nar/gkt1043.
41. Fu L, Niu B, Zhu Z, Wu S, Li W. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565.
42. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using clustalw and clustalx. Curr Protoc Bioinf. 2003;1:2–3. https://doi.org/10.1002/0471250953.bi0203s00.
43. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. Cov-abdab: the coronavirus antibody database. Bioinformatics. 2021;37(5):734–5. https://doi.org/10.1101/2020.05.15.077313.
44. Mitchell LS, Colwell LJ. Comparative analysis of nanobody sequence and structure data. Proteins: Structure, Function, and Bioinformatics 2018;86(7):697–706. https://doi.org/10.1002/prot.25497
45. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7. https://doi.org/10.1093/bioinformatics/btq134.
46. Rego NB, Xi E, Patel AJ. Identifying hydrophobic protein patches to inform protein interaction interfaces. Proc Natl Acad Sci. 2021;118(6):2018234118. https://doi.org/10.1073/pnas.2018234118.

47. Burbach SM, Briney B. Improving antibody language models with native pairing. Patterns 0, 100967 (2024) https://doi.org/10.1016/j.patter.2024.100967

Publisher's Note