



Flexibility-aware graph-based algorithm improves antigen epitopes identification

Chuang Gao¹, Yiqi Wang⁴, Jie Luo³, Ziyi Zhou³, Zhiqiang Dong^{2,4,*}, Liang Zhao^{1,3,5,*}

1 School of Computing and Electronic Information, Guangxi University, Guangxi, China

2 Brain Research Institute, Taihe Hospital, Hubei University of Medicine, Hubei, China

3 Precision Medicine Research Center, Taihe Hospital, Hubei University of Medicine, Hubei, China

4 College of Life Science and Technology, Huazhong Agricultural University, Hubei, China

5 Hubei Key Laboratory of Embryonic Stem Cell Research, Hubei University of Medicine, Hubei, China

 These authors contributed equally to this work.

* dongz@mail.hzau.edu.cn; S080011@e.ntu.edu.sg

Abstract

Epitopes of an antigen are the surface residues in the spatial proximity that can be recognized by antibodies. Identifying such residues has shown promising potentiality in vaccine design, drug development and chemotherapy, thus attracting extensive endeavors. Although great efforts have been made, the epitope prediction performance is still unsatisfactory. One possible issue accounting to this poor performance could be the ignorance of structural flexibility of antigens. Flexibility is a natural characteristic of antigens, which has been widely reported. However, this property has never been used by existing models. To this end, we propose a novel flexibility-aware graph-based computational model to identify epitopes. Unlike existing graph-based approaches that take the static structures of antigens as input, we consider all possible variations of the side chains in graph construction. These flexibility-aware graphs, of which the edges are highly enriched, are further partitioned into subgraphs by using a graph clustering algorithm. These clusters are subsequently expanded into larger graphs for detecting overlapping residues between epitopes if exist. Finally, the expanded graphs are classified as epitopes or non-epitopes via a newly designed graph convolutional network. Experimental results show that our flexibility-aware model markedly outperforms existing approaches and promotes the F1-score to 0.656. Comparing to the state-of-the-art, our approach makes an increment of F1-score by 16.3%. Further in-depth analysis demonstrates that the flexibility-aware strategy contributes the most to the improvement. The source codes of the proposed model is freely available at <https://github.com/lzhlab/epitope>.

Author summary

Epitope prediction is helpful to many biomedical applications so that dozens of models have been proposed aiming at improving prediction efficiency and accuracy. However, the performances are still unsatisfactory due to its complicated nature, particularly the

noteworthy flexible structures, which makes the precise prediction even more challenging. The existing approaches have overlooked the flexibility during model construction. To this end, we propose a graph model with flexibility heavily involved. Our model is mainly composed of three parts: i) flexibility-aware graph construction; ii) overlapping subgraph clustering; iii) graph convolutional network-based subgraph classification. Experimental results show that our newly proposed model markedly outperforms the existing best ones, making an increment of F1-score by 16.3%.

Introduction

A B-cell epitope is a specific region at the surface of an antigen that can be neutralized by antibodies, and this neutralization can consequently elicit crucial immune response [1]. Identification of epitopes can be useful to design vaccines, drugs, reagents and so on [2-5]. Hence, intensive efforts have been made to develop epitope prediction models, including experimental-based and computational-based approaches. Experimental methods, such as X-ray crystallography, nuclear magnetic resonance and phage display [6], are accurate but labor intensive as well as costly, while computational models are more efficient and economical, but suffer from lower accuracy. Due to the difficulty of improving experimental approaches and the fast-evolving computational techniques, massive efforts have been made from the computational perspective.

A general protocol of the computational-based epitope prediction models is: i) collecting antigen-antibody interaction complexes to obtain epitopes; ii) engineering features of epitopes from chemo-physical and statistical perspectives, such as residue's polarity [7], hydrophobicity [8], protrusion index [9], relative frequency [10], et al.; iii) building computational models based on the features, such as machine learning models [11], graph models [12], statistical models [13], et al.; iv) identifying new epitopes via the well trained models. Although dozens of methods have been proposed, the prediction accuracy still has a big room to improve. For instance, the F1-score of the state-of-the-art epitope prediction model is still less than 0.6 [6]. One possible reason could be the ignorance of structural flexibility of antigens. To our best knowledge, all existing approaches, either experimental or computational, are based on the static structures of antibody-antigen interacting complexes, that is, the structures of antigens and antibodies are fixed when epitopes are determined. However, the structure of an antigen can be flexible [14,15], particularly the side chain of the surface residues. Therefore, incorporating flexibility into epitope prediction models could be helpful and promising.

Protein flexibility has been widely reported, and there are three types of flexibility, i.e., local flexibility (at atom/residue level) [16], regional flexibility (at intra-domain/multi-residue level) [17], and global flexibility (at multi-domain level) [18]. The local and regional flexibility are mainly caused by the movement of chemical bonds and bond angles [19], while the global flexibility mostly comes from the hinges, helical-to-extended conformations and side chains [20-22]. Taking as an example shown in Fig 1, the two structures (having the Protein Data Bank (PDB) [23] identifier of 1A14 and 7NN9, respectively) have exactly the same sequence of residues, but notably different structures after aligned. The 1A14 is a bonded structure interacting with the anti-influenza virus neuraminidase, while the 7NN9 is an unbound structure. After alignment, the averaged RMSD (root mean square deviation) of the atoms between the two structures is $0.434 \pm 0.636 \text{ \AA}$, while the RMSD of the atoms at the epitope region is $0.576 \pm 0.748 \text{ \AA}$, which is 1.33 times larger than the former one, indicating a markedly movement of conformation upon binding. More interestingly, the side chain RMSD of the epitope residues is $0.853 \pm 0.993 \text{ \AA}$, showing a big fluctuation; cf. Fig 1(A) and (C). These observations indicate that using bonded epitope conformations as training data to

figure out epitopes from antigens that have unknown binding partners, to some extent, is inaccurate. This has also been argued by Zhao et al [6]. Unfortunately, all existing models, both experimental and computational, have overlooked this property.

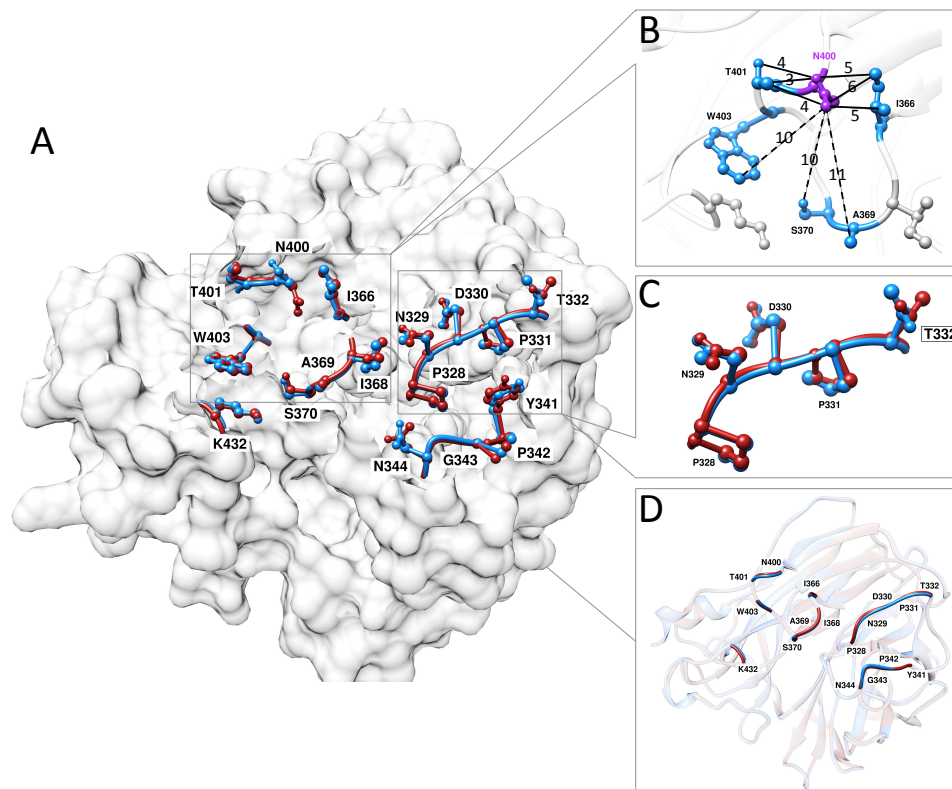


Fig 1. Flexibility illustration. Panel A shows the superimposed structures of a neuraminidase from an influenza virus with (PDB ID 1A14) and without (PDB ID 7NN9) associated antibodies, in which the epitopes are highlighted in balls and sticks. Panel B shows the connections between the residue asparagine (on chain N of 1A14 at position 400) and its neighbors. The solid lines are the connections without considering flexibility, while the dash lines are the connections after flexibility is incorporated. The number on the lines are the distance between the linked atoms measured in angstrom (\AA) based on the complex of 1A14. Panel C illustrates the spatial discrepancy between the epitopes on 1A14 and 7NN9, and Panel D shows the superimposed structure between the two in cartoon without side chains. Structures are downloaded from the Protein Data Bank [23] and the figures are generated using the Chimera [24].

To overcome this problem, we propose a graph model that fully considers the local flexibility for epitope prediction; see Fig 2. This model starts with flexibility-aware graphs construction by exploring the distortion of the side chain of all surface residues. These edge-enriched graphs are subsequently partitioned into non-overlapping subgraphs by a message flowing algorithm [25] and expanded into overlapping subgraphs via a community detection algorithm [26]. Finally, a graph convolutional network (GCN) is built to discriminate the expanded subgraphs into epitopes or non-epitopes. Experimental results show that the proposed model elevates the F1-score to 0.656, making an increment of 16.3% compared to the state-of-the-art. In addition, this model is able to identify all single, multiple and overlapping epitopes simultaneously.

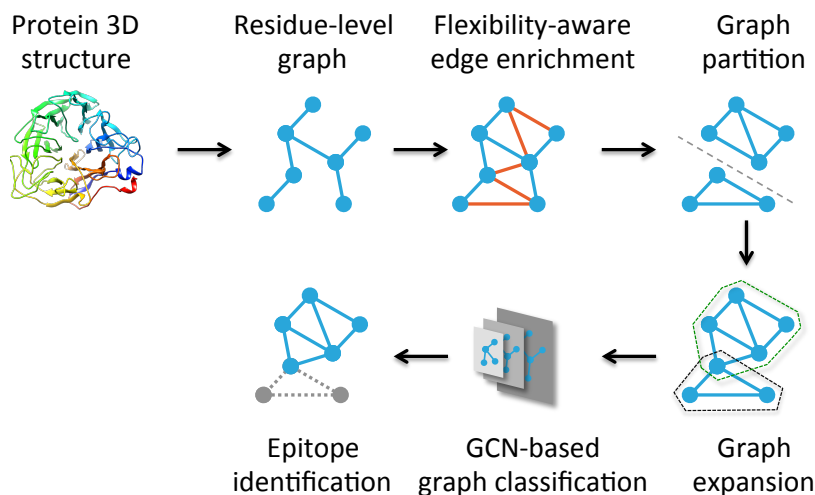


Fig 2. The diagram of the proposed flexibility-aware epitope identification model. It contains graph construction from antigen tertiary structures, flexibility-aware graph enrichment, edge-enriched graph partition, subgraph expansion and graph convolutional network-based classification. The core idea is the inclusion of flexibility in graph construction, which enables big improvement on all the downstream processes.

Materials and methods

Data collection

The antigen-antibody interacting complexes are obtained from published data [6], in which all redundant antigens as well as epitopes are removed, and single, multiple, even overlapping epitopes have been well recorded. This dataset contains 258 antigen-antibody complexes, in which 163 antigens have only single epitope (each antigen has only one epitope), 42 have multiple epitopes (each antigen has more than one separated epitopes) and 53 have overlapping epitopes (each antigen has at least one pair of epitopes that are overlapped with each other).

Flexibility calculation

The conformation of the composing residues of an antigen are not fixed, rather they are quite flexible. This flexibility mainly comes from the rotation of substructures around the covalent bonds [27], particularly the local flexibility (at atom/residue level) [16] and regional flexibility (at intra-domain/multi-residue level) [17], which can be quantified by the torsion angles of residues [28-30]. See Table S1 for the detailed torsion angles of the twenty standard amino acids.

A residue can have several torsion angles based on the number of atoms located at the side chain. For instance, Alanine has no torsion angle, while Proline has five torsion angles; cf. Table S1. A torsion angle is determined by a tuple of four consecutive atoms within a residue [31], in which the angle is the dihedral angle between the two planes formed by three continuously connected atoms. Take the tuple $(C, C_\alpha, C_\beta, C_{\gamma_2})$ in Fig 3 as an example, the torsion angle χ_1 can be considered as the angle between the plane (C, C_α, C_β) and $(C_\alpha, C_\beta, C_{\gamma_2})$. All torsion angles are determined analogously by a sliding a widow of four atoms along a residue.

To compute the oscillation of each atom due to distortion, we build a tree structure in which the root node is the C_α atom, the leaf node is the farthest atom, and the intermediate nodes are the atoms between the two. Note that, there can have multiple

leaf nodes; cf. Fig 3. The content contained in each node is the trajectory of the atom after distorted. Suppose the vector of an arbitrary atom is $\mathbf{v} = \langle x, y, z \rangle$ and the unit vector of the rotation axis is \mathbf{k} , then the rotated vector \mathbf{v}_{rot} of \mathbf{v} is

$$\mathbf{v}_{rot} = \cos\theta\mathbf{v} + (1 - \cos\theta)(\mathbf{v} \cdot \mathbf{k})\mathbf{k} + \sin\theta\mathbf{k} \times \mathbf{v},$$

where θ is the angle of the rotation. By this means, we are able to determine the possible spatial locations of each atom from its static structure after distortion.

Note that, the flexibility of a buried atom is markedly reduced, even to none, after protein folding. Thus, we only calculate the flexibility of the surface atoms in this study. That is, the root node is replaced by the most superficially buried atom that is connecting with the exposed atom during flexibility determination; cf. Fig 3.

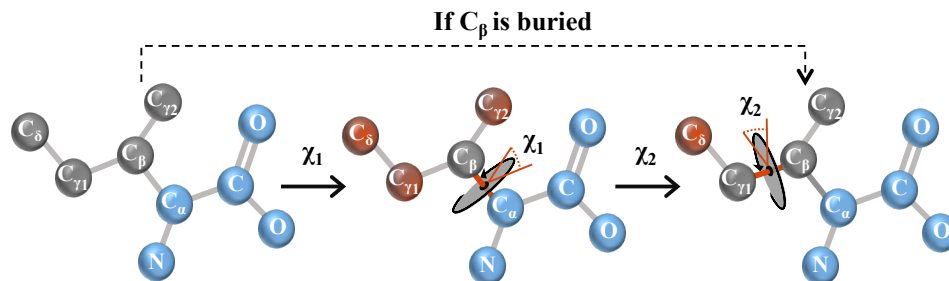


Fig 3. Flexibility calculation illustrated by isoleucine. The possible locations of the atoms at the side chain of the residue are determined by rotating the side chain around axes one-by-one from the C_α to the farthest atom C_δ . The backbone atoms that are almost static are shown in blue, while the side chain atoms are in gray. The atoms to be affected by rotating the side chain around an axle are shown in red. In case C_β is buried, we will skip the torsion of χ_1 .

Model construction

The proposed model is mainly composed of three steps: (i) flexibility-aware graph construction, (ii) overlapping subgraph clustering and (iii) graph classification. The details of each step are as follows.

Step 1: flexibility-aware graph construction

A graph $G = (V, E)$ is generated from the surface atoms of an antigen, where V is the set of nodes (surface atoms) and E is the set of edges (connections) between atoms in V . An atom is deemed as a surface atom if its accessible surface area is no less than 10\AA^2 [12] calculated by NACCESS [32] with the default probe size, while an edge is generated if the Euclidean distance between any two nodes is less than 8\AA [12].

Unlike existing approaches of building the graph from a static structure of an antigen (*a snapshot of huge number of possible states*), we consider its flexibility, and generate additional edges for the graph by distortion; see the section ‘Flexibility calculation’. By this means the edges of the graphs can be highly enriched; see Fig 1(B). For instance, after enrichment on edges of the graph generated from PDB 1A14, the number of edges is increased from 635 to 936. To accelerate the determination of edges, the KD-tree data structure is borrowed to calculate the distance between atoms.

The atom-level graphs are upgraded into residue-level graphs by removing redundant and self-contained edges. An edge is deemed as redundant if there exist more than one edge connecting the two nodes, while the self-contained edges are the ones generated from the atoms of the same residue.

Step 2: overlapping subgraph clustering

The residue-level graph is clustered into overlapping subgraphs via two steps: partition and expansion. Partition splits the whole graph into non-overlapping subgraphs by using the Markov clustering algorithm [25], while the expansion enlarges separated subgraphs into overlapping subgraphs by using the local community detection algorithm DMF-R [26].

Graph partition is conducted on the weighted graph by using MCL [25] with default parameters, in which the weight $w_{i,j}$ of an edge $e_{i,j}$ is calculated as

$$w_{i,j} = \sum_c \alpha_c * w_{i,j}^c, \quad s.t. \sum_c \alpha_c = 1,$$

where

$$w_{i,j}^c = |e_{i,j}^c| / \sum |e_{i',j'}^c|,$$

i, i', j and j' belong to the twenty types of amino acids, c is the type of an edge from ‘epitope’, ‘non-epitope’ or ‘boundary’, and $|x|$ denotes the size of x . Note, boundary edges should be cut off in reality, thus $w_{i,j}^{boundary} = 1 - w_{i,j}^{boundary}$.

The ground truth label of an edge is determined from the aforementioned dataset, where an edge $e_{i,j}^c$ is labeled as ‘epitope’ if and only if both the node i and j are epitopic residues. Similarly, $e_{i,j}^c$ is marked as ‘non-epitope’ if both the node i and j are non-epitopic residues. An edge $e_{i,j}^c$ is deemed as a boundary edge if one node is epitopic and the other is non-epitopic.

Subgraph expansion is carried out on the partitioned subgraphs by using the DMF-R algorithm. Unlike existing community detection algorithms that require global information, DMF-R only takes the local community as input and considers its characteristics as well. Hence, it is suitable to our subgraph expansion problem, which is in line with the nature of spatially proximity of epitopes. This step is critical to identify overlapping epitopes as they have been widely reported [33,34] while graph partition cannot solve this problem intrinsically. To our knowledge, there is only one study aiming at overlapping epitope identification [6].

Note that, the quality of graph partition and expansion are highly related to the way of graph construction, i.e., flexibility-aware or flexibility-agnostic. With flexibility-aware graph construction, hubs may appear; otherwise, all nodes will have evenly distributed degree. See results later.

Step 3: graph classification

Expanded overlapping subgraphs are further classified as epitopes or non-epitopes by a newly designed GCN-based algorithm [35,36], which is composed of two graph convolutional layers and two fully connected layers. The graph convolutional layer accounts for graph embedding, while the fully connected layer is for feature learning.

Let $G = (F, A)$ be an expanded subgraph having $F \in R^{n \times d}$ be the feature matrix and $A \in R^{n \times n}$ be the weighted adjacency matrix, where n is the number of nodes and d is the number of dimensions. The i -th graph convolutional layer is calculated as

$$F^{(i)} = \sigma(A \cdot F^{(i-1)} \cdot W^{(i)}),$$

where $W^{(i)} \in R^{d^{(i-1)} \times d^{(i)}}$ is a trainable weight matrix to be optimized and $\sigma(\cdot)$ is the activation function realized as ReLU [37]. Regarding the output of the j -th fully connected layer, it is computed as

$$F^{(j)} = \sigma(F^{(j-1)} \cdot W^{(j)} + b^{(j)}),$$

where $W^{(j)} \in R^{d^{(j-1)} \times d^{(j)}}$ is the parameter matrix to be learned, $b^{(j)} \in R^{n \times 1}$ is the bias and $\sigma(\cdot)$ is ReLU as well. All the weights $W^{(\cdot)}$ are optimized by the stochastic gradient descent algorithm. 139
140
141

Due to the imbalanced nature of epitopic and non-epitopic subgraphs, we use the focal loss [38] to optimize the training process, which is defined as

$$L(p_t) = -\alpha_t \cdot (1 - p_t)^\gamma \cdot \log(p_t),$$

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{otherwise,} \end{cases}$$

$$\alpha_t = \begin{cases} \alpha & \text{if } y = 1, \\ 1 - \alpha & \text{otherwise,} \end{cases}$$

where $y \in \{0, 1\}$, $p \in [0, 1]$, $\alpha \in [0, 1]$ and γ equals 4. 142

Experimental results 143

Evaluation metrics 144

The F1-score, recall and precision are used to quantify the performance of our model as well as existing models, which are defined as:

$$F1\text{-score} = 2 \cdot \text{Recall} \cdot \text{Precision} / (\text{Recall} + \text{Precision}),$$

$$\text{Recall} = TP / (TP + FN),$$

$$\text{Precision} = TP / (TP + FP),$$

where TP (True Positive) is the number of epitopes fished out correctly, FP (False Positive) is the number of non-epitopes predicted as epitopes incorrectly, and FN (False Negative) is the number of epitopes deemed as non-epitopes. Among these metrics, F1-score is more robust and meaningful as the number of non-epitopes is far larger than that of epitopes. 145
146
147
148
149

Performance qualification 150

Our proposed model is evaluated by using leave-one-out cross validation on the 258 complexes. The averaged F1-score, recall and precision are 0.656 ± 0.138 , 0.587 ± 0.191 and 0.818 ± 0.150 , respectively. Comparing to the exiting best mode Glep [6], the proposed model has made remarkable advancement on epitope prediction, achieving an increment of F1-score by 16.3%. On the same dataset, the F1-score achieved by existing approach ElliPro [39], DiscoTope 2.0 [40], EpiPred [41] and Glep [6] is 0.372, 0.159, 0.353, and 0.564, respectively. The detailed performances are shown in Table 1. 151
152
153
154
155
156
157

Table 1. Performance comparison of epitope prediction between our method and the existing state-of-the-art models.

Method	F1-score	Recall	Precision
ElliPro	0.372 ± 0.183	0.383 ± 0.230	0.456 ± 0.260
DiscoTope 2.0	0.159 ± 0.173	0.262 ± 0.309	0.168 ± 0.215
EpiPred	0.353 ± 0.208	0.474 ± 0.283	0.296 ± 0.188
Glep	0.564 ± 0.135	0.493 ± 0.171	0.722 ± 0.163
Ours	0.656 ± 0.138	0.587 ± 0.191	0.818 ± 0.150

From the table we can see that the performance of our model significantly outperforms existing models. In terms of F1-score, the increment is 76.3%, 312.6%, 158
159

85.8% and 16.3% compared with ElliPro, DiscoTope 2.0, EpiPred and Glep, respectively. Besides, the recall and precision obtained by our model are markedly better than that of others as well. The performances of other models are generated from their source codes with default parameters.

Flexibility enables performance advancement

The key contribution of this study is the incorporation of flexibility into the epitope prediction model. Thus, we carefully examine the impact of flexibility by keeping all other steps unchanged but toggling flexibility only.

On average, the F1-score is 0.656 ± 0.138 and 0.599 ± 0.157 for the flexibility-aware and flexibility-agnostic model, respectively. Comparing to the later, the former one lifts the F1-score by 9.5%. Without flexibility the F1-score of the proposed model is only increased by 0.035 compared to the state-of-the-art models [6]. However, this value is increased to 0.092 in case flexibility is considered, which is around 3 times higher. See Fig 4 for the detailed performance comparison.

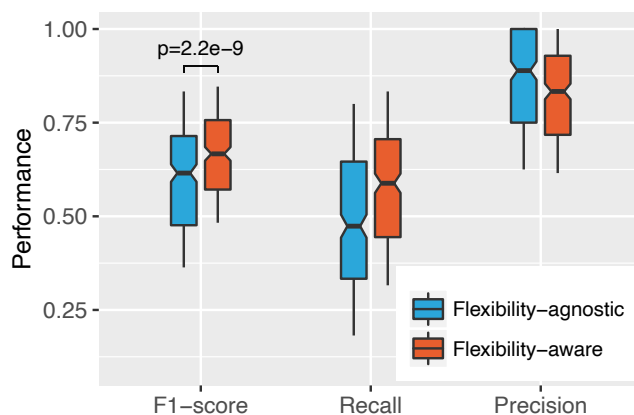


Fig 4. The performance comparison between the flexibility-agnostic and flexibility-aware model.

It is obvious that flexibility has great impact on epitope prediction. We speculate that the impact is mainly from the denser connections of the graphs when flexibility is considered. Based on the data, we found that the averaged degree of each node (residue) was increased from 6.39 ± 1.71 to 9.04 ± 2.58 , achieving an increment of 41.5%; see Fig 5. Interestingly, the edge enrichment favors nodes with medial degrees, which can be observed from the peak of the bottom left panel of Fig 5. For instance, the averaged difference between the top ten nodes having largest discrepancy of degree is 16.6 ± 0.8 , in which the averaged degree of these nodes is 4.9 ± 1.2 and 21.5 ± 0.7 for flexibility-agnostic and flexibility-aware graph, respectively. Notably, these nodes are mainly of Arginine and Lysine, which have been reported as epitope-favored residues [12], indicating that flexibility is particularly helpful to identify epitope-enriched residues. From Fig 5 (the top right panel) we can also see that flexibility enables hubs generation. Hubs are very helpful in graph partition [42], expansion [43] and classification. Without flexibility, the edge-enriched graphs are degraded into ordinary graphs, in which many essential patterns will disappear.

Graph expansion identifies various epitopes

Most existing epitope prediction models are actually epitopic-residue oriented although they are claimed as epitope oriented. This is because the interaction between antigens

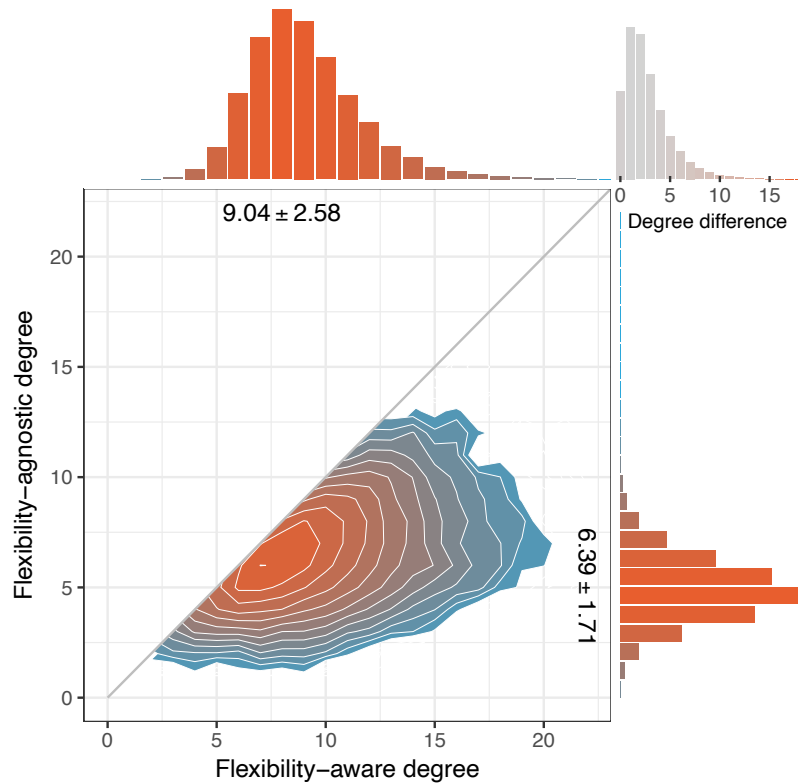


Fig 5. Degree distribution of flexibility-aware and flexibility-agnostic graphs. The color indicates the density of the distribution, of which orange is denser. The degree difference is determined on the same residue with and without flexibility being considered.

and antibodies is not necessarily a one-to-one correspondence. In fact, an antigen can interact with multiple antibodies simultaneously or competitively, resulting in several epitopes either separated or overlapped [44,45]. These epitopes together, however, do not form a big epitope, rather a set of epitopic residues. To solve this problem, we expand subgraphs that are generated from graph partition into larger ones so that shared epitopic residues can be included by more than one epitope, which facilitates the identification of overlapping epitopes.

Among the dataset, there have 53 overlapping epitopes. By using our approach, we can identify 96.2% of them, achieving an averaged F1-score of 0.678 at residue level. Compared to the unexpanded version, the averaged F1-score is increased by 31.6%.

Expansion is also helpful to improve the performance of single and multiple epitopes identification; see Fig 6. Results show that expansion lifts the averaged F1-score by 23.3% and 27.9% for single and multiple epitopes, respectively. Expansion is therefore more helpful for multiple, and overlapping epitopes prediction, particularly the overlapping ones. The F1-score is 0.643, 0.678 and 0.678 for single, multiple and overlapping epitopes identification, respectively; see Fig 6.

GCN guarantees good distinguishability

The expanded subgraphs are finally classified as epitopes or non-epitopes by using a GCN-based model. The parameters of this model are optimized based on 3,289 subgraphs, in which 935 are deemed as epitopes and the rest as non-epitopes. Since

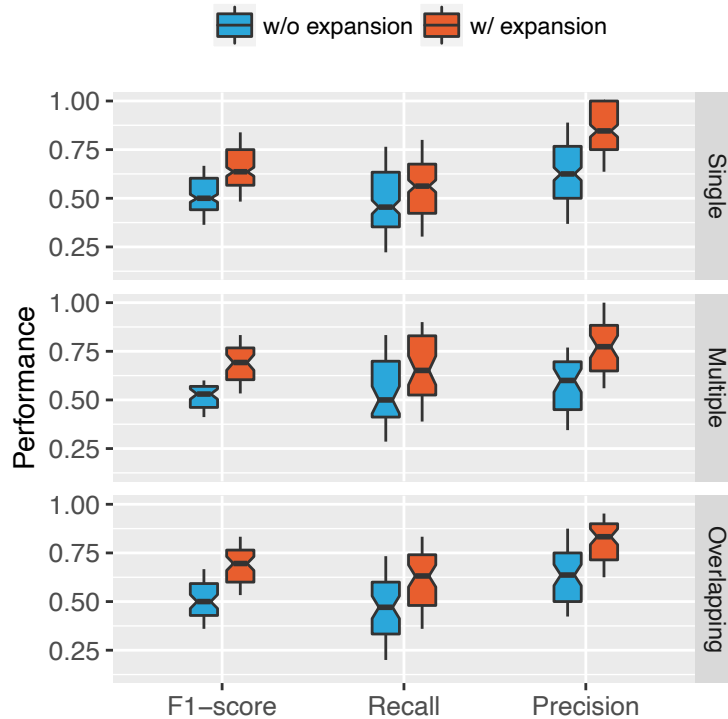


Fig 6. Prediction performance comparison on graphs with and without expansion. The results are broken down into three categories according to the types of epitopes, i.e., single, multiple and overlapping epitopes, to examine the effectiveness of expansion under various circumstances.

graph partition and subgraph expansion are not perfect, some residues within an epitopic graph are not necessarily epitopic residues. Hence, we consider a subgraph as an epitope if the proportion of epitopic residues is no less than 30% during model training.

The training and validation loss as well as the classification accuracy is show in Fig 7. After two rounds of learning rate decrement, the model is converged to a stable minimum. At this state the best classification accuracy of the test set is 83.1%, and the corresponding F1-score is 0.704.

Concluding remarks

Epitopes of an antigen can be neutralized by antibodies, identification of these epitopes can be helpful to various applications, e.g., vaccine design and drug development. Hence, intensive efforts have been made to solve this thorny problem, from both experimental [46, 47] and computational perspectives [48, 49], particularly the later one. Although chemo-physical properties of antigens, such as hydropathy index [50], protrusion index [39] and statistical proximity [51], have been well explored to identify epitopes, all of them are determined from the static structure of antigens, either bonded or unbound. However, the structure of an antigen is not fixed, particularly the side chain of the surface residues [52]. To incorporate such important information, we have proposed a novel graph-based model that is flexibility-aware. Our model starts with residue-level graphs construction from antigens having tertiary structures, and enriches the edges of the graphs via side chain distortion. These edge-enriched graphs are further

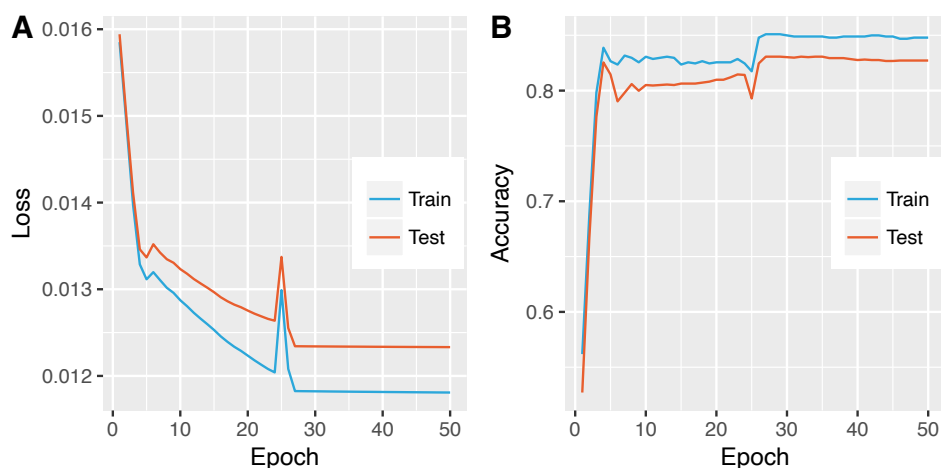


Fig 7. The loss and accuracy of the GCN-based epitope classifier. The model is trained with manually declined learning rate starting from 0.01, and reduced by a degree of 10 in case the loss is unchanged.

partitioned into subgraphs by using MCL [25], and are expanded into overlapping subgraphs by DMF-R [26]. Finally, these expanded graphs are classified into epitopes or non-epitopes by a newly built graph convolutional network. Experimental results show that the proposed approach outperforms all existing models, achieving a high F1-score of 0.656. This score makes a 16.3% increment compared to that of the state-of-the-art models. Besides, our model can identify single, multiple and overlapping epitopes simultaneously. Although there still has notably room to improve the performance, this study pioneers a new direction for improving epitopes identification.

Supporting information

Table S1. The torsion angles of the twenty standard amino acids.

Acknowledgments

This study was collectively supported by the Natural Science Foundation of China [32060150; 31871481], the Natural Science Foundation of Guangxi [2018GXNSFAA281275], the Free Exploration Fund of Hubei University of Medicine [FDFR201805], the Scientific Research Found of Guangxi University [XGZ150316] and Taihe Hospital [2016JZ11].

References

1. Zhang J, Zhao X, Sun P, Gao B, Ma Z. Conformational B-cell epitopes prediction from sequences using cost-sensitive ensemble classifiers and spatial clustering. *BioMed Research International*. 2014;2014:1–12.
2. Qamar MTU, Saleem S, Ashfaq UA, Bari A, Alqahtani S. Epitope-based peptide vaccine design and target site depiction against Middle East Respiratory Syndrome Coronavirus: an immune-informatics study. *Journal of Translational Medicine*. 2019;17(1):1–14.

3. Wu X, Yin Z, McKay C, Pett C, Yu J, Schorlemer M, et al. Protective epitope discovery and design of MUC1-based vaccine for effective tumor protections in immunotolerant mice. *Journal of the American Chemical Society*. 2018;140(48):16596–16609.
4. Wadood A, Mehmood A, Khan H, Ilyas M, Ahmad A, Alarjah M, et al. Epitopes based drug design for dengue virus envelope protein: a computational approach. *Computational Biology and Chemistry*. 2017;71:152–160.
5. Bai H, Li Y, Michael NL, Robb ML, Rolland M. The breadth of HIV-1 neutralizing antibodies depends on the conservation of key sites in their epitopes. *PLOS Computational Biology*. 2019;15(6):1–14.
6. Zhao L, Wu S, Jiang J, Li W, Luo J, Li J. Novel overlapping subgraph clustering for the detection of antigen epitopes. *Bioinformatics*. 2018;34(12):2061–2068.
7. Ofran Y, Schlessinger A, Rost B. Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *The Journal of Immunology*. 2008;181(9):6230–6235.
8. Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner BD, Hu X, Adachi Y, et al. RosettaAntibodyDesign (RABD): A general framework for computational antibody design. *PLOS Computational Biology*. 2018;14(4):1–38.
9. Bansal N, Zheng Z, Merz KM. Advances in In-silico B-cell Epitope Prediction. *Current Topics in Medicinal Chemistry*. 2019;19(2):105–115.
10. Zhao L, Li J. Mining for the antibody-antigen interacting associations that predict the B cell epitopes. *BMC Structural Biology*. 2010;10(S1):S6.
11. Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B-cell epitopes. *Molecular Immunology*. 2009;46(5):840–847.
12. Zhao L, Wong L, Lu L, Hoi SC, Li J. B-cell epitope prediction through a graph model. *BMC Bioinformatics*. 2012;13(S17):S20.
13. Osajima T, Suzuki M, Neya S, Hoshino T. Computational and statistical study on the molecular interaction between antigen and antibody. *Journal of Molecular Graphics and Modelling*. 2014;53:128–139.
14. Tuffery P, Derreumaux P. Flexibility and binding affinity in protein–ligand, protein–protein and multi-component protein interactions: limitations of current computational approaches. *Journal of The Royal Society Interface*. 2012;9(66):20–33.
15. Moraes AH, Simonelli L, Pedotti M, Almeida FC, Varani L, Valente AP. NMR investigation of domain III of Dengue virus E protein: antibody binding modulates conformational exchange in the antigen. *Journal of Virology*. 2015;90(4):1802–1811.
16. Davis IW, Arendall III WB, Richardson DC, Richardson JS. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*. 2006;14(2):265–274.
17. Turton DA, Senn HM, Harwood T, Laphorn AJ, Ellis EM, Wynne K. Terahertz underdamped vibrational motion governs protein-ligand binding in solution. *Nature Communications*. 2014;5(1):3999.

18. Bu Z, Callaway DJ. Proteins move! Protein dynamics and long-range allostery in cell signaling. *Advances in Protein Chemistry and Structural Biology*. 2011;83:163–221.
19. Fenwick RB, van den Bedem H, Fraser JS, Wright PE. Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences*. 2014;111(4):E445.
20. Taylor D, Cawley G, Hayward S. Quantitative method for the assignment of hinge and shear mechanism in protein domain movements. *Bioinformatics*. 2014;30(22):3189–3196.
21. Cousin SF, Kadeřávek P, Bolik-Coulon N, Gu Y, Charlier C, Carlier L, et al. Time-Resolved Protein Side-Chain Motions Unraveled by High-Resolution Relaxometry and Molecular Dynamics Simulations. *Journal of the American Chemical Society*. 2018;140(41):13456–13465.
22. Karch R, Stocsits C, Ilieva N, Schreiner W. Intramolecular Domain Movements of Free and Bound pMHC and TCR Proteins: A Molecular Dynamics Simulation Study. *Cells*. 2019;8(7):720.
23. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*. 2002;58(6):899–907.
24. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. 2004;25(13):1605–1612.
25. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*. 2008;30(1):121–141.
26. Luo W, Zhang D, Jiang H, Ni L, Hu Y. Local community detection with the dynamic membership function. *IEEE Transactions on Fuzzy Systems*. 2018;26(5):3136–3150.
27. Gupta SP. Protein Flexibility: A Challenging Issue of Drug Discovery. *Current Chemical Biology*. 2018;12(1):3–13.
28. Dolenc J, van Gunsteren WF, Protá AE, Steinmetz MO, Missimer JH. Conformational Properties of the Chemotherapeutic Drug Analogue Etoposide A: How to Model a Flexible Protein Ligand Using Scarcely Available Experimental Data. *Journal of Chemical Information and Modeling*. 2019;59(5):2218–2230.
29. Loshbaugh AL, Kortemme T. Comparison of Rosetta flexible-backbone computational protein design methods on binding interactions. *Proteins: Structure, Function, and Bioinformatics*. 2020;88(1):206–226.
30. Ravikumar A, de Brevern AG, Srinivasan N. Conformational Strain Indicated by Ramachandran Angles for the Protein Backbone Is Only Weakly Related to the Flexibility. *The Journal of Physical Chemistry B*. 2021;125(10):2597–2606.
31. IUPAC-IUB commission on biochemical nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry*. 1970;9(18):3471–3479.
32. Hubbard SJ, Thornton JM. NACCESS: program for calculating accessibilities; 1992. Available from: <http://wolf.bms.umist.ac.uk/naccess/>

33. Ciurana CL, Hack CE. Competitive binding of pentraxins and IgM to newly exposed epitopes on late apoptotic cells. *Cellular Immunology*. 2006;239(1):14–21.
34. Wang K, Zheng B, Zhang L, Cui L, Su X, Zhang Q, et al. Serotype specific epitopes identified by neutralizing antibodies underpin immunogenic differences in Enterovirus B. *Nature Communications*. 2020;11(1):4419.
35. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. In: *Proceedings of the Second International Conference on Learning Representations (ICLR)*; 2014. p. 1–14.
36. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: *Proceedings of the Fifth International Conference on Learning Representations (ICLR)*; 2017. p. 1–14.
37. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (ICAIS)*. vol. 15; 2011. p. 315–323.
38. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; 2017. p. 2999–3007.
39. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics*. 2008;9(1):514.
40. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLOS Computational Biology*. 2012;8(12):1–10.
41. Krawczyk K, Liu X, Baker T, Shi J, Deane CM. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*. 2014;30(16):2288–2294.
42. Shiokawa H, Fujiwara Y, Onizuka M. SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-Scale Graphs. *Proceedings of the VLDB Endowment*. 2015;8(11):1178–1189.
43. Bóta A, Krész M. A high resolution clique-based overlapping community detection algorithm for small-world networks. *Informatica*. 2015;39(2):177–187.
44. Gieras A, Cejka P, Blatt K, Focke-Tejkl M, Linhart B, Flicker S, et al. Mapping of conformational IgE epitopes with peptide-specific monoclonal antibodies reveals simultaneous binding of different IgE antibodies to a surface patch on the major birch pollen allergen, Bet v 1. *The Journal of Immunology*. 2011;186(9):5333–5344.
45. Bhandari D, Chen FC, Hamal S, Bridgman RC. Kinetic Analysis and Epitope Mapping of Monoclonal Antibodies to Salmonella Typhimurium Flagellin Using a Surface Plasmon Resonance Biosensor. *Antibodies*. 2019;8(1):22.
46. Carboni F, Adamo R, Fabbrini M, De Ricco R, Cattaneo V, Brogioni B, et al. Structure of a protective epitope of group B Streptococcus type III capsular polysaccharide. *Proceedings of the National Academy of Sciences*. 2017;114(19):5017–5022.

47. Maritan M, Veggi D, Cozzi R, Dello Iacono L, Bartolini E, Lo Surdo P, et al. Structures of NHBA elucidate a broadly conserved epitope identified by a vaccine induced antibody. *PLOS ONE*. 2018;13(8):1–21.
48. Kringelum JV, Nielsen M, Padkjær SB, Lund O. Structural analysis of B-cell epitopes in antibody: protein complexes. *Molecular Immunology*. 2013;53(1-2):24–34.
49. Zheng W, Ruan J, Hu G, Wang K, Hanlon M, Gao J. Analysis of conformational B-cell epitopes in the antibody-antigen complex using the depth function and the convex hull. *PLOS ONE*. 2015;10(8):1–16.
50. Adekiya TA, Aruleba RT, Khanyile S, Masamba P, Oyinloye BE, Kappo AP. Structural Analysis and Epitope Prediction of MHC Class-1-Chain Related Protein-A for Cancer Vaccine Development. *Vaccines*. 2018;6(1):1.
51. Jespersen MC, Mahajan S, Peters B, Nielsen M, Marcatili P. Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. *Frontiers in Immunology*. 2019;10:298.
52. Bansal N, Zheng Z, Merz KM. Incorporation of side chain flexibility into protein binding pockets using MT_{flex} . *Bioorganic & Medicinal Chemistry*. 2016;24(20):4978–4987.

Table S1: The torsion angles of the twenty standard amino acids.

Amino acid	Torsion angle (°)					Reference
	χ_1	χ_2	χ_3	χ_4	χ_5	
GLY	–	–	–	–	–	–
ALA	–	–	–	–	–	–
CYS	–	–	–	–	–	[1]
VAL	21	–	–	–	–	[1]
SER	16	–	–	–	–	[1]
THR	16	–	–	–	–	[1]
ILE	45	47	–	–	–	[2]
LEU	36	40	–	–	–	[2]
ASP	29	71	–	–	–	[2]
ASN	21	59	–	–	–	[2]
HIS	22	56	–	–	–	[2]
TRP	30	38	–	–	–	[2]
PHE	29	71	–	–	–	[1, 2]
TYR	17	46	–	–	–	[1, 2]
GLU	30	39	71	–	–	[1, 2, 3]
GLN	30	39	59	–	–	[1, 2, 3]
MET	36	39	38	–	–	[2]
LYS	29	36	32	40	–	[2]
ARG	31	30	37	40	7	[2]
PRO	26.8	36.5	32.2	17.3	9.8	[4]

Note, the side chain torsion angle must be formed from at least two heavy atoms, thus the GLY and ALA have no torsion angle.

References

- [1] Janin J, Wodak S, Levitt M, Maigret B. Conformation of amino acid side-chains in proteins. *Journal of Molecular Biology*. 1978;125(3):357–386.
- [2] McGregor MJ, Islam SA, Sternberg MJ. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *Journal of Molecular Biology*. 1987;198(2):295–310.
- [3] Bhat T, Sasisekharan V, Vijayan M. An analysis of side-chain conformation in proteins. *International Journal of Peptide and Protein Research*. 1979;13(2):170–184.
- [4] Ho BK, Coutsiias EA, Seok C, Dill KA. The flexibility in the proline ring couples to the protein backbone. *Protein Science*. 2005;14(4):1011–1018.