The Protein Journal
https://doi.org/10.1007/s10930-023-10121-9

DeepBindPPI: Protein–Protein Binding Site Prediction Using Attention Based Graph Convolutional Network

Sharon Sunny[1] · Pebbeti Bhanu Prakash[2] · G. Gopakumar[1] · P. B. Jayaraj[1]

Accepted: 25 April 2023

Abstract

Due to the importance of protein-protein interactions in defence mechanism of living body, attempts were made to investigate its attributes, including, but not limited to, binding affinity, and binding region. Contemporary strategies for binding site prediction largely resort to deep learning techniques but turned out to be low precision models. As laboratory experiments for drug discovery tasks utilize this information, increased false positives devalue the computational methods. This emphasize the need to develop enhanced strategies. DeepBindPPI employs deep learning technique to predict the binding regions of proteins, particularly antigen–antibody interaction sites. The results obtained are applied in a docking environment to confirm their correctness. An integration of graph convolutional network with attention mechanism predicts interacting amino acids with improved precision. The model learns the determining factors in interaction from a general pool of proteins and is then fine-tuned using antigen–antibody data. Comparison of the proposed method with existing techniques shows that the developed model has comparable performance. The use of a separate spatial network clearly improved the precision of the proposed method from 0.4 to 0.5. An attempt to utilize the interface information for docking using the HDOCK server gives promising results, with high-quality structures appearing in the top10 ranks.

Keywords Protein–protein interaction · Binding site prediction · Antigen–antibody complexes · Protein–protein docking, graph convolutional network · Attention mechanism

1 Introduction

Protein–protein interactions are elemental in functioning of a host of biological systems. Any variance from their natural binding character can lead to diseases like Alzheimer's disease and Parkinson's disease. To our discredit, a cure for such diseases is still in the developmental stage, hence the demand for ardent efforts in protein-related research studies. The biological community started searching for alternatives to the precise experimental techniques as they are laborious, and expensive. Nowadays, computational methods are gaining popularity due to their high throughput and reduced cost. However, their results are not on par with experimental standards and thus require improvement. The studies around protein interactions mainly focus on the interactability of the proteins [1], identifying the regions of interaction [2, 3], identifying the partner-specific binding region [4], predicting the interacting pairs in given proteins [5], predicting the contact map or the expected structure of the complex generated during the interaction [6, 7].

Interactions between proteins are specific and are guided by the characteristics of amino acids, like polarity, in the binding region. In other terms, it is not the individual amino acids but the collective nature of amino acids in a region decides the probability of interaction. Apart from the physicochemical characteristics, the shape and size of the binding region also affect interaction. Computational docking is a technique to predict the structure of a complex formed by the interaction between two proteins. Its results can be

✉ Sharon Sunny
ssharon099@gmail.com

Pebbeti Bhanu Prakash
bhanuprakash_b190607ec@nitc.ac.in

G. Gopakumar
gopakumarg@nitc.ac.in

P. B. Jayaraj
jayarajpb@nitc.ac.in

[1] Department of CSE, National Institute of Technology, Calicut, Kerala 673601, India

[2] Department of ECE, National Institute of Technology, Calicut, Kerala 673601, India

Published online: 18 May 2023

utilized to study the physicochemical properties and biological characteristics of the newly generated protein complex. Despite the introduction of many docking methods, the problem remains unsolved.

The present era witnessed a surge in protein-related studies, using both experimental and computational techniques especially deep learning, which was triggered by the outbreak of the pandemic Covid 19 [8–10]. One notable advantage of automated techniques is the speed with which they can model rapid mutations to the external protein (antigen). Though computational methods are improving, none of them currently can replace experimental techniques as accurate and precise solutions are not assured to date.

Antigen–antibody interactions are a particular case of molecular interactions that happen as an immune response by the living body. Antigens are usually heavy-weight molecules that are proteins, nucleic acids, lipids, or polysaccharides, while antibodies are biological proteins whose production is triggered by an antigen's presence. When antigen–antibody interaction occurs, their binding regions are called epitope and paratope, respectively, and are chosen in such a way as to neutralize the effect of antigen. As it is known, paratope predictions are relatively easy since they mainly occur in the tip of the Y-shaped antibody [11]. Information on the paratope-epitope region can assist in drug discovery tasks [12, 13].

A surge in the volume of sequence data triggered the broad application of deep learning in protein-related studies; the addition of relevant structural details supplements the developed models. As a classification task, protein–protein binding site prediction can also be solved using such a model. Due to the importance of binding sites in drug design and to accommodate the rapid mutations that can happen, conventional techniques pave way for specialized deep learning models. Some of the existing works address the prediction as a partner independent task. GraphPPI proposed by Yuan et al. [14] uses graph convolution layers to predict the probability of interaction of an amino acid in a protein. It uses both sequential and structural features to make accurate classification. The method proved to have better performance than other existing methods. But, it does not consider partner specific information for prediction.

EpiPred [15], proposed by Krawczyk et al. to predict B-cell epitope, searches for matching regions on the interface to make correct predictions. It takes advantage of the existing knowledge on antigen–antibody complexes to get an improved performance. As the method uses partner information, the results are more specific and better than partner independent prediction methods. Lin et al. [16] proposed a Support Vector Machine (SVM) based method that uses the features, like evolutionary information and amino acid propensity scale, for the same purpose. LIBSVM, a package to easily implement SVM, with radial basis function as the kernel is used in the proposed method. Though the model generalize well for the linear epitope, no par excellence performance can be expected in case of conformational epitopes since it is not trained for the same.

Lebris et al. [17] proposed Parapred, a sequence based method for paratope prediction that exploits the capabilities of convolutional neural network and recurrent neural network to carry out its intended task. It incorporates exponential linear units activation function and residual connections to better the prediction capability. The applicability of the prediction in docking is checked by associating with PatchDock and has promising results. The proposed deep learning network can be easily extended to process structural features by altering the layers and expanding the training set with the corresponding modifications.

Lo et al. [18] proposed a two fold procedure that uses two sequential modules for matching and prediction of epitope. Matching involves the application of BLAST to find identical antigen sequences and identification of high scoring epitope sequences. It may also be implemented as spiral feature vector search. Prediction may be implemented on the basis of knowledge-based energy or by comparing the spiral vectors of patches with known conformational epitopes. The method shows superior performance but does not consider partner specific information.

Vecchio et al. [11] describes the design of models for epitope and paratope predictions. The combination of Convolutional Neural Network (CNN) and Message Passing Neural Network (MPNN) [19] models employed along with the final linear and sigmoid layers identifies paratope using antibody Fv sequence and Fv graph. Epitope predictor, on the other hand, uses two separate Graph Convolutional Networks (GCN) to encode the features of antibody and antigen relevant for its counterpart. The work emphasizes that the interaction should be treated as asymmetric.

Daberdaku et al. [20] proposed a method that exploits the ability of 3D Zernike Descriptors (ZD) [21] to efficiently represent the geometric as well as the physicochemical characteristics of proteins to predict the interacting amino acids in antibodies. It utilized the undersampling of the majority class and oversampling of the minority class to deal with the data imbalance problem. The application of an SVM classifier integrated with an Isolation Forest (IF) [22] algorithm correctly identifies the interacting regions with continuity. IF algorithm avoids the isolated amino acids from the interface region as the chance of such isolated interaction is negligible in antibody interactions. The method leaves out partner-specific information.

Oh et al. [23] attempt to find the hotspot on the antigen surface through which it binds to a potential antibody. It uses GCN layers with pooling, upsampling, and attention mechanisms to identify partner-specific epitope regions. While the pooling layers help to get a coarse-grained representation of the input, upsampling layer ensures the required dimension at the output. The attention layer confirms that the partner-dependent information is conveyed appropriately and accurately to determine the interacting amino acids. The method includes a clustering step to consolidate the probable regions of binding than identifying probable residues involved in antigen–antibody complex formation. The model is trained using the same dataset as in [15].

Lu et al. [24] use CNN to capture neighborhood information of sequential neighbors, and the result is fed to GCN to compute the spatial neighborhood of amino acids. An attention layer is introduced to get the partner information of the antibody. The final fully connected layer performs the task of classifying amino acids as interacting or non-interacting. The authors observe that adding more layers to the network may not assure an improvement in results. Since the model is trained with a paratope prediction dataset, it performs better than many state-of-the-art paratope prediction methods.

Dai et al. [25] approach the interaction prediction problem as a segmentation task where the protein surface is segmented as interacting and non-interacting regions. The method learns the geometric features of the input proteins using PointNet [26] modules. The Spatial Transformer Network [27] ensures that the input is transformation invariant. The network extracts local and global features from the input to make predictions. Despite being trained on protein–protein interactions rather than antigen–antibody complexes, the proposed Protein Interface Network (PInet) could achieve state-of-the-art performance in epitope-paratope identification, validating the model's generalizability.

Pittala et al. [28] proposes a transfer learning approach in antigen–antibody interaction prediction, where a combination of convolution layers, graph convolution layers, and attention layers achieves the task. Convolution and graph convolution layers deal with the spatial proximity of residues, while attention layers extract context information. The authors observe that including residual features can improve the performance of the model.

Many existing methods do not consider the partner protein in binding site prediction, especially epitope and paratope prediction, and hence the obtained results may not be specific. Also, the number of false positives in their predictions is too high that it may depreciate the utility of computational methods. To our knowledge, none of the existing methods has given special attention to spatial features which is crucial in case of antigen–antibody complexes. This work attempts to predict binding region, with special attention given to spatial features in addition to the sequential features. A blend of graph convolution with attention mechanism is conducive to identifying the interface. The model trained on a general set of proteins (base model) is enhanced for epitope/paratope prediction by finetuning with antigen–antibody complexes (transfer learning model).

2 Materials and Methods

The proposed deep learning network is first trained using a general set of protein–protein complexes. Now, this base model has learned the characteristic features of amino acids in the interface region of proteins. Existing studies show that using a pre-trained model can help when the number of available training data is less [29]. Thus, in our method, the weights obtained by the base model are fine-tuned by a second round of training using antigen–antibody complexes. The use of such a pre-trained model helps to deal with the limited data availability in the case of epitope-paratope prediction. The predicted binding site information is used to generate protein–protein complex structures using HDOCK server. Workflow of the proposed method is given in Fig. 1 and the description of the approach follows.

2.1 Dataset

The proposed method uses targets from different datasets for training and testing the deep learning model. It includes experimentally predicted structures and structures obtained by homology modeling; the details of these datasets follows:

Protein–Protein Docking Benchmark set (DB5.5): The latest docking benchmark set 5.5 [30] is used to train and test the model. It contains 257 complexes, of which 88

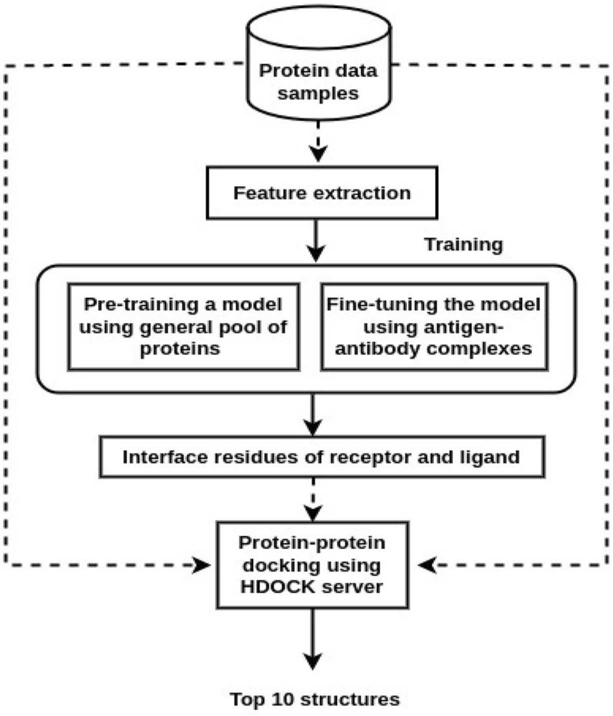
Protein data samples
Feature extraction
Training
Pre-training a model using general pool of proteins
Fine-tuning the model using antigen-antibody complexes
Interface residues of receptor and ligand
Protein-protein docking using HDOCK server
Top 10 structures

Fig. 1 The proposed pipeline for binding site identification and generation of protein–protein complexes

are enzyme-inhibitor complexes, 67 are antigen–antibody complexes, and 102 are other complexes. Antigen–antibody complexes in common with epitope and paratope prediction datasets, which are mentioned later in this section, are removed from the set. DB5.5 is non-redundant according to Structural Classification of Proteins (SCOP).

PPI4Dock [31]: This dataset of heterodimeric models contains structures obtained through homology modeling. Out of the total 1417 structures, no two samples have more than 70% sequence identity. Among these, 138 targets are antigen–antibody complexes. A filtering step ensures that the intersection of samples in the PPI4Dock dataset and epitope-paratope prediction sets is null.

PRISM [32]: This dataset is generated from structures in the protein data bank after applying different levels of filtering. The dimers in Protein Data Bank (PDB) are analyzed and clustered based on interfacial similarity. From the retained non-redundant complexes, a random subset of targets is selected to train the model. Care is taken to avoid samples from the epitope-paratope prediction set.

SAbDAb [33]: Samples from this dataset are used for epitope prediction. Complexes having antibody identity greater than 99% and corresponding antigen greater than 90% are eliminated. All the selected complexes have at least 3Å resolution. The filtered sample set from the dataset contains 148 non-redundant complexes. More details on this dataset can be found in [15].

Dataset for paratope prediction: Samples from [20] are widely used in checking the potential of paratope prediction methods. This set of 417 samples, which is categorically divided as train, validation, and test data, is filtered to guarantee that the sequence identity of included samples is less than 95%.

2.2 Preprocessing

Data from different datasets are carefully chosen so as to avoid identical sequences. These data are preprocessed to extract residue-level features before feeding them to the model. Each amino acid in the input is represented using the Cα atom for ease of computation. Input having sequence length less than 80 or greater than 999 are eliminated to avoid too short or long samples in the sample set. The following features are extracted from the input data:

- *3D coordinates*: The XYZ coordinates of Cα atoms are collected, which is pivotal in the geometric feature extraction of proteins.
- *Depth of the residue*: The deeper an amino acid is, the lesser is its interaction probability. The proposed method calculates the minimum depth of each amino acid from the surface of a protein using the tool MSMS [34]. It fails to generate the surface for some proteins, and such samples are removed from the dataset. Also, the tool can work for proteins containing at most 1000 residues. Hence larger inputs are also removed.

Table 1 Kyte–Doolittle scale

Amino acid	Kyte–Doolittle	Amino acid	Kyte–Doolittle
Alanine	1.8	Cysteine	2.5
Aspartic acid	− 3.5	Glutamic acid	− 3.5
Phenylalanine	2.8	Glycine	− 0.4
Histidine	− 3.2	Isoleucine	4.5
Lysine	− 3.9	Leucine	3.8
Methionine	1.9	Asparagine	− 3.5
Proline	− 1.6	Glutamine	− 3.5
Arginine	− 4.5	Serine	− 0.8
Threonine	− 0.7	Valine	4.2
Tryptophan	− 0.9	Tyrosine	− 1.3

Table 2 List of input features and their lengths

Feature	Length
3D coordinates	3
Depth of a residue	1
Hydrophobicity	1
Polarity	1
Percentage of amino acids in the neighborhood	20
One-hot encoded amino acid	20
Receptor/ligand flag	1
PSSM	20
Total length of the feature vector	67

- *Hydrophobicity*: It is calculated using the Kyte–Doolittle (KD) scale [35], which is summarised in Table 1. A positive KD value indicates that the residue is hydrophobic. Hydrophobic residues tend to get buried away from water and may occupy interface regions [36].
- *Polarity*: Polar amino acids show high propensity to interact than non-polar ones, with glycine as an exception.
- *Percentage of amino acids in k nearest neighborhood*: Neighborhood information of amino acids is extracted as the collective nature of residues in a region influences the probability of interaction.
- *One-hot encoded amino acid*: The constituent amino acids in a protein are one-hot encoded.
- *Receptor/ ligand flag*: A flag is set to distinguish between receptor and ligand features.
- *PSSM*: Position Specific Scoring Matrix of the input protein sequence is calculated using PSI-BLAST.

Table 2 lists all the features and their lengths. Input to the model will be of shape Lx67, where L is the number of amino acids which varies with the sample and 67 is the length of feature vector. Amino acids that appear in the interface of the complex are identified based on their proximity to the partner protein's amino acids. A threshold of 10Å is set as only Cα atoms are considered.

It is ensured that there is no intersection of samples in train and test sets. Input proteins' features are fed separately to the model. Since a protein can be easily represented as a graph, the same is used to represent the input data. Each node in the graph represents an amino acid and hence contains the residue level features. Intraprotein edges are calculated based on the distance between their Cα atoms. The labels are checked to ensure that both interacting and non-interacting amino acids are present in the sample. Proteins with single class labels are removed from the dataset.

2.3 The Proposed Architecture for Interface Prediction

The architecture of the proposed deep learning model for interface prediction is shown in Fig. 2. In the figure, X_r and X_l are receptor and ligand features, respectively, and Y is the classification model's output, representing whether an amino acid in the protein–protein complex is interacting with a partner or not. The method uses a graph convolutional network and attention module to obtain the prediction.

2.3.1 Graph Convolutional Network Layer

When the relations between entities are crucial in decision making, a categorical solution would be to use graph neural networks. In docking, features of spatially neighboring nodes affect the value of a particular node. Equally important is its node features. Hence graph convolution operation on a graph with a self-loop could be a natural choice. A basic GCN layer uses Eq. (1) to calculate the node features at layer *l*.

$$h^{l} = \sigma\left(D^{-1/2}\,\tilde{A}\,D^{-1/2}\,h^{l-1}\,W^{l}\right) \tag{1}$$

σ is the nonlinear activation function that filters out its input. Here, the adjacency matrix is modified to include self loops, i.e. $\tilde{A} = A + I_n$ where A is the original adjacency matrix, I_n is the identity matrix and n is the number of nodes in the input

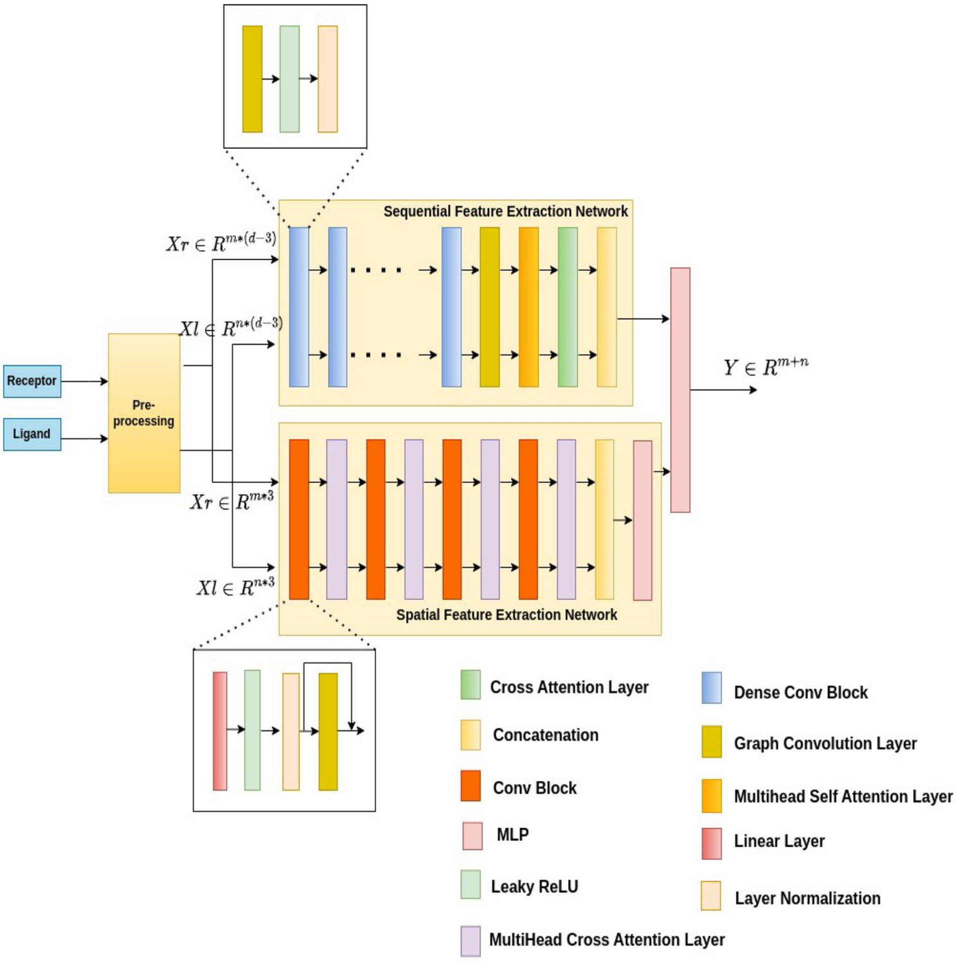

Fig. 2 The proposed network architecture for binding site prediction

graph. Degree matrix D is calculated as $D_{\mathrm{ii}} = \sum_j A_{\mathrm{ij}}$. The inverse of degree matrix is the normalized degree matrix which is calculated to avoid prioritizing nodes having high degrees. In the proposed method, each graph convolutional layer l calculates a result as given in Eq. (2).

$$a^l = LeakyReLU\left(D^{-1/2}\ \tilde{A}\ D^{-1/2}\ h^l\ W^l\ \right) \tag{2}$$

2.3.2 Multihead Attention Layer

A basic attention [37] module is shown in Fig. 3. Multihead attention essentially considers the different aspects of a feature; each head treats it differently. Let key k, value v, and query q be input to this attention layer. Latent representations of all these values are first computed as given in Eqs. (3), (4), and (5).

$$\tilde{q} = W_{\mathrm{q}}\ q + b_{\mathrm{q}} \tag{3}$$

$$\tilde{k} = W_{\mathrm{k}}\ k + b_{\mathrm{k}} \tag{4}$$

$$\tilde{v} = W_{\mathrm{v}}\ v + b_{\mathrm{v}} \tag{5}$$

Here, Ws are the weights and bs are the biases. Then the attention scores are calculated using Eq. (6).

$$Attention\_Score = Softmax\left(\frac{\tilde{q}\tilde{k}^T}{\sqrt{d}}\right)\tilde{v} \tag{6}$$

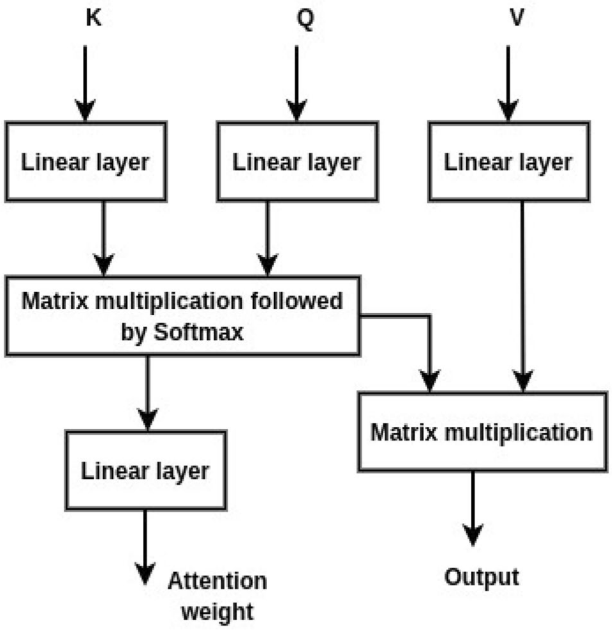

Fig. 3 Diagram of a basic attention module

where d is the number of heads. Equation (6) calculates the attention score of a single head. Similar calculations are carried out for each head before combining them to a final score. A mapping of merged result to latent vector follows.

2.3.3 Cross-Attention Layer

The cross-attention layer allows the exchange of information between ligand and receptor features. Let R and L be receptor and ligand features, respectively. The attention score for a receptor calculated using Eq. (10) apply the values obtained in Eqs. (7), (8), and (9). Cross attention score of ligand is calculated in a similar way.

$$Q = LeakyReLU(W_{\mathrm{qr}}\ R + b_{\mathrm{qr}}) \tag{7}$$

$$K = LeakyReLU(W_{\mathrm{kr}}\ L + b_{\mathrm{kr}}) \tag{8}$$

$$V = LeakyReLU(W_{\mathrm{vr}}\ L + b_{\mathrm{vr}}) \tag{9}$$

$$Cross\_Attention\_Score = Softmax(\ Q\ *\ K^T\ )\ *\ V\ +\ R \tag{10}$$

2.3.4 Loss Function

Protein–protein interface prediction can be mapped as a binary classification task, and thus the last layer should output the interaction probability of amino acids. Here, the interaction threshold is taken as 0.5. The proposed method uses a binary cross-entropy loss with logits function as it offers better numerical stability than using a sigmoid layer followed by a binary cross entropy loss. Since the interaction data is imbalanced, a weightage is assigned to the positive class. This work uses a constant weight of 10 for the positive labels.

2.4 Implementation

The proposed model is implemented using Pytorch Lightning on an NVIDIA-DGX-A100 station with 4x40GB Graphics RAM. A description of the model (shown in Fig. 2) follows:

Spatial feature extraction network: The coordinates extracted from the receptor and ligand are used in extracting the geometric features. Let us consider the case of the receptor first. Its features are treated in a pipeline containing a linear layer, a non-linear activation function, and layer normalization to generate a latent vector representation. The geometric features may have neighborhood dependency; graph convolutional layers are employed to deal with this. The inclusion of residual connection adds information from the activation function to the results of the graph convolutional layer. The same network transforms ligand features to extract hidden geometric features required for interaction prediction. Now cross-attention features of both inputs are calculated. Four such blocks are stacked in a spatial network. The spatial features of both inputs thus extracted are concatenated, and the same is passed to an MLP (Multi-Layer Perceptron).

Sequential feature extraction network: The model processes the ligand and receptor features separately before concatenating them in the last layers. All features but the coordinates of individual proteins are initially fed through a graph convolutional network. As the input size varies, the batch size is taken as one. In such a case, the variance would be zero, nullifying the effect of batch normalization. Therefore, layer normalization is used to normalize the result of the LeakyReLU activation function. A chain of six such blocks is utilized to extract the necessary features from sequence data. The result is then passed through the multi-head attention layer as query, key, and value to prioritize essential features. Receptor and ligand features are calculated using the above network. Now cross-attention is applied to both features and is concatenated to get the features of the complex.

Features obtained at this stage are further explored using a stack of fully connected layers and LeakyRELU to calculate the interaction probability. A threshold of 0.5, used in existing systems, is chosen to distinguish interacting from non-interacting amino acids. The network is trained for 500 epochs with a learning rate of 0.0001. Adam optimizer adjusts the model parameters to obtain optimal feature values. The next phase of training is specific to antigen–antibody complexes. The model parameters learned through 500 epochs of training are fine-tuned using the data in [15, 20]. The pre-trained model is fine-tuned for 150 epochs using this data.

3 Docking Based on Binding Site Information

The binding site details of the receptor and ligand can be effectively utilized to predict the protein–protein complex structure. The proposed pipeline uses the HDOCK [38] server that uses a hybrid method (template-based and ab initio methods) to generate plausible complex structures. It requires the individual receptor and ligand structures and the residue number with the corresponding chain of amino acids in the binding site region as input data. It should be ensured that the residue number in the binding site and input pdb file should match to get the desired output. The generated top 10 predictions are retrieved from the HDOCK server.

4 Results and Analysis

An analysis of the proposed pipeline in binding site prediction (using the base model) and the ability of the transfer learning model in epitope-paratope prediction is carried out. Potential of the identified residues in comlex formation is also investigated using the HDOCK server.

Figure 4 shows the range of F1-score, Matthews Correlation Coefficient (MCC), and AUC-ROC values obtained by the base model when tested using a sample set of targets in DB5.5. Clearly, the base model trained on general protein–protein complexes learned the interface's features and is successful for more than half of the targets; a median value greater than 0.5 for F1-score, MCC, and AUC-ROC illustrates this. Also, the reported AUC-ROC values show that the model can predict interacting and non-interacting amino acids.

Obtained binding site information is utilized to predict the complex structure using the HDOCK server. Best results obtained for the targets 1PXV and 2O8V in top 10 predictions are shown in Fig. 5. There is a close resemblance between the predicted structure and crystal structure of these samples. As shown in Fig. 6, the method could successfully predict complex structures for 31.707% of 41 samples used. It is clear from the graph that, in the top 10, there is at least one good (high or medium) quality structure in prediction.

4.1 Comparison of the Proposed Method with State of the Art Techniques for Binding Site Prediction

This section compares the proposed base model with existing techniques like BIPSPI [39] and PInet [25] using samples in DB5.5. PInet approaches the protein interaction prediction problem as a segmentation task where the protein

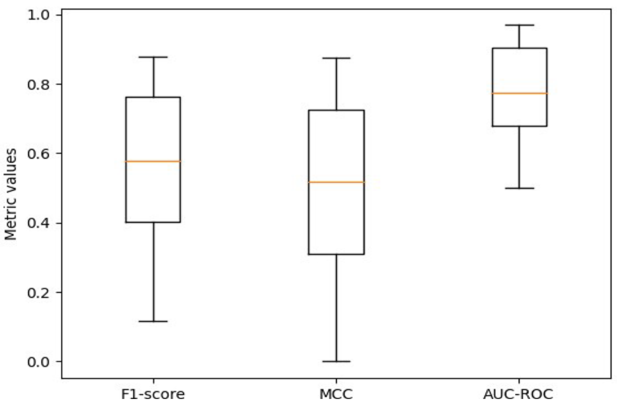

Fig. 4 Performance metric values for the proposed base model on a set of samples in DB5.5

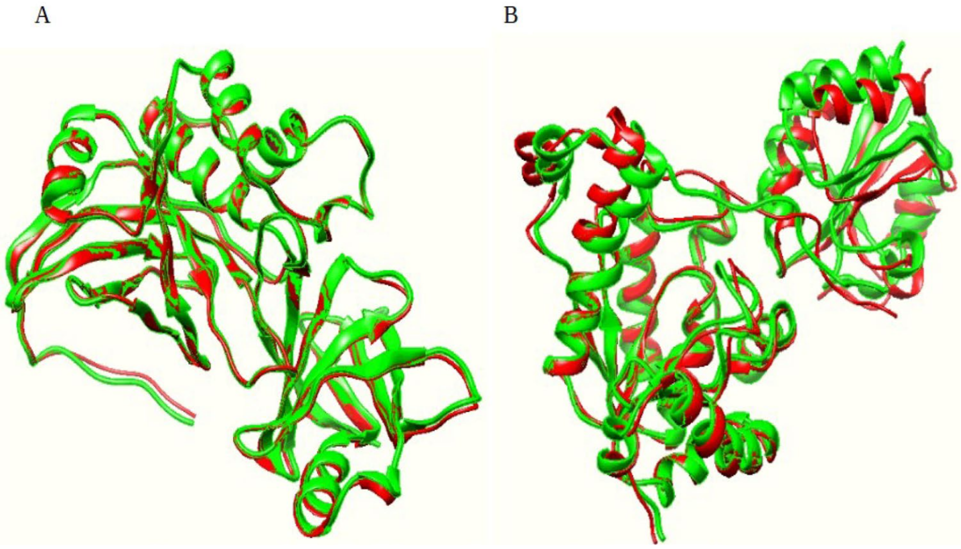

Fig. 5 Output obtained by HDOCK server for the targets, 1pxv, and 2o8v. Ribbon like structure shown in green is the original crystal structure and in red is the predicted structure (Color figure online)

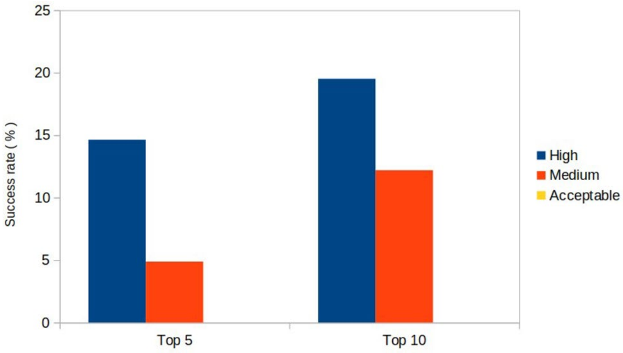

Fig. 6 Performance of the proposed method in generating the protein–protein complex structures for samples in DB5.5. Success rate in top 5 and top 10 ranks is depicted in this Figure

Table 3 Comparison of proposed method for binding site prediction with state of the art techniques based on their performance on DB5.5 samples. 'Aug 50' shows the augmentation of training data

Struct	Method	Precision	Recall	AUC-ROC
U	Proposed base model (without spatial network)	0.401	0.546	0.700
U	Proposed base model	**0.500**	0.598	0.753
U	PInet (Aug 50)	0.492	**0.723**	0.753
U	BIPSPI	0.391	0.558	**0.822**
B	PInet	**0.511**	0.749	**0.837**
B	BIPSPI	0.394	0.599	0.827
B	Proposed base model (without spatial network)	0.462	**0.829**	0.824

Best performance values obtained are given in bold

surface is segmented as interacting and non-interacting regions. The network extracts local and global features from the input and uses PointNet and spatial transformer network to make the predictions. BIPSPI uses XGBoost in identifying the interfacial residues.

Table 3 compares the proposed base model with existing techniques using samples in DB5.5. The values are on par with the existing systems. It is clear from the values that the spatial network improved the prediction capability of the model. In the table, column 'struct' represents whether the input structures are in bound or unbound form. The proposed model without the spatial network has improved the recall value for bound structures. The better results for PInet is attributed to the augmentation of training data.

We know that antigen–antibody interactions have a significant role in the development of vaccines. But the number of such complexes in different datasets is very small. Hence, to get improved performance for antigen–antibody interface prediction, a transfer learning technique is adopted in the proposed method. Comparison of the proposed transfer learning model with state-of-the-art techniques like PECAN [28], PInet [25], EpiPred [15], and DiscoTope [40] based on its performance in predicting the correct binding sites for samples in [15] is shown in Table 4. The reported values of the proposed method are the average of 5-fold cross-validation. When tested for epitope prediction, the fine-tuned model performs better than the base model.

Table 5 compares the transfer learning model with state-of-the-art paratope prediction technique. The highlight of the proposed transfer learning model is that it

Table 4 Comparison of the proposed method with state-of-the art techniques based on their performance on the epitope prediction dataset

Method	Precision	Recall	AUC-ROC
Proposed method (Base model)	0.105	0.198	0.523
Proposed method (Transfer learning)	**0.315**	0.413	**0.665**
EpiPred	0.136	0.436	NA
DiscoTope	0.214	0.110	NA
PECAN	0.157	0.730	0.655
PInet	0.181	**0.931**	0.654

Best performance values obtained are given in bold

Table 5 Comparison of the proposed method with state-of-the art techniques based on their performance on the paratope prediction dataset

Method	Precision	Recall	AUC-ROC
Proposed transfer learning model	**0.60**	0.75	0.84
PECAN (Conv1 Layer + Attntn)	0.48	0.89	**0.95**
PECAN (Conv2 Layer)	0.36	**0.95**	**0.95**

predicts comparatively fewer false positives and hence improves precision.

5 Discussion

A deep learning model requires appropriate training using relevant data. When tested for epitope prediction, the base model reported a precision of 0.105, recall of 0.198, and AUC-ROC of 0.523. These values point to the fact that it seldom encountered an antigen–antibody complex during its training. Expecting a deep learning model to predict an unseen distribution is abortive. A pre-trained model helps at this juncture. The model trained on general pool of protein–protein complexes shows improved performance once calibrated exclusively for antigen–antibody complexes; this is evident from the results reported in Table 4.

An attempt to check the importance of the spatial network in the proposed method is carried out by feeding the whole set of features to the sequence network. The model obtained after 500 epochs has a precision of 0.401, recall of 0.546, and AU-ROC of 0.700 for the test samples in DB5.5. It is clear from Table 3 that adding a spatial network improves the prediction.

To deal with the class imbalance problem in data, the proposed method adopts a scheme that imposes an increased penalty for the misclassification of the positive class. Among the different weights (5, 10, 15, and 20) tried, 10 is found as the best weight as it guarantees better results.

It is known that a classification model ideally generate fewer false positives and zero false negatives. A higher cut-off yields high precision, while lower cut-off favors recall value [41]. A considerable gap in precision and recall values indicates several false positives in prediction. Table 4 shows that the proposed method includes fewer false positives in its prediction compared to other techniques. An attempt to predict protein–protein complex structure is relatively successful as the HDOCK server could make a wise decision based on the available true positives. The plot showing the quality of generated structures against the number of false positives and false negatives for a set of samples in DB5.5 is given in Fig. 7. For the sake of representation, the quality classes high, medium, acceptable, and incorrect are mapped to floating-point numbers 0.3, 0.2, 0.1, and 0.0, respectively. An increased number of false positives or false negatives can degrade structure prediction quality.

A docking tool that can work on all types of proteins would be appreciable. Hence the proposed base model is used to predict the structure of complexes in a sample test set from DB5.5. The quality of the best results obtained in the top10 are shown in Fig. 6. The quality of successfully generated structures is either high or medium, which implies the competence of the proposed method.

Figure 8 shows a comparison of crystal structure and predicted structure of samples for which the proposed method could successfully generate plausible structures. In

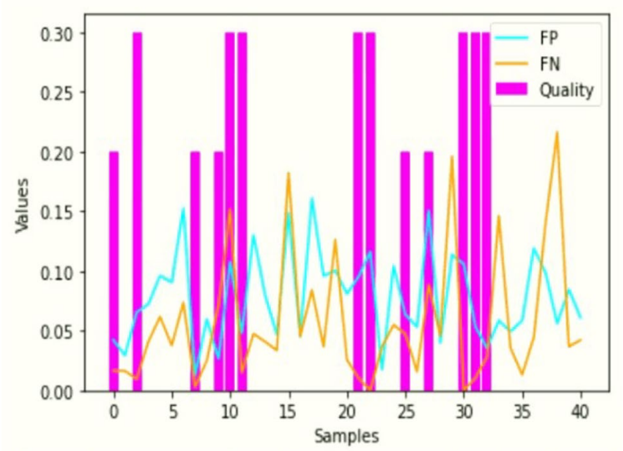

Fig. 7 The quality of generated structures against the number of false positives and false negatives for the targets in DB5.5. The values on the y-axis shows the false positives and false negatives in the predicted binding sites. It also corresponds to different quality classes. ’High’ quality is represented using 0.3, ’medium’ quality as 0.2 and ’acceptable’ quality as 0.1

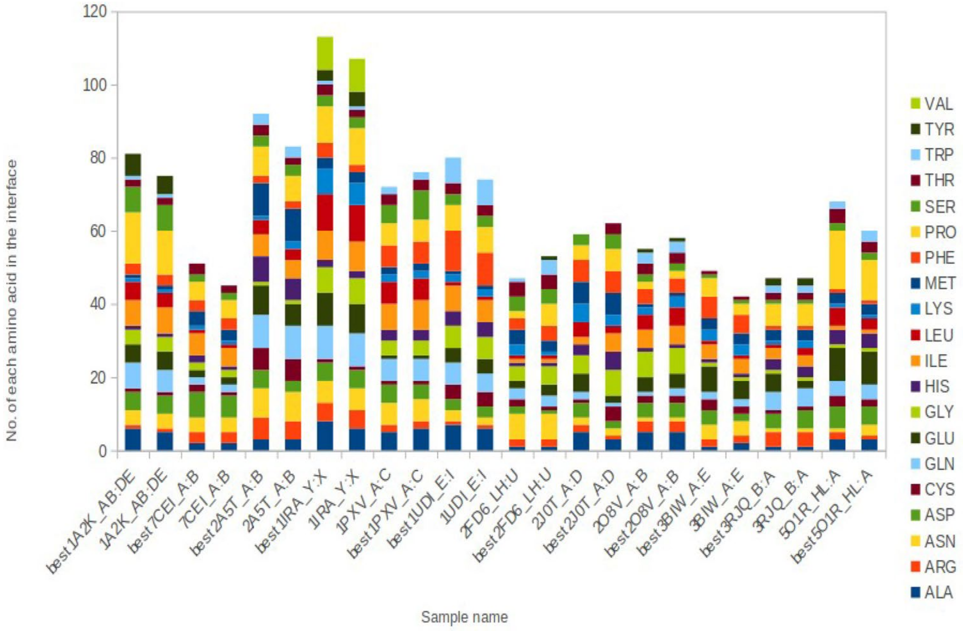

Fig. 8 Comparison of predicted structure and crystal structure of a set of samples in DB5.5 based on the number of each amino acid in the interface

the figure, the bar corresponding to the name prefixed with ’best’ shows the predicted result, and the other corresponds to the crystal structure. The same figure illustrates the size of the interface and the count of each amino acid in the interfacial region. The results show that the prediction is similar to the original interface region and hence we can infer that the method can efficiently predict protein complex structures.

6 Conclusion

The difference in characteristics of amino acids in the interface region is crucial, and it makes the attempts to predict the behavior of proteins hard. Computational docking gets its popularity as the structure and function of protein are correlated. The proposed method solves the docking problem in two stages: identifying the partner-specific binding sites using a deep learning model and generating the complex structure using the HDOCK server by utilizing the predicted interface region. Since a graph in its capacity can skillfully represent the spatial relations between amino acids, a graph convolutional network is employed in the proposed deep learning model. The integration of attention mechanism enhances the important features in hidden layer representations. Sequential and spacial features of proteins are captured separately, and this enhances the performance of the model. The proposed method can assist experimental drug discovery because of the reduced number of false positives in prediction. The method is analyzed in terms of its correctness in binding site prediction and structure prediction. Results are promising and may be improved further by refining the deep learning model. The proposed method is trained using handpicked features extracted from training data. Employing an autoencoder network to generate efficient feature representation may improve the performance of this model. The application of self-supervision may contribute to the improvement of the deep learning model’s performance. Introducing a second deep learning network for structure prediction may also better the results.

Acknowledgements The authors gratefully recognize the Department of Computer Science, NIT Calicut for providing infrastructure facilities. The authors thank the Centre for Computational Modelling and Simulation (CCMS) and Central Computer Centre (CCC) at NITC for the NVIDIA DGX station facility. Special thanks to Mr. Sai Manoj Akondi and Ms. Seenia Francis for their support.

Author Contributions SS: Conceptualization, design, implementation, manuscript preparation and editing PBP: Conceptualization, implementation GG.: Conceptualization, manuscript review JPB: Manuscript review, supervision

Funding The authors have no relevant financial or non-financial interests to disclose.

Data Availability Not applicable

Code Availability Source code of DeepBindPPI is available at: https://github.com/Sharon1989Sunny/DBP

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Approval Not applicable

Consent to Participate Not applicable

Consent for PÑublication The manuscript is approved by all the authors for publication.

References

1. Singh A, Kumar A, Uversky VN, Giri R (2018) Understanding the interactability of chikungunya virus proteins via molecular recognition feature analysis. RSC Adv 8(48):27293–27303
2. Liberis E, Veličković P, Sormanni P, Vendruscolo M, Lió P (2018) Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics 34(17):2944–2950
3. Tubiana J, Schneidman-Duhovny D, Wolfson HJ (2022) Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods 4:1–103
4. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V (2012) Predicting protein–protein interface residues using local surface structural similarity. BMC Bioinform 13(1):1–14
5. Liu J, Gong X (2019) Attention mechanism enhanced lstm with residual architecture and its application for protein–protein interaction residue pairs prediction. BMC Bioinform 20(1):1–11
6. Zeng H, Wang S, Zhou T, Zhao F, Li X, Wu Q, Xu J (2018) Complexcontact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res. 46(W1):432–437
7. Quadir F, Roy RS, Halfmann R, Cheng J (2021) Dncon2_inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning. Sci. Rep. 11(1):1–10
8. Ong E, Wong MU, Huffman A, He Y (2020) Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning. Front Immunol 11:1581
9. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, Yu L, Ni Q, Chen Y, Su J et al (2020) A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering 6(10):1122–1129
10. La Gatta V, Moscato V, Postiglione M, Sperli G (2020) An epidemiological neural network exploiting dynamic graph structured data applied to the covid-19 outbreak. IEEE Trans Big Data 7(1):45–55
11. Vecchio A, Deac A, Liò P, Veličković P (2021) Neural message passing for joint paratope-epitope prediction
12. Zhang MM, Huang RY-C, Beno BR, Deyanova EG, Li J, Chen G, Gross ML (2020) Epitope and paratope mapping of pd-1/nivolumab by mass spectrometry-based hydrogen-deuterium exchange, cross-linking, and molecular docking. Anal Chem 92(13):9086–9094
13. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH et al (2021) A compact vocabulary of paratope–epitope interactions enables predictability of antibody–antigen binding. Cell Rep 34(11):108856
14. Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y (2022) Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics 38(1):125–132
15. Krawczyk K, Liu X, Baker T, Shi J, Deane CM (2014) Improving b-cell epitope prediction and its application to global antibody–antigen docking. Bioinformatics 30(16):2288–2294
16. Lin SY-H, Cheng C-W, Su EC-Y (2013) Prediction of b-cell epitopes using evolutionary information and propensity scales. In: BMC Bioinformatics, vol. 14, pp. 1–9. BioMed Central
17. Liberis E, Veličković P, Sormanni P, Vendruscolo M, Liò P (2018) Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics 34(17):2944–2950
18. Lo Y-T, Shih T-C, Pai T-W, Ho L-P, Wu J-L, Chou H-Y (2021) Conformational epitope matching and prediction based on protein surface spiral features. BMC Genom 22(2):1–16
19. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: International Conference on Machine Learning, pp. 1263–1272. PMLR
20. Daberdaku S, Ferrari C (2019) Antibody interface prediction with 3d zernike descriptors and svm. Bioinformatics 35(11):1870–1876
21. Novotni M, Klein R (2003) 3d zernike descriptors for content based shape retrieval. In: Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications, pp. 216–225
22. Liu FT, Ting KM, Zhou Z-H (2008) Isolation forest. In: 2008 Eighth IEEE international conference on data mining, pp. 413–422. IEEE
23. Oh L, Dai B, Bailey-Kellogg C (2021) A multi-resolution graph convolution network for contiguous epitope prediction. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1–10
24. Lu S, Li Y, Wang F, Nan X, Zhang S (2021) Leveraging sequential and spatial neighbors information by using cnns linked with gcns for paratope prediction. IEEE/ACM Trans Comput Biol Bioinform 19(1):68–74
25. Dai B, Bailey-Kellogg C (2021) Protein interaction interface region prediction by geometric deep learning. Bioinformatics 37(17):2580–2588
26. Qi CR, Su H, Mo K, Guibas LJ (2016) Pointnet: Deep learning on point sets for 3d classification and segmentation. arXiv preprint arXiv:1612.00593
27. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. Adv Neural Inform Process Syst 28
28. Pittala S, Bailey-Kellogg C (2020) Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics 36(13):3996–4003
29. Tran M, Soleymani M (2022) A pre-trained audio-visual transformer for emotion recognition. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4698–4702. IEEE
30. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J et al (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427(19):3031–3041
31. Yu J, Guerois R (2016) Ppi4dock: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. Bioinformatics 32(24):3760–3767
32. Cukuroglu E, Gursoy A, Nussinov R, Keskin O (2014) Non-redundant unique interface structures as templates for modeling protein interactions. PLoS ONE 9(1):86738
33. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM (2014) SAbDab: the structural antibody database. Nucleic Acids Res 42(D1):1140–1146
34. Sanner MF, Olson AJ, Spehner J-C (1996) Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38(3):305–320
35. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
36. Xie Z, Deng X, Shu K (2020) Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. Int J Mol Sci 21(2):467
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst 30

38. Yan Y, Zhang D, Zhou P, Li B, Huang S-Y (2017) Hdock: a web server for protein-protein and protein-dna/rna docking based on a hybrid strategy. Nucleic Acids Res 45(W1):365–373
39. Sanchez-Garcia R, Sorzano COS, Carazo JM, Segura J (2019) Bipspi: a method for the prediction of partner-specific protein–protein interfaces. Bioinformatics 35(3):470–477
40. Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous b-cell epitopes using protein 3d structures. Protein Sci 15(11):2558–2567
41. Krawczyk K, Baker T, Shi J, Deane CM (2013) Antibody i-patch prediction of the antibody binding site improves rigid local antibody–antigen docking. Protein Eng Des Select 26(10):621–629