frontiers | Frontiers in Immunology

METHODS
published: 01 July 2022
doi: 10.3389/fimmu.2022.890943

A Structure-Based B-cell Epitope Prediction Model Through Combing Local and Global Features

Shuai Lu[1], *Yuguang Li*[1], *Qiang Ma*[2], *Xiaofei Nan*[1]* *and Shoutao Zhang*[2,3]*

[1] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China, [2] School of Life Sciences, Zhengzhou University, Zhengzhou, China, [3] Longhu Laboratory of Advanced Immunology, Zhengzhou, China

B-cell epitopes (BCEs) are a set of specific sites on the surface of an antigen that binds to an antibody produced by B-cell. The recognition of BCEs is a major challenge for drug design and vaccines development. Compared with experimental methods, computational approaches have strong potential for BCEs prediction at much lower cost. Moreover, most of the currently methods focus on using local information around target residue without taking the global information of the whole antigen sequence into consideration. We propose a novel deep leaning method through combing local features and global features for BCEs prediction. In our model, two parallel modules are built to extract local and global features from the antigen separately. For local features, we use Graph Convolutional Networks (GCNs) to capture information of spatial neighbors of a target residue. For global features, Attention-Based Bidirectional Long Short-Term Memory (Att-BLSTM) networks are applied to extract information from the whole antigen sequence. Then the local and global features are combined to predict BCEs. The experiments show that the proposed method achieves superior performance over the state-of-the-art BCEs prediction methods on benchmark datasets. Also, we compare the performance differences between data with or without global features. The experimental results show that global features play an important role in BCEs prediction. Our detailed case study on the BCEs prediction for SARS-Cov-2 receptor binding domain confirms that our method is effective for predicting and clustering true BCEs.

Keywords: Bi-LSTM, GCN, SARS-CoV-2, structure-based, attention, B-cell epitopes prediction

OPEN ACCESS

Edited by:
Roland Dunbrack,
Fox Chase Cancer Center,
United States

Reviewed by:
Philippe Auguste Robert,
University of Oslo, Norway
Kannan Sankar,
Novartis Institutes for BioMedical Research, United States

***Correspondence:**
Xiaofei Nan
iexfnan@zzu.edu.cn
Shoutao Zhang
zhangst@zzu.edu.cn

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *07 March 2022*
Accepted: *23 May 2022*
Published: *01 July 2022*

Citation:
Lu S, Li Y, Ma Q, Nan X and Zhang S (2022) A Structure-Based B-cell Epitope Prediction Model Through Combing Local and Global Features. Front. Immunol. 13:890943. doi: 10.3389/fimmu.2022.890943

1 INTRODUCTION

The humoral immune system protects the body from foreign objects like bacteria and viruses by developing B-cells and producing antibodies (1). Antibodies play a crucial role in immune response through recognizing and binding the disease-causing agents, called antigen. B-cell epitopes (BCEs) are a set of certain residueson the antigen surface that are bound by an antibody (2). BCEs of protein antigens can be roughly classified into two categories, linear and conformational (3). Linear BCEs consist of residues that are contiguous in the antigen primary sequence, while the conformational BCEs comprise residues which are not contiguous in sequence but folding together in three-

dimensional structure space. About 10% of BCEs are linear and about 90% are conformational (4). In our study, we focus on conformational BCEs of protein antigens.

The localization and identification of epitopes is of great importance for the development of vaccines and for the design of therapeutic antibodies (5, 6). However, traditional experimental methods to identify BCEs are still expensive and time-consuming (7). Therefore, great efforts for computational approaches based on machine learning algorithms have been developed to predict BCEs. These approaches can be divided in two categories: sequence-based and structure-based methods. As the name implies, the sequence-based approaches predict BCEs only based on the antigen sequence, while the structure-based approaches also consider its structural features. Currently, various structure-based predictors have been developed to predict and analyze BCEs including BeTop (8), Bpredictor (9), DiscoTope-2.0 (10), CE-KEG (11), CeePre (12), EpiPred (13), ASE_Pred (14) and PECAN (15).

Some of those methods improve model performance by introducing novel features such as statistical features in BeTop, thick surface patch in Bpredictor, new spatial neighborhood definition and half-sphere exposure in DiscoTope-2.0, knowledge-based energy and geometrical neighboring residue contents in CE-KEG, B factor in CeePre and surface patches in ASE_Pred. Except novel features, antibody structure information and suitable model also improve the performance of BCEs prediction. EpiPred utilizes antibody structure information to annotate the epitope region and improves global docking results. PECAN represents antigen or antibody structure as a graph and employ graph convolution operation on them to make aggregation of spatial neighboring residues. An additional attention layer is used to encode the context of the partner antibody in PECAN for predicting the antibody-specific BCEs rather than antigenic residues. Because antibody structure information is required, these methods are not applicable to a novel virus when its antibody is unknown. However, all the currently structure-based BCEs prediction methods only use local information around target amino acid residue without considering the global information of the whole antigen sequence.

Global features have been proved to be effective in some biology sequence analysis models such as protein-protein interaction sites prediction model DeepPPISP (16) and protein phosphorylation sites prediction model DeepPSP (17). However, which model is used for extracting global features is important. DeepPPISP utilizes TextCNNs processing the whole protein sequence for protein-protein interaction sites prediction. DeepPSP employs SENet blocks and Bi-LSTM blocks to extract the global features for protein phosphorylation sites. In our study, we take advantage of the Attention based Bidirectional Long Short-Term Memory (Att-BLTM) networks. Att-BLTM networks are first introduced for relation classification in the field of natural language processing (NLP) (18). Att-BLSTM networks are also employed for some chemical and biomedical text processing tasks including chemical named entity recognition (19) and biomedical event extraction (20). Given the excellent performance of Att-BLSTM, we combine it with the novel deep learning model Graph Convolution Networks (GCNs) (21) for BCEs prediction.

In this study, we propose a structure-based BCEs prediction model utilizing both antigen local features and global features. The source code of our method is available at https://github.com/biolushuai/GCNs-and-Att-BLSTM-for-BCEs-prediction. By combining Att-BLSTM and GCNs, both local and global features are used in our model to improve its prediction performance. We implement our model on some public datasets and the results show that global features can provide useful information for BCEs prediction.

2 MATERIALS AND METHODS

2.1 Datasets

In order to make fair comparison, we use the same antibody-antigen complexes as PECAN (15). It should be noted that those bound conformations are only used to identify epitope residues and no-epitope residues. Same as previous works (13, 15), residues are labeled as part of the BCEs if they have any heavy atom within 4.5Å away from any heavy atom on the antibody. As our model is partner independent, it only takes antigen structure as input for predicting BCEs.

Those complexes are from two separate datasets: EpiPred (13) and Docking Benchmarking Dataset (DBD) v5 (22). The 148 antibody-antigen complexes from EpiPred share no more than 90% pairwise sequence identity. Among them, 118 complexes are used for training and 30 for testing. For constructing a separate validation set, PECAN filters the antibody-antigen complexes in DBD v5 and selects 162 complexes which have no more than 25% pairwise sequence identity to every antigen in the testing set. Antigens in training set are used for training our model, antigens in validation set are used to tune the hyperparameters of our proposed method, and antigens in testing set are uesd for evaluation our model and making comparison with competing methods. The size of datasets and number of BCEs are shown in **Table 1**.

2.2 Input Features Representation

For global features, we construct the input antigen sequence as a set of sequential residues:

$$S = [r_1, r_2, r_3, \cdots, r_i, \cdots, r_l]^T, S \in R^{(l*d)} \quad (1)$$

where each residue is represented as a vector $r_i \in R^d$ corresponding to the i-th residue in the antigen sequence, l is the antigen sequence length, and d is the residue feature dimension.

For local features, each antigen structure is represented as a graph as related studies (13, 15, 23). The residue is a node in the protein graph whose features represent its properties. For residue r_i, the local environment N_i consists of k spatial neighboring residues:

$$N_i = \{n_1, \cdots, n_k\} \quad (2)$$

TABLE 1 | Summary of datasets.

Datasets	NO. of Complexes	NO. of BCEs	NO. of non-BCEs
Training Set	103	2708	19567
Validation Set	29	839	5553
Testing Set	30	758	6434

And, $\{ r_{n_1},\cdots,r_{n_k} \}$ are the neighbors of residue r_i which define the operation field of the graph convolution. The distance between r_i and are r_{n_k} calculated by averaging the distance between non-hydrogen atoms in r_i and r_{n_k}. In this study, node features and edge features in antigen graph are used for characterizing the local environment of target residue. The node features are represented as a 128-dimension vector encoding important properties as in our earlier work (24). All those node features can be divided into two classes: sequence-based and structure-based. Sequence-based features consist of the one-hot encoding of the amino acid residue type, seven physicochemical parameters (25) and evolutionary information. We utilize python script to encode the residue type and physicochemical parameters of each antigen sequence. The features that contain evolutionary information such as position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM) are returned by running PSI-BLAST (26) against nr database (27) using three iterations and an E-value threshold of 0.001. The structure-based features are calculated for each antigen structure isolated from the antibody-antigen complex by DSSP (28), MSMS (29), PSAIA (30) and Biopython (31).

The edge features between two residues r_i and r_j are representing as e_{ij}. e_{ij} reflects the spatial relationships including the distance and angle between residue pair r_i and r_j and it is computed by their C_a (23).

2.3 Model Architecture

Our model solves a binary classification problem: judging an antigen residue binding to antibody or not. As shown in **Figure 1**, our model consists of two parallel parts: GCNs and Att-BLSTM networks. The former captures local features of target antigen residue from its spatial neighbors by using graph convolutional layer, and the latter extracts global features from the whole antigen sequence by using Bi-LSTM layer and attention layer. The outputs are concatenated and fed to fully connected layer to predict the binding probability for each antigen residue.

2.3.1 Graph Convolutional Networks

Figure 2 shows the flow of convolution operation using the information of nodes and edges. At first, each protein is represented as a graph, and a residue is a node in the graph. The local environment of the target residue is a set of residues which are adjacent in space. And then, node and edge are represented by a vector as our previous work (24). Actually, the graph convolution operation on the local environment of target residue is the aggregation of neighboring residues and its edges. Every node in the graph is updated through repeated aggregation operation. Based on edges are used or not, we utilize two graph convolution operators in this study:

$$z_i = \sigma(W_t r_i + \frac{1}{|N_i|}\sum_{j\in N_i} W_n r_j + b_n) \tag{3}$$

$$z_i = \sigma(W_t r_i + \frac{1}{|N_i|}\sum_{j\in N_i} W_n r_j + \frac{1}{|N_i|}\sum_{j\in N_i} W_e e_{\mathrm{ij}} + b_{\mathrm{ne}}) \tag{4}$$

Where N_i is the receptive field, i.e. a set of neighbors of target residue r_i, W_t is the weight matrix associated with the target node, W_n is the weight matrix associated with neighboring nodes, σ is a non-linear activation function, and b_n is a bias vector. Formula 3 groups the node information in receptive filed. Formula 4 utilizes not only node features but also edge features between two residues, where W_e is the weight matrix associated with edge features, e_{ij} represents the edge features between residue r_i and r_j, and b_{ne} is a vector of biases.

2.3.2 Attention-Based Bidirectional Long Short-Term Memory Networks

Besides local features, global features are crucial in BCEs prediction as well. In our work, Attention-based Bidirectional Long Short-Term Memory (Att-BLSTM) networks are used to capture global sequence information of input antigen sequence. Currently, Att-BLSTM has been used for processing chemical and biomedical text (19, 20). It can capture the most important semantic information in a sequence. However, its advantage has not been exploited in biology sequence analysis such as BCEs prediction.

Figure 3 shows the architecture of Att-BLSTM. At first, the input antigen matrix S is fed into a Long Short-Term Memory (LSTM) network which learns long-range dependencies in a sequence (32, 33). Typically, the structure of an LSTM unit at each time t is calculated by the following formulas:

$$i_t = \sigma(W_i * [h_{t-1}, r_t] + b_i) \tag{5}$$

$$f_t = \sigma\left(W_f * [h_{t-1}, r_t] + b_f\right) \tag{6}$$

$$o_t = \sigma(W_o * [h_{t-1}, r_t] + b_o) \tag{7}$$

$$c_t = f_t * c_{t-1} + i_t * (tanh(W_c * [h_{t-1}, r_t] + b_c)) \tag{8}$$

$$h_t = o_t * tanh(c_t) \tag{9}$$

where *tanh* is the element-wise hyperbolic tangent, σ is the logistic sigmoid function, r_t, h_{t-1} and c_{t-1} are inputs, and h_t and c_t are outputs. There are three gates consisting of one input gate i_t with corresponding weightmatrix W_i, and a bias b_i; one forget gate f_t with corresponding weight matrix W_f, and a bias b_f one output gate o_t with corresponding weight matrix W_o, and a bias b_o.

Bidirectional LSTM (Bi-LSTM) can learn forward and backward information of input sequence. As shown in **Figure 2**, the networks contain two sub-networks for the left

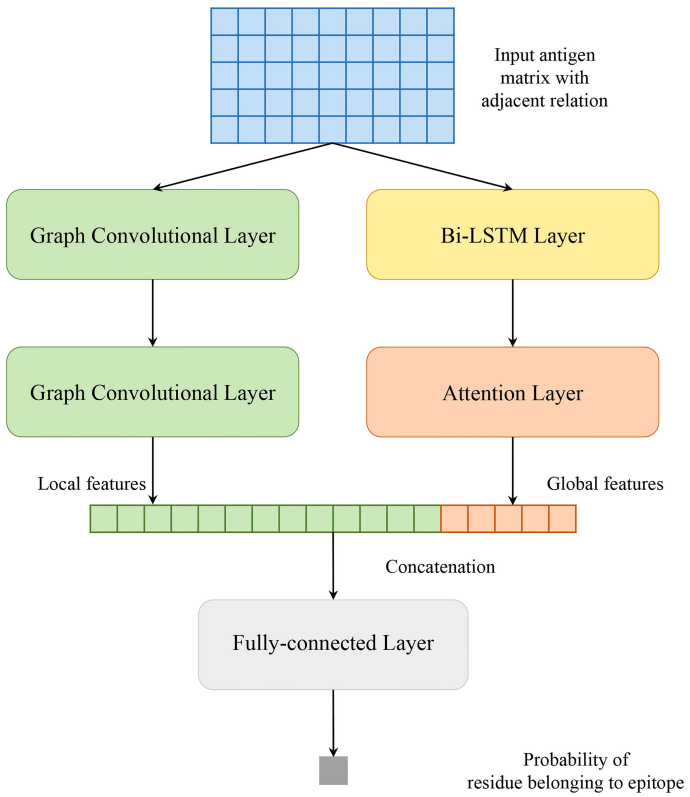

FIGURE 1 | Model architecture of proposed method in this study.

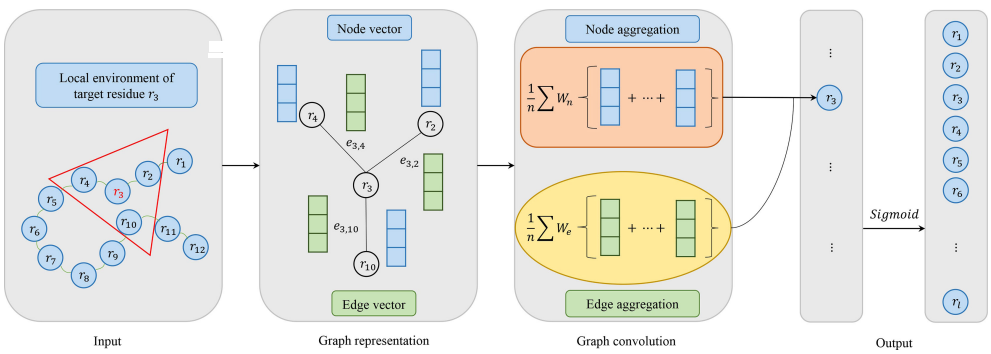

FIGURE 2 | Model architecture of graph convolutional networks consisting of four parts: input, graph representation, graph convolution and output. In the input, the target residue is red and the local environment of the target residue is in a red triangle. For graph representation, node vector is blue and edge vector is green. For graph convolution, if edge vector is used, it corresponds to ^#the formula 4. Otherwise, it corresponds to the formula 3. In the output, we use the Sigmoid activation function.

and right sequence contexts. For the i-th residue in the input antigen sequence, we combine the forward pass output $\overrightarrow{h_i}$ and backward pass output $h_i^{\leftarrow}$ by concatenating them:

$$h_i^{'} = [\overrightarrow{h_i} \oplus \overleftrightarrow{h_i}]$$

The output of Bi-LSTM layer is matrix H which consists of all output vectors of input antigen residues: $H = [h_1^{'}, h_2^{'}, h_3^{'}, \ldots, h_l^{'}]^T, H \in R^{(l*2d)}$,where l is the input antigen sequence length, and d is the residue features dimension.

Attention mechanism has been used in a lot of biology tasks ranging from compound-protein interaction prediction (34), paratope prediction (15) and protein structure prediction (35). The attention layer in our model employs a classical additive model in which α is the attention weight. After attention layer of Att-BLSTM, the novel representation $S^{'}$ as well as the output y_g of

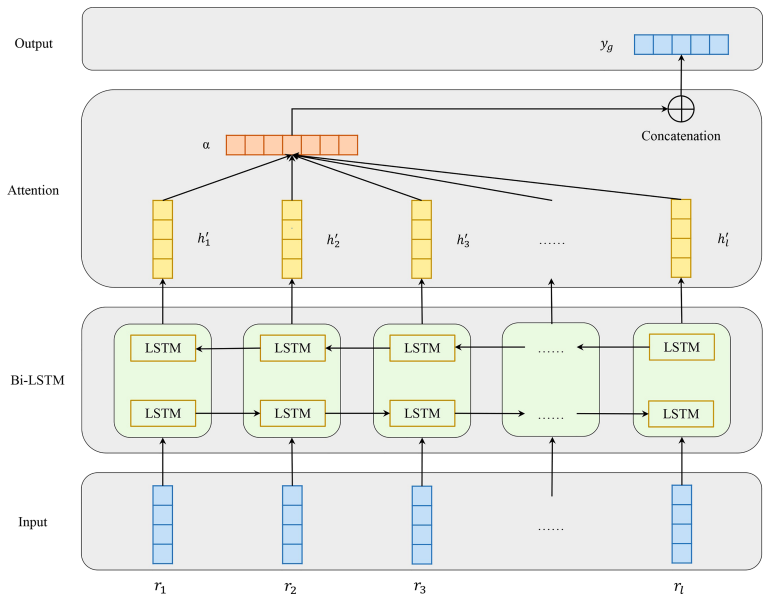

FIGURE 3 | Model architecture of attention-based bidirectional long short-term memory networks which consists of four parts: input, Bi-LSTM layer, Attention layer and output.

the input antigen is formed by a weighted sum of those output vectors H:

$$M = tanh(H) \tag{11}$$

$$\alpha = softmax\left(W_\mu M\right) \tag{12}$$

$$y_g = S' = H\alpha^T \tag{13}$$

2.3.3 Fully-Connected Networks

As shown in **Figure 1**, the local features z_i extracted by GCNs and the global features y_g derived from Att-BLSTM networks are concatenated. And then, they are fed to fully-connected layer. The calculation of probability y_i for each input antigen residue belonging to BCEs is shown as:

$$y_i = f\left(W\left(y_g \oplus z_i\right) + b\right) \tag{14}$$

2.4 Performance Evaluation

In order to make comparison with state-of-the-art structure-based BCEs predictors, we use three evaluation metrics to evaluate the performances of the BCEs prediction models: Precision, Recall and Matthews Correlation Coefficient (MCC) which are shown as followings:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{17}$$

where, TP (True Positive) is the number of interacting residues that are correctly predicted as BCEs, FP (False Positive) is the number of non-interacting residues that are falsely predicted as BCEs, TN (True Negative) denotes the number of non-interacting sites that are identified correctly, and FN (False Negative) denotes the number of interacting sites that are identified falsely. Precision and recall reflect the prediction tendencies of classifiers. Recall indicates the percentage of correct predictions for positive and negative samples. Precision shows the percentage of correct positive samples. There is a trade-off between precision and recall. Recall favors positive-bias predictions, while precision favors negative predictions.

Because precision, recall and MCC are threshold-dependent, we also utilize the area under the receiver operating characteristics curve (AUC ROC) and precision-recall curve (AUC PR) which gives a threshold-independent evaluation on the overall performance. Moreover, AUC PR is more sensitive than AUC ROC on imbalanced data (36). And, the datasets used for BCEs prediction are roughly 90% negative class. Therefore, we take AUC PR as the most import metric for model evaluation and selection.

It should be noted that the precision, recall and MCC shown in **Table 2** are averaged over all antigens in the testing set. And, the AUC ROC and AUC PR reported in **Figures 4**, **5** are calculated among all antigen residues in the testing set.

2.5 Implementation Details

We implement our model using PyTorch. The training details of these neural networks are as follows: optimization: Momentum

TABLE 2 | Performances of BCEs prediction methods.

Method	Precision	Recall	MCC
DiscoTope-2.0	0.214	0.110	0.096
EpiPred	0.136	0.436	0.156
PECAN	0.154	**0.691**	NA
Our Method	**0.657**	0.671	**0.319**

Best values are in bold.

optimizer with Nesterov accelerated gradients; learning rate: 0.1, 0.01, 0.001 and 0.0001; batch size: 32, 64 and 128; dropout: 0.2,0.5 and 0.7; spatial neighbors in the graph: 20; number of LSTM layers in Att-BLSTM networks: 1, 2 or 3; number of graph convolution networks layers: 1, 2 or 3. Training time of each epoch varies from roughly 1 to 3 minutes depending on network depth, using a single NVIDIA RTX2080 GPU.

For each combination, networks are trained until the performance on the validation set stops improving or for a maximum of 250 epochs. Graph convolution networks have the following number of filters for 1, 2 and 3 layers, respectively: (256), (256, 512), (256, 256, 512). All weight matrices are initialized as (23) and biases are set to zero.

3 RESULTS AND DISCUSSION

3.1 The Effects of Different Network Combinations

In this section, we focus on which network combinations are most effective. The AUC ROC and AUC PR are shown in **Figure 4.**

First, we train our model of 1-layer Att-BLSTM with varying GCNs depths with or without residue edge features. From **Figures 4**, we observe that the 2-layer GCNs with residue edge features perform best (AUC ROC = 0.804, AUC PR = 0.376). This draws the same conclusion with our earlier work for antibody paratope prediction (24). We also find that residue edge features can always provide better performance as the GCNs depths vary. The same results are found in protein interface prediction task using GCNs as well (23).

Second, 2-layer GCNs and Att-BLSTM networks of different depths are combined in our model. **Figures 4** show the performance evaluated by AUC ROC and AUC PR. It can be found that the combination of 1-layer Att-BLSTM network and 2-layer GCNs with residue edge features still has the best results. In general, the deeper the Att-BLSTM networks grow, the results get worse. As discussed in DeepPPISP (16), global features may cover the relationships among residues of longer distances. However, as Att-BLSTM networks become deeper, these relationships may become weaker.

In summary, our model with 2-layes GCNs and 1-layer Att-BLSTM network performs best, and it is the proposed model in this paper and used for comparison with competing methods in the following sections.

3.2 The Effects of Global Features

The global feature has been shown to improve the performance of protein-protein interaction sites prediction in DeepPPISP (16) and protein phosphorylation sites prediction in DeepPSP (17). In order to verify whether global features are effective in BCEs prediction as well, we remove the Att-BLSTM networks in our model for comparison. As shown in **Figure 4**, label G0 means there is only GCNs in our model, and no global features are used. Without global features, the AUC ROC is 0.787, which is lower than the proposed model G1NE2 (also lower thanG2NE2, but slightly better than G3NE2). Without global features, the AUC PR is 0.335, which is significantly worse than the proposed model (but slightly better than G2NE2 and G3NE2). The model without global features performs worse on both AUC ROC and AUC PR metrics than our proposed model. Therefore, global features improve the performance of our model for BCEs prediction.

However, in our experiments, models with global features are not always superior to models without global features. Similar observation has been found in DeepPSP, but DeepPPISP reaches a contrary conclusion. This situation might be caused by different models processing global features. In DeepPPISP, a simple fully-connected network is used, and in DeepPSP, SENet blocks and Bi-LSTM blocks are used.

3.3 The Effects of Different Types of Input Features

Different types of input features (sequence and structure-based) play different roles in our model. The input features can be divided into four types: (a) residue type one-hot encoding at alphabetical order, (b) evolutionary information of antigen sequence such as PSSM and PSFM, (c) seven physicochemical parameters returned by machine leaning model and (d) structural features consisting of solvent accessibility, secondary structure, dihedral angle, depth, protrusion and B-value of every residue calculated by various bioinformatic tools. To discover what role each feature typeplays in our method, we delete each input feature type and compare their performances on our proposed model (G1NE2, i.e., 1-layer Att-BLSTM network and 2-layersGCNs with residue edge features). **Figure 5** shows the experimental results. As **Figure 5** shows, the AUC ROC without features b is 0.754, significantly lower than the best performance 0.804. The AUC PR without features b drops biggest from 0.376 (all features) to 0.289. It indicates that evolutionary information profile features (feature type b) are most important in our model for BCEs prediction. The model using all the features still performs best on both AUC ROC and AUC PR metrics.

3.4 Comparison With Competing Methods

To evaluate the performance of our method for BCEs prediction, we compare our proposed model with three competing structure-based BCEs prediction methods: DiscoTope-2.0 (10), EpiPred (13) and PECAN (15). Note that these methods all used local features but did not consider global features. The precision, recall and MCC calculated in this study using a threshold 0.116 at which our method achieves best performance on the testing set. **Table 2** shows the experimental results of our method and the competing models. The results on three competing models are taken from (13). Although our model gets lower recall than PECAN, it is higher than all other competing methods on precision and MCC.

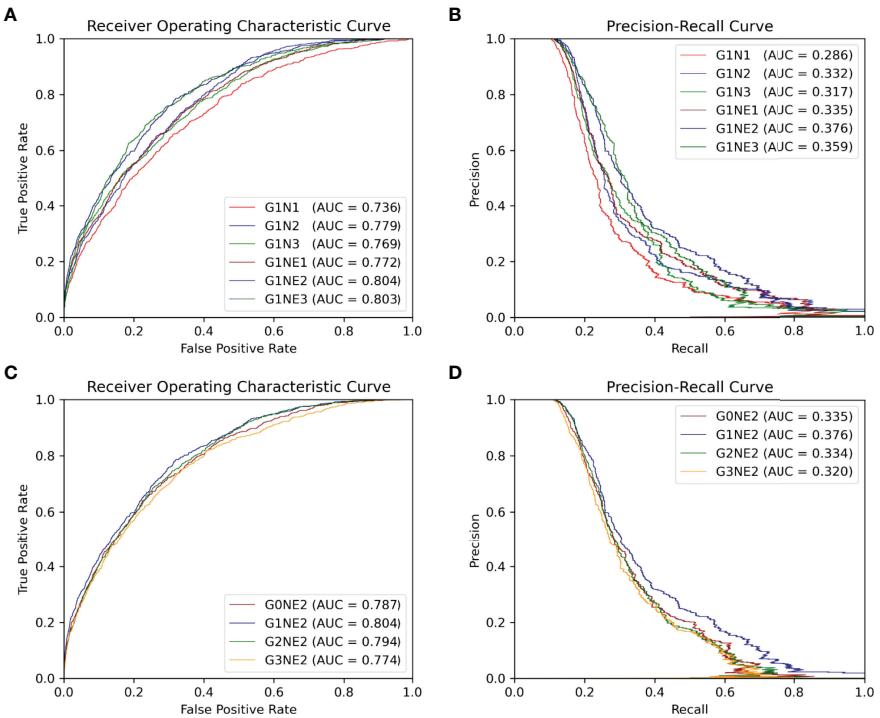

FIGURE 4 | ROC and PR curve of different network combinations among all antigen residues in the testing set. **(A, B)** ROC and PR curve using different networks processing local features. **(C, D)** ROC and PR curve of the using different networks processing gobal features.

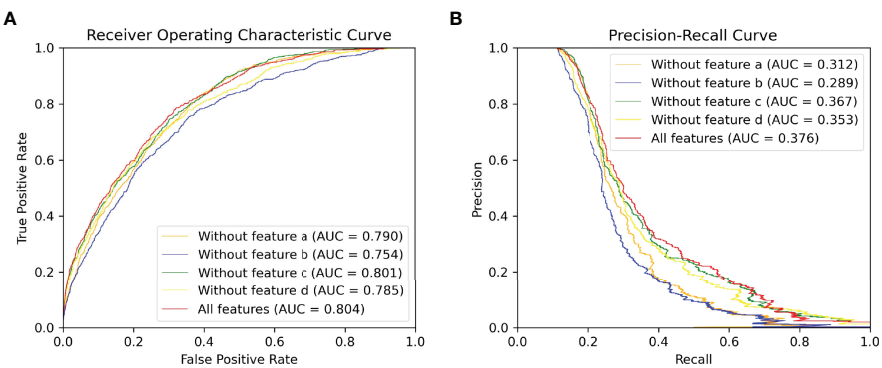

FIGURE 5 | ROC and PR curve of different combinations of input features among all antigen residues in the testing set. **(A)** ROC curve. **(B)** PR curve.

We also compare the results of each antigen in testing set with DiscoTope-2.0 and EpiPred. The results presented of DiscoTope-2.0 and EpiPred in **Table S1** are taken form (13). The values in bold indicate the best prediction result. We find that our model achieves best precision on 26 antigens, best recall on 20 antigens and best MCC on 20 antigens of all 30 antigens in testing set. We also observe that our model produces usable prediction even for the long antigen target as the global features provide information from long distance effect.

3.5 Case Study

We also employ our method for predicting BCEs of SARS-Cov-2 which caused the coronavirus disease 2019 (COVID-19) pandemic. The entry of SARS-CoV-2 into its target cells depends on binding between the Receptor Binding Domain (RBD) of the viral Spike (S) protein and its cellular receptor, angiotensin-converting enzyme 2 (ACE2) (37). A number of neutralizing antibodies (NAbs) are reported and most bind the RBD of the S protein. According to the published works and determined complexes, NAbs target

SARS-CoV-2 with various conformations and neutralization mechanisms. These NAbs can be divided into five types (type 1 to type 5) based on different epitopes they target (38). **Table 3** shows the five types of antibodies and the neutralization mechanisms of them. And, we randomly select a representative complex structure from Protein Data Bank (PDB) (33) of each type for predicting the corresponding five types of BCEs.

The BCEs prediction results are listed in **Table 4**. Compared with the competing predictors, our method achieves the best performance for every metric when predicts BCEs type 2 and type 4. For BCEs type 3, the recall and MCC of our method are highest. Higher recall indicates that more true epitopes are predicted and higher MCC states the overall performance of our method is better. For BCEs type 1, Discotope-2.0 performs best and our method ranks only second to it on recall and MCC. For BCEs type 5, PECAN achieves best precision and MCC and the results of our method are not good. It should be noted that epitopes type 5 are located in the N-terminal domain (NTD) of S1 protein rather than RBD region. And, it's different with all other four BCEs types.

In order to visually show the prediction results for all 5 types of BCEs, we show in **Figure 6** the true and predicted BCEs by our model and other competing methods. For each BCEs type, we utilize the representative antigen structure as show in **Table 3**.

4 CONCLUSIONS

Accurate prediction of BCEs is helpful for understanding the basis of immune interaction and is beneficial to therapeutic design. In this work, we propose a novel deep learning framework combining local and global features which are extracted from antigen sequence and structure to predict BCEs. GCNs are used for capturing the local features of a target residue. Att-BLSTM networks are used to extract global features, which figure the relationship between a target residue and the whole antigen. We employ our model on a public and popular dataset and the results show improvement of BCEs prediction. Moreover, our results declare that the global features are useful for improving the prediction of BCEs.

For deep case study, we apply our method to the BCEs prediction for SARS-Cov-2. According to summarized works and analyzed complex structures, there are many different types of SARS-Cov-2 BCEs. However, our method doesn't perform

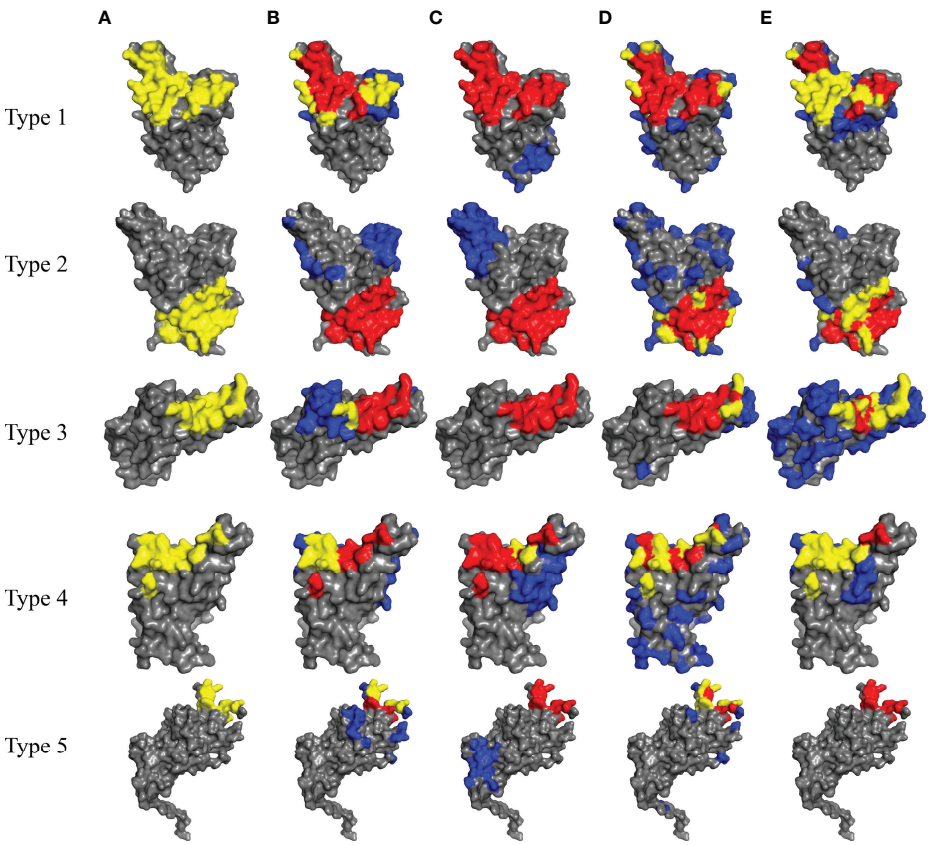

FIGURE 6 | Prediction results for SARS-Cov-2 of five types of BCEs (type 1 to type 5). **(A)** The true epitope residues. **(B–E)** Prediction results by Discotope-2.0, EpiPred, PECAN and our method, respectively. TP predictions are in yellow, FN predictions are in red, FP predictions are in blue and the background grey represents TN predictions.

TABLE 3 | Five types of antibodies neutralizing by SARS-Cov-2.

Type	Antibody name	PDB ID	Neutralizing mechanism	References
1	C102	7K8M	Block the hACE2-RBD interaction	(39)
2	CR3022	6YOR	Trap the RBD in the up conformation	(40)
3	S2M11	7K43	Lock the RBD in the down conformation	(41)
4	P2B-2F6	7BWJ	Compete with ACE2 and prevent the RBD from binding	(42)
5	FC05	7CWU	Target non-RBD regions	(43)

TABLE 4 | Prediction performances on five types of SARS-Cov-2 BCEs and best values are in bold.

Epitopes (PDB ID and chains name)	Methods	Precision	Recal	MCC
Type 1 (7K8M_AB_E)	Discotope-2.0	**0.660**	**0.649**	**0.297**
	Epipred	0.402	0.401	-0.199
	PECAN	0.638	0.583	0.143
	Our method	0.576	0.631	0.200
Type 2 (6YOR_HL_E)	Discotope-2.0	0.432	0.430	-0.140
	Epipred	0.431	0.423	-0.153
	PECAN	0.494	0.492	-0.016
	Our method	**0.664**	**0.658**	**0.322**
Type 3 (7K43_HL_A)	Discotope-2.0	0.526	0.554	0.095
	Epipred	0.493	0.489	-0.021
	PECAN	**0.531**	0.545	0.086
	Our method	0.525	**0.773**	**0.166**
Type 4 (7BWJ_HL_E)	Discotope-2.0	0.589	0.615	0.226
	Epipred	0.494	0.491	-0.017
	PECAN	0.555	0.591	0.171
	Our method	**0.684**	**0.847**	**0.506**
Type 5 (7CWU_PI_C)	Discotope-2.0	0.547	**0.641**	0.226
	Epipred	0.494	0.490	-0.019
	PECAN	**0.579**	0.626	**0.240**
	Our method	0.494	0.500	0.000

The PDB ID and chains name of representative complex structures are shown as PDB ID_Antibody heavy chain and light chain name_ Antigen (SARS-Cov-2) chain name. The results of Discotope-2.0 and Epipred are obtained from their websites using suggested threshold. For PECAN, we download its source code and run it for making comparison. It should be noted that Epipred and PECAN take both antigen and its partner antibody structure as input. Discotope-2.0 and our method only utilize isolated antigen structure.

best for every BCEs type, but it achieves best results for three types of SARS-Cov-2 BCEs.

Though our method outperforms other competing computational methods for BCEs prediction, it also has some disadvantages. The first one is that our predictor needs antigen structure as it takes structure-based residue features as input. The second one is that our model∈it consumes long computer time because PSI-BLAST (26) needs to be performed at the stage of extracting residue features. The third one is that although our method performs better than comparative models for predicting BCEs of SARS-Cov-2, it can be observed that our method is not very good at predicting non-overlapping BCEs.

In this study, we show that combing local and global features can be useful for BCEs prediction. In the future, we would further improve BCEs prediction by expanding the training set and utilizing the partner antibody structure of the antigen.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

SL designed the study. SL and YL performed the method development. SL, XN and QM performed the data analysis. XN and SZ wrote and revised the manuscript. All authors reviewed the manuscript.

FUNDING

This work was funded by Bingtuan Science and Technology Project (2019AB034), 'Created Major New Drugs' of Major National Science and Technology (2019ZX09301-159), Leading Talents Fund in Science and Technology Innovation in Henan Province (194200510002), and Natural Science Foundation of Henan Province of China (202300410381).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.890943/full#supplementary-material

REFERENCES

1. Getzoff ED, Tainer JA, Lerner RA, Geysen HM. The Chemistry and Mechanism of Antibody Binding to Protein Antigens. *Adv Immunol* (1988) 43:1–98. doi: 10.1016/S0065-2776(08)60363-6
2. Michnick SW, Sidhu SS. Submitting Antibodies to Binding Arbitration. *Nat Chem Biol* (2008) 4:326–9. doi: 10.1038/nchembio0608-326
3. Barlow DJ, Edwards MS, Thornton JM. Continuous and Discontinuous Protein Antigenic Determinants. *Nature* (1986) 322:747–8. doi: 10.1038/322747a0
4. Caoili SEC. Hybrid Methods for B-Cell Epitope Prediction Approaches to the Development and Utilization of Computational Tools for Practical Applications. *Methods Mol Biol* (2014) 1184:245–83. doi: 10.1007/978-1-4939-1115-8_14
5. Akbar R, Bashour H, Rawat P, Robert PA, Smorodina E, Cotet TS, et al. Progress and Challenges for the Machine Learning-Based Design of Fit-for-Purpose Monoclonal Antibodies. *mAbs* (2022) 14:2008790. doi: 10.1080/19420862.2021.2008790
6. Chan AC, Carter PJ. Therapeutic Antibodies for Autoimmunity and Inflammation. *Nat Rev Immunol* (2010) 10:301–16. doi: 10.1038/nri2761
7. Abbott WM, Damschroder MM, Lowe DC. Current Approaches to Fine Mapping of Antigen-Antibody Interactions. *Immunology* (2014) 142:526–35. doi: 10.1111/imm.12284
8. Zhao L, Wong L, Lu L, Hoi SC. Li J. B-Cell Epitope Prediction Through a Graph Model. *BMC Bioinf* (2012) 13:1–12. doi: 10.1186/1471-2105-13-S17-S20
9. Zhang W, Xiong Y, Zhao M, Zou H, Ye X, Liu J. Prediction of Conformational B-Cell Epitopes From 3D Structures by Random Forests With a Distance-Based Feature. *BMC Bioinf* (2011) 12:1–10. doi: 10.1186/1471-2105-12-341
10. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Comput Biol* (2012) 8:e1002829. doi: 10.1371/journal.pcbi.1002829
11. Lo YT, Pai TW, Wu WK, Chang HT. Prediction of Conformational Epitopes With the Use of a Knowledge-Based Energy Function and Geometrically Related Neighboring Residue Characteristics. *BMC Bioinf* (2013) 14:1–10. doi: 10.1186/1471-2105-14-S4-S3
12. Ren J, Liu Q, Ellis J, Li J. Tertiary Structure-Based Prediction of Conformational B-Cell Epitopes Through B Factors. *Bioinformatics* (2014) 30:264–73. doi: 10.1093/bioinformatics/btu281
13. Krawczyk K, Liu X, Baker T, Shi J, Deane CM. Improving B-Cell Epitope Prediction and its Application to Global Antibody-Antigen Docking. *Bioinformatics* (2014) 30:2288–94. doi: 10.1093/bioinformatics/btu190
14. Jespersen MC, Mahajan S, Peters B, Nielsen M. Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. *Front Immunol* (2019) 10:298. doi: 10.3389/fimmu.2019.00298
15. Pittala S, Bailey-Kellogg C. Learning Context-Aware Structural Representations to Predict Antigen and Antibody Binding Interfaces. *Bioinformatics* (2020) 36:3996–4003. doi: 10.1093/bioinformatics/btaa263
16. Zeng M, Zhang F, Wu FX, Li Y, Wang J, Li M. Protein-Protein Interaction Site Prediction Through Combining Local and Global Features With Deep Neural Networks. *Bioinformatics* (2020) 36:1114–20. doi: 10.1093/bioinformatics/btz699
17. Guo L, Wang Y, Xu X, Cheng KK, Long Y, Xu J, et al. DeepPSP: A Global-Local Information-Based Deep Neural Network for the Prediction of Protein Phosphorylation Sites. *J Proteome Res* (2021) 20:346–56. doi: 10.1021/acs.jproteome.0c00431
18. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016). Stroudsburg, PA USA: Association for Computational Linguistics. p. 207–12. doi: 10.18653/v1/p16-2034
19. Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H, et al. An Attention-Based BiLSTM-CRF Approach to Document-Level Chemical Named Entity Recognition. *Bioinformatics* (2018) 34:1381–8. doi: 10.1093/bioinformatics/btx761
20. Li L, Wan J, Zheng J, Wang J. Biomedical Event Extraction Based on GRU Integrating Attention Mechanism. *BMC Bioinf* (2018) 19:93–100. doi: 10.1186/s12859-018-2275-2
21. Kipf TN, Welling M. Semi-Supervised Classfication With Graph Convolutional Networks. In: *Proceedings of the 5th International Conference on Learning Representations* (2017). OpenReview.net. p. 1–10.
22. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* (2015) 427:3031–41. doi: 10.1016/j.jmb.2015.07.016
23. Fout A, Byrd J, Shariat B, Ben-Hur A. Protein Interface Prediction Using Graph Convolutional Networks. In: *Proceedings of the 31st Annual Conference on Neural Information Processing Systems* (2017). LA California USA: Neural Information Processing Systems. p. 6531–40.
24. Lu S, Li Y, Wang F, Nan X, Zhang S. Leveraging Sequential and Spatial Neighbors Information by Using CNNs Linked With GCNs for Paratope Prediction. *IEEE/ACM Trans Comput Biol Bioinf* (2022) 19:68–74. doi: 10.1109/TCBB.2021.3083001
25. Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and Evaluation of Dimension-Reduced Amino Acid Parameter Representations by Artificial Neural Networks. *J Mol Model* (2001) 7:360–9. doi: 10.1007/s008940100038
26. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res* (1997) 25:3389–402. doi: 10.1093/nar/25.17.3389
27. McGinnis S, Madden TL. BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res* (2004) 32:20–5. doi: 10.1093/nar/gkh435
28. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* (1983) 22:2577–637. doi: 10.1002/bip.360221211
29. Sanner MF, Olson AJ, Spehner J. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* (1996) 38:305–20. doi: 10.1002/(sici)1097-0282(199603)38:3<305::aid-bip4>3.3.co;2-8
30. Mihel J, Šikić M, Tomić S, Jeren B, Vlahoviček K. PSAIA - Protein Structure and Interaction Analyzer. *BMC Struct Biol* (2008) 8:1–11. doi: 10.1186/1472-6807-8-21
31. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* (2009) 25:1422–3. doi: 10.1093/bioinformatics/btp163
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* (2000) 28:235–42. doi: 10.1093/nar/28.1.235
33. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735
34. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T, et al. TransformerCPI: Improving Compound–Protein Interaction Prediction by Sequence-Based Deep Learning With Self-Attention Mechanism and Label Reversal Experiments. *Bioinformatics* (2020) 36:4406–14. doi: 10.1093/bioinformatics/btaa524
35. Berkeley UC, Meier J, Sercu T, Rives A. Transformer Protein Language Models Are Unsupervised Structure Learners. In: *Proceedings of the 9th International Conference on Learning Representations* (2021). OpenReview.net. p. 1–24.
36. Staeheli LA, Mitchell D. The Relationship Between Precision-Recall and ROC Curves Jesse. In: *Proceedings of the 23rd International Conference on Machine Learning* (2006). New York, NY, USA: Association for Computing Machinery. p. 233–40. doi: 10.1145/1143844.1143874
37. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A Pneumonia Outbreak Associated With a New Coronavirus of Probable Bat Origin. *Nature* (2020) 579:270–3. doi: 10.1038/s41586-020-2012-7
38. Xue JB, Tao SC. Epitope Analysis of Anti-SARS-CoV-2 Neutralizing Antibodies. *Curr Med Sci* (2021) 41:1065–74. doi: 10.1007/s11596-021-2453-8
39. Barnes CO, Jette CA, Abernathy ME, Dam KMA, Esswein SR, Gristick HB, et al. SARS-CoV-2 Neutralizing Antibody Structures Inform Therapeutic Strategies. *Nature* (2020) 588:682–7. doi: 10.1038/s41586-020-2852-1
40. Huo J, Zhao Y, Ren J, Zhou D, Duyvesteyn HM, Ginn HM, et al. Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host Microbe* (2020) 28:445–54. doi: 10.1016/j.chom.2020.06.010

41. Tortorici MA, Beltramello M, Lempp FA, Pinto D, Dang HV, Rosen LE, et al. Ultrapotent Human Antibodies Protect Against SARS-CoV-2 Challenge *via* Multiple Mechanisms. *Science* (2020) 370:950–7. doi: 10.1126/science.abe3354
42. Ju B, Zhang Q, Ge J, Wang R, Sun J, Ge X, et al. Human Neutralizing Antibodies Elicited by SARS-CoV-2 Infection. *Nature* (2020) 584:115–9. doi: 10.1038/s41586-020-2380-z
43. Wang N, Sun Y, Feng R, Wang Y, Guo Y, Zhang L, et al. Structure-Based Development of Human Antibody Cocktails Against SARS-CoV-2. *Cell Res* (2021) 31:101–3. doi: 10.1038/s41422-020-00446-w

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.