scientific reports

Check for updates

OPEN

B cell epitope prediction by capturing spatial clustering property of the epitopes using graph attention network

Sungjin Choi✉ & Dongsup Kim✉

Knowledge of B cell epitopes is critical to vaccine design, diagnostics, and therapeutics. As experimental validation for epitopes is time-consuming and costly, many in silico tools have been developed to computationally predict the B cell epitopes. While most methods show poor performance, deep learning methods in recent years have shown promising results. We developed a method called EpiGraph that outperformed previous methods, including those that showed a significant improvement in performance in recent years. Our model's performance can be attributed to the following factors: (1) a combination of structure and sequence feature embeddings obtained from pretrained ESM-IF1 and ESM-2 models could capture the structural and evolutionary features of B cell epitopes, (2) a graph attention network could learn the spatial proximity of B cell epitopes with high graph homophily, and (3) residual connections in the model framework mitigate the over-smoothing problem in the graph neural network. Our model achieved the highest performance on an independent benchmark dataset. The results were also consistent on a different dataset. The datasets and source codes are available at https://github.com/sj584/EpiGraph. A user-friendly web server is freely available at http://epigraph.kaist.ac.kr.

Keywords Conformational, B cell epitope prediction, Graph neural network, ESM

B cells have a vital role in adaptive immunity. B cells provide a specific and long-term immune response against pathogens by generating antibodies. Antibodies neutralize pathogens by recognizing certain residues of the antigen called epitopes. Thus, knowledge of B cell epitopes is fundamental for vaccine design, diagnostics, and therapeutics for enhanced immunological response[1]. However, epitopes can only be acquired through an experimental process, which is usually time-consuming and costly[2]. To address this issue, many in silico tools based on available experimental data have been developed. Computational methods are efficient for obtaining epitope information without the need for an experiment[3–12].

There are two types of B cell epitopes: linear and conformational epitopes. Linear epitopes are continuous in the amino acid sequence, and can therefore be modeled using sequence-based methods[13]. Conformational epitopes, on the other hand, are discontinuous in the primary sequence but closely located in 3D space[14]. Structure-based methods are generally used for conformational epitope prediction. It is known that approximately 90% of B cell epitopes are conformational epitopes[15], and the structure-based methods typically outperform the sequence-based methods.

Recent advances in artificial intelligence in the protein domain are revolutionizing the fields of biology and drug discovery[16,17]. B cell epitope prediction was also accelerated by various deep learning methods[18–27]using the graph neural network (GNN)[28]and the large language model such as evolutionary scale modeling (ESM)[29,30]. As a structure-based method, GNN can capture the structural information of the protein as a molecular graph[31], enhancing feature learning in conformational B cell epitope prediction models.

To date, a number of computational methods have been developed to predict B cell epitopes. However, model performance was hampered by the lack of data and the ambiguity of the epitope characteristics. Even though there were studies reporting the statistical properties of B cell epitopes[15], what clearly differentiates epitopes from non-epitopes is not known. This led to the widespread notion that every surface residue could be an epitope. However, as previously seen in the case of SARS-CoV-2, epitopes are mostly located in the receptor-binding domain rather than all surface regions[32]. This indicates that certain surface residues are more likely to be

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. ✉email: ; kds@kaist.ac.kr

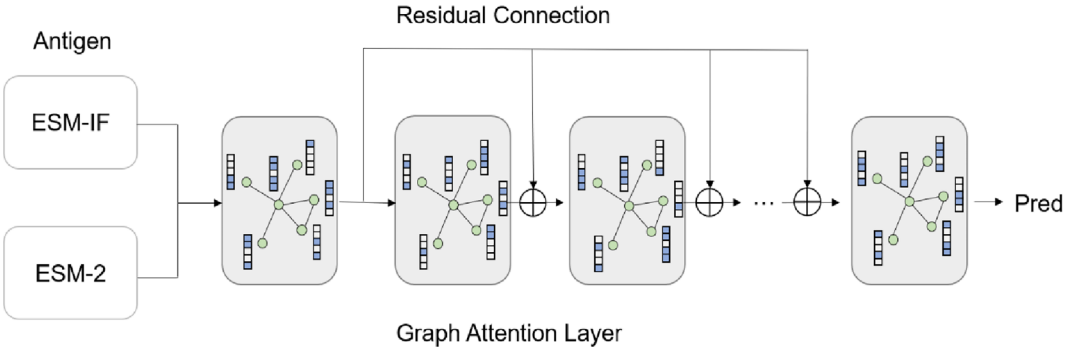

Fig. 1. Model architecture schematics.

Method	F1	MCC	BACC	AUC-ROC
SEPPA 3	0.14	0.02	0.52	0.52
DiscoTope-2.0	0.11	-0.01	0.50	0.49
ElliPro	0.11	-0.06	0.44	0.44
BepiPred-2.0	0.15	0.04	0.55	0.54
epitope3D	0.02	-0.02	0.49	0.49
Bepipred-3.0	0.19	0.08	0.57	0.71
DiscoTope-3.0	0.20	0.09	0.57	0.71
GraphBepi	0.28	0.16	0.62	0.64
EpiGraph	**0.29**	**0.19**	**0.62**	**0.73**

Table 1.. Benchmark result on epitope3D test set.

epitopes than others[33,34]. Based on these facts, we hypothesized that identifying the evolutionary and structural characteristics of epitopes using the pretrained protein model embeddings would be helpful for circumventing the limitation in this field. In addition, epitopes are spatially close to each other, but this property has not been considered in most studies. Using the homophily property of GNN, i.e., the node features of the connected nodes tend to get more similar after each round of GNN layers, we reasoned that this clustering property of B cell epitopes in 3D space could be captured by GNN. As GNN might have an over-smoothing problem where the neighboring node features become too similar as the number of layers increases[35], we adjusted the model architecture to solve it by adding the residual connections.

In this study, we present EpiGraph, which harnesses feature embedding from pretrained protein models and the graph attention network (GAT) model to capture B cell epitope properties. By using the feature embeddings from ESM-IF1[36]and ESM-2[29] pretrained protein models, we were able to capture yet undefined structural and sequential features of B cell epitopes. GAT model with a residual connection could learn the conformational property of the B cell epitope while diminishing the over-smoothing problem in the GAT. EpiGraph could achieve state-of-the-art results on an independent benchmark dataset. The ablation study showed model superiority in terms of feature engineering and model construction.

Results

The overview of the model architecture is presented in Fig. 1. The model predicts epitope probability based on node embedding. Each residue is initially represented by the features extracted from ESM-IF1 and ESM-2 models. ESM-IF1 embedding contains structural features, and ESM-2 embedding contains evolutionary features of the protein. Using these two features, the model could capture the representation of B cell epitopes. The graph attention layer then aggregates the information from neighboring nodes by message passing with attention. In general, the GAT models suffer from over-smoothing problem. To prevent this problem, we utilized residual connections to make the model remember the initial features.

Benchmark on epitope3D dataset

We tested EpiGraph on epitope3D benchmark set in Table 1. Only surface residues were used for model evaluation, as buried residues cannot be the candidate for the binding interface. Surface residues were collected with an RSA threshold of 0.15[37].

Epitope3D was published in 2022 and reported to outperform other models[38]. However, contrary to the paper, we could get close to random results from web server prediction. The result was consistent with other reports[23,32]. We speculated that the model has errors not appropriately considering the surface residues in both training and testing. We further tested recently developed BepiPred-3, DiscoTope-3, GraphBepi, and EpiGraph for benchmark test. For BepiPred-3, DiscoTope-3, and GraphBepi evaluation, we removed the redundancy

Method	AUC-ROC	AUC-PR
Bepipred-3.0	0.730	0.193
DiscoTope-3.0	0.725	0.204
EpiGraph	**0.730**	**0.215**

Table 2. Benchmark result on DiscoTope-3 test set.

Feature Ablation	AUC-ROC	AUC-PR
One-hot	0.53	0.12
ESM-2	0.70	0.18
ESM-IF	0.70	0.21
EpiGraph	**0.73**	**0.23**
Architecture Ablation	AUC-ROC	AUC-PR
MLP	0.69	0.20
w/o residual connection	0.72	0.21
EpiGraph	**0.73**	**0.23**

Table 3.. Ablation Study on epitope3D test set in terms of feature engineering and model architecture.

from the test set with a sequence similarity of 70%. While four models showed significantly better results than previous methods, EpiGraph performed best. The AUC-PR was measured as 0.20, 0.19, 0.24, and 0.23, respectively. In addition, BepiPred-3.0 and DiscoTope-3.0 showed relatively low F1, MCC, and BACC scores compared to threshold-independent scores. We conjectured that two models being trained on the all-residue condition made the difference compared to our model trained on surface residues. However, classification results may not be important because the threshold for the individual case is usually arbitrary. The order of the probability could be a more important criterion, which is represented by the threshold-independent metrics. In the case of GraphBepi, the result showed relatively good scores for classification and AUC-PR, but AUC-ROC score was inferior to other models. When our model was evaluted on the ESMFold-generated models with the same dataset, the score showed a less accurate result (Supplementary Table 1). We inferred that poor quality of some generated structures, with an average plddt score ranging from 0.25 to 0.7, influenced the performance (Supplementary Fig. 1).

Benchmark on DiscoTope-3 dataset

To investigate model superiority in the same condition, we retrained our model on DiscoTope-3 training set and evaluated it on the DiscoTope-3 test set in Table 2. The test set shares less than 20% sequence similarity to the training set. Our model only used the subset of the training set with a 70% sequence similarity cutoff from X-ray crystallography data, while the DiscoTope-3 used both AlphaFold model and X-ray crystal dataset with some redundancy. Same as before, the surface residues with RSA 0.15 were tested. While the AUC-ROC scores for three models were almost the same, AUC-PR scores showed some difference. As the B cell epitope prediction is a highly imbalanced classification task, AUC-PR could give more information when the AUC-ROC is similar. EpiGraph was reported to perform best, while the DiscoTope-3 is better than BepiPred-3. Given that our model used only a portion of the DiscoTope-3 training set, we concluded that EpiGraph has strength compared to DiscoTope-3. Furthermore, our model showed competence independent of the dataset with consistency.

Ablation study

To examine the contributions from feature engineering factors and model architecture factors to the performance, we conducted an ablation study on the test set, and the results are shown in Table 3. Surface residues with RSA 0.15 cutoff were used. Threshold-independent AUC-ROC and AUC-PR were measured.

For feature engineering, we tested four different conditions: (1) baseline one-hot encoding (2) only ESM-2 features were used, (3) only ESM-IF1 features were used, and (4) both ESM-2 and ESM-IF1 features were used. We observed that both features were beneficial for model performance. In terms of AUC-PR, structural features from ESM-IF1 contributed more than evolutionary features from ESM-2. As the final model, we could get better performance when structural and evolutionary features were combined with concatenation.

We further performed an ablation study to examine the model architecture factors. The original model was compared with two subset models without GAT layer or residue connection. To examine the contribution of GAT layer, a simple multi-layer perceptron (MLP) was used to see the difference. The graph attention network model outperforms the MLP model. In addition, we assumed that the graph over-smoothing effect from many graph layers makes neighboring node embeddings too similar to each other. We observed that by adding the residue connection to make the nearby node embeddings different from each other, the model performed better.

B cell epitope spatial proximity modeling using GAT homophily

As conformational B cell epitopes are closely located to each other in 3D space, we tried to impose our model to capture this clustering property. To examine whether our model really captured this inductive bias, we measured

the homophily scores. The following simple equation was used to check the degree of nodes with the same label being adjacent to each other.

$$\text{homophily score} = \frac{1}{|V|}\sum_{v\in V}\frac{Number\ of\ v'\ s\ neighbors\ who\ have\ the\ same\ label\ as\ v}{Number\ of\ v'\ s\ neighbors}$$

where v is a node and $V$is a set of nodes[39]. In this equation, the score of 1.0 means that the graph consists of nodes with the same label. In the antigen graph data where the edges are connected within 10 A distance, the average node homophily ratio for the train set and test set is 0.88, 0.86, respectively. To explore our model capability of modeling conformational B cell epitopes in 3D space, we calculated the homophily score from the prediction result. We compared the homophily scores of top-k from EpiGraph, MLP model, and random sampling.

We observe in Fig. 2 that homophily scores from EpiGraph keep decreasing but stay high with increasing top-k. Compared to MLP model and the random sampling cases, the homophily scores from EpiGraph are much higher, indicating that our model induced the desired inductive bias to capture the spatial proximity of the B cell epitopes. We also visualized the prediction result from the test set using PyMol[40] in Fig. 3. We sampled various cases where prediction accuracy is good, random, or bad. In all cases, we observed that the residues predicted to have a high epitope probability are close to each other. We concluded that the homophily-inducing property of GAT forces nodes with the same label to stay spatially close together, which is beneficial in conformational B cell epitope prediction.

Discussion

In this study, we showed that our approach of using ESM-based pretrained protein models and GAT model with residual connections enabled model improvement in terms of feature engineering and model construction. Despite its strengths, our model has some limitations. First, EpiGraph employed an antibody-agnostic prediction approach. As the model could predict the epitopes overlapped by several antibodies, it could not be used to predict epitopes specific to particular antibodies. To address this issue, antibody-specific prediction tools have been developed, but they are far from being useful[32]. Even the recently developed model, SEPPA-mAb[41],was reported to have little sensitivity to antibody information. It is expected that antibody-specific prediction methods will be further developed for practical use. Second, as threshold could be arbitrary for individuals, there could be some cases where almost all residues are labeled as either epitope or non-epitope. In this regard, the order of the epitope probability could be more relevant than the current threshold. Third, although the performance of our model is better than the other previous methods, the prediction accuracy still needs to be improved.

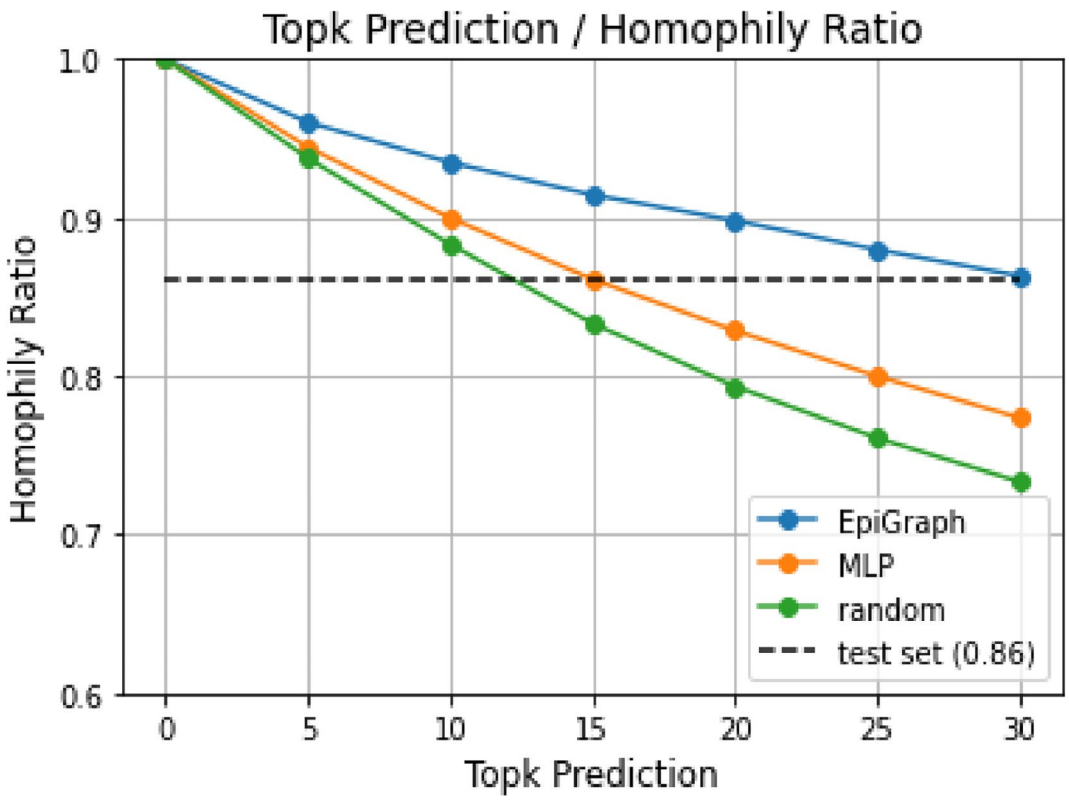

Fig. 2. Graph homophily plot from prediction result on the test set. Homophily ratio from the top-k prediction compared to MLP model and randomly sampled graph data are shown.

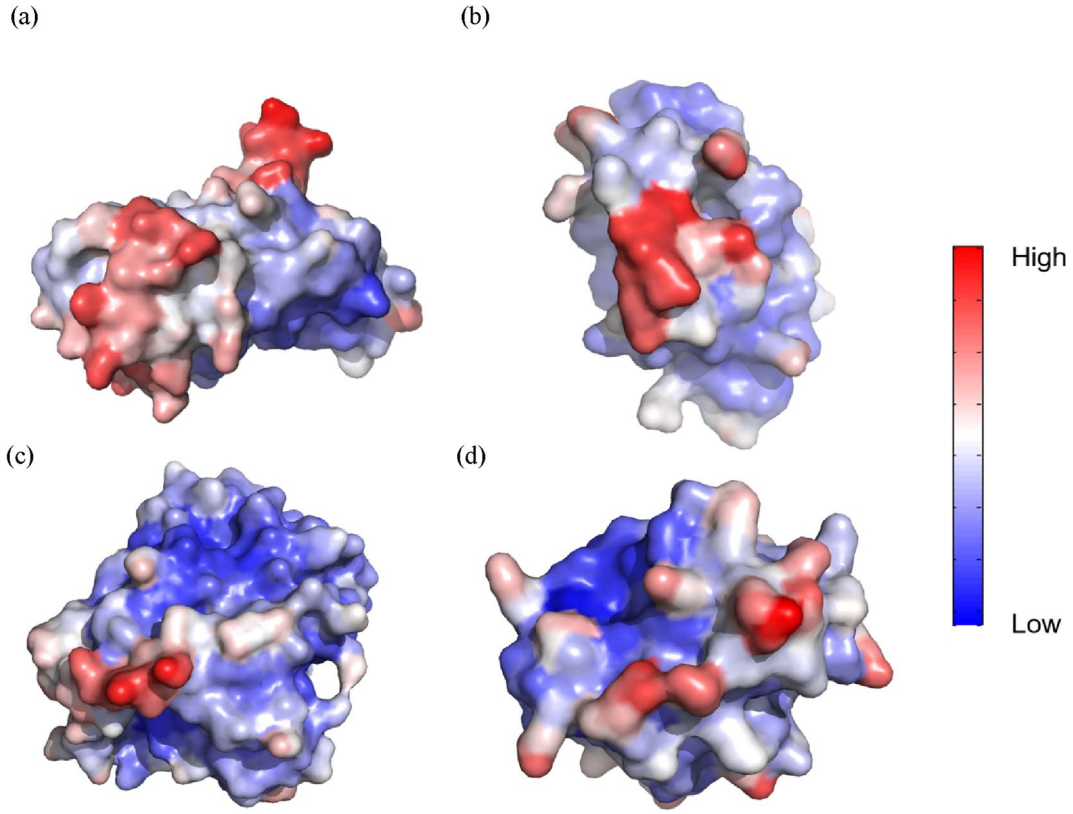

Fig. 3. Individual prediction visualization from the test set using PyMol. Residue with high probability is marked as red and residue with low probability is marked as blue. (a) 1DQT, AUC 0.92 (b) 2B4J, AUC 0.75 (c) 4E9O, AUC 0.52 (d) 4FNZ, AUC 0.17.

The reason for the insufficient accuracy of our model is partially due to potential errors associated with precise annotation of epitopes/non-epitopes in the available datasets. It is often observed that for homo-multimeric antigen structures, because epitope annotation procedures did not consider symmetry information properly, only one of many equivalent residues in different chains is labeled as an epitope. Moreover, annotation on the negative class biases to label potential epitopes as non-epitopes, given the complex data. Positive-unlabeled (PU) learning could be used to detour false-negative annotation problems[42]. Lastly, some portion of residues labeled epitopes was found to be buried residues for a given RSA threshold. We suggest that either the annotation method or surface measurement be modified in the future. An alternative approach could be RSA-based annotation[32]or residue depth-based surface measurement[43]. In vaccine design, the identification of pathogen surfaces targeted by B cells is important. With in silico prediction tools, researchers could efficiently design and produce vaccines that contain specific regions to stimulate the production of antibodies by the immune system[44]. EpiGraph could expedite the antibody design by providing the potential epitopes with high accuracy. We anticipate our model to be used for future vaccine design and antibody engineering applications.

Materials and methods

Dataset

The dataset was taken from epitope3D[38]. Data curation methodology was the same as the previous study[45]. Specifically, antibody-antigen complexes were collected from the protein data bank (PDB)[46]up until May 2021. Complex structures were filtered by resolution 3Å and chain size 25 residues cutoff. Epitopes were identified based on the distance. Antigen residues with at least one heavy atom within 4Å distance from the antibody residues were labeled as epitopes. Unbound structures were collected based on the bound structures used for epitope identification. Unbound structures were filtered only to have 100% structural alignment and a minimum of 70% sequence similarity to bound structures. To reduce redundancy, the unbound structures were screened to have less than 70% sequence similarity among them, resulting in 245 antigens. We trained the model with 200 antigens and evaluated on the test set with 45 antigens. We only used the chain that contains epitopes from the multi-chain complex. An imbalanced dataset that sampled all surface residues was used. For training, surface residues were collected from each structure with a relative surface accessibility (RSA) value of 0.10 cutoff. RSA is a measure of exposure to the surface. RSA was calculated by DSSP[47].

For DiscoTope-3[24]benchmark, we retrained the model on DiscoTope-3 training set. The training set has 1406 chains from 582 PDBs in both Alphafold-generated structure and crystal structure. Epitopes were annotated in the same way as epitope3D dataset. After removing sequence similarity with 70% sequence similarity cutoff

using CD-HIT[48], nonredundant 397 chains from X-ray crystallography were used for training. After training the model on a nonredundant dataset, the model was evaluated on the test set. The test set comprises 24 chains with less than 20% sequence similarity with the training set.

Node representation

Nodes were represented by two different ESM-based pretrained protein models. ESM-IF1 feature extraction from "*esm_if1_gvp4_t16_142M_UR50*" model was used for structural features. ESM-2 feature extraction from "*esm2_t33_650M_UR50D*" model with per-residue representation was used for evolutionary features. Each embedding has dimensions of 512 and 1280, respectively. Two features were concatenated to have an initial node feature representation with 1792 dimensions.

Evaluation

The model was trained on epitope3D 200 antigens with 10-fold cross-validation and evaluated on benchmark sets. After training, 10 models were saved by monitoring validation loss and validation area under the curve of receiver operating characteristics (AUC-ROC) from each fold. Only when both metrics was improved, the model was updated. Every model predicts the per-residue epitope probability on the test set. The average value from each fold becomes the final prediction. AUC-ROC, area under the curve of precision and recall (AUC-PR), F1, Matthews correlation coefficient (MCC), and balanced accuracy (BACC) scores were used to examine the model performance. As the dataset has an imbalanced class ratio, AUC-PR, MCC, and BACC scores were employed. These metrics enabled the robust assessment of the model performance[49]. A threshold that maximized the MCC value in the holdout set was chosen in each fold. In the final prediction, the average threshold was used.

Web server from epitope3D, DiscoTope-3, BepiPred-3, and GraphBepi

The test set was used in each web server. Prediction from the web server was conducted with default settings. Epitope3D results only showed binary classification results; "Prediction" instead of score, "Prediction" was used for evaluation. For DiscoTope-3[24], "DiscoTope-3.0_score" was used for calculating AUC, AUC-PR, and "epitope" for binary classification. For BepiPred-3[25], "BepiPred-3.0 score" from "raw_output.csv" was used for calculating AUC, AUC-PR, and FASTA file "Bcell_epitope_preds.fasta" for binary classification. In the FASTA file, predicted epitopes are denoted as capital letters, while others are non-capital letters. For GraphBepi, "score" was used for calculating AUC, AUC-PR, and "is epitope" for binary classification.

Hyperparameters

Hyperparameters were tuned to have the lowest validation loss. We used epoch 50, batch_size 8, binary cross-entropy loss, adam optimizer, 10-fold cross-validation, hidden dimension 128, 8 multi-attention heads, 8 layers, learning rate 1e-5, weight decay 1e-8, and elu activation function.

Implementation of Web Server

We utilized a Web Server Gateway Interface (WSGI) connection between the Nginx (https://www.nginx.com) web server and the Django (https://www.djangoproject.com) web application framework via Gunicorn (https://gunicorn.org). Front-end was made by HTML and CSS.

Data availability

Preprocessed datasets are available in https://github.com/sj584/EpiGraph. The original dataset of epitope3D and DiscoTope-3 is available in https://biosig.lab.uq.edu.au/epitope3d/data and https://services.healthtech.dtu.dk/services/DiscoTope-3.0/, respectively. The contact detail for data access request is csungjin@kaist.ac.kr.

Received: 4 August 2024; Accepted: 31 October 2024
Published online: 11 November 2024

References

1. Potocnakova, L., Bhide, M. & Pulzova, L. B. An introduction to B-cell epitope mapping and in silico epitope prediction. *Journal of immunology research* (2016). (2016).
2. El-Manzalawy, Y. & Honavar, V. Recent advances in B-cell epitope prediction methods. *Immunome Res.* **6**, 1–9 (2010).
3. Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput. Biol.* **8**, e1002829 (2012).
4. Ponomarenko, J. et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinform.* **9**, 1–8 (2008).
5. Sweredoski, M. J. & Baldi, P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics.* **24**, 1459–1460 (2008).
6. Rubinstein, N. D., Mayrose, I., Martz, E. & Pupko, T. Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinform.* **10**, 1–6 (2009).
7. Sela-Culang, I., Ashkenazi, S., Peters, B. & Ofran, Y. PEASE: predicting B-cell epitopes utilizing antibody sequence. *Bioinformatics.* **31**, 1313–1315 (2015).
8. Ansari, H. R. & Raghava, G. P. Identification of conformational B-cell epitopes in an antigen from its primary sequence. *Immunome Res.* **6**, 1–9 (2010).
9. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* **45**, W24–W29 (2017).
10. Zhou, C. et al. SEPPA 3.0—enhanced spatial epitope prediction enabling glycoprotein antigens. *Nucleic Acids Res.* **47**, W388–W394 (2019).
11. Liang, S. et al. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. *BMC Bioinform.* **11**, 1–6 (2010).
12. Krawczyk, K., Liu, X., Baker, T., Shi, J. & Deane, C. M. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics.* **30**, 2288–2294 (2014).

13. Ponomarenko, J. V. & Van Regenmortel, M. H. B cell epitope prediction. *Struct. Bioinf.* **2**, 849–879 (2009).
14. Sanchez-Trincado, J. L., Gomez-Perosanz, M. & Reche, P. A. Fundamentals and methods for T-and B-cell epitope prediction. *J. Immunol. Res.* **2017** (2017).
15. Kringelum, J. V., Nielsen, M., Padkjær, S. B. & Lund, O. Structural analysis of B-cell epitopes in antibody: protein complexes. *Mol. Immunol.* **53**, 24–34 (2013).
16. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **373**, 871–876 (2021).
18. Park, M., Seo, S., Park, E. & Kim, J. EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. *bioRxiv*, 2002. 2027.481241 (2022). (2022).
19. Collatz, M. et al. EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics*. **37**, 448–455 (2021).
20. Shashkova, T. I. et al. SEMA: Antigen B-cell conformational epitope prediction using deep transfer learning. *Front. Immunol.*, 5272 (2022).
21. Del Vecchio, A., Deac, A., Liò, P. & Veličković, P. Neural message passing for joint paratope-epitope prediction. *arXiv preprint arXiv:2106.00757* (2021).
22. Pittala, S. & Bailey-Kellogg, C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*. **36**, 3996–4003 (2020).
23. Zeng, Y. et al. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics*. **39**, btad187 (2023).
24. Høie, M. H. et al. DiscoTope-3.0-Improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations. *bioRxiv*, 2002. 2005.527174 (2023). (2023).
25. Clifford, J. N. et al. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci.* **31**, e4497 (2022).
26. Ivanisenko, N. V. et al. SEMA 2.0: web-platform for B-cell conformational epitopes prediction using artificial intelligence. *Nucleic Acids Res.*, gkae386 (2024).
27. Israeli, S. & Louzoun, Y. Single-residue linear and conformational B cell epitopes prediction using random and ESM-2 based projections. *Brief. Bioinform.* **25**, bbae084 (2024).
28. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
29. Lin, Z. et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* (2022).
30. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118, e2016239118 (2021).
31. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
32. Cia, G., Pucci, F. & Rooman, M. Critical review of conformational B-cell epitope prediction methods. *Brief. Bioinform.* **24**, bbac567 (2023).
33. Kunik, V. & Ofran, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng. Des. Sel.* **26**, 599–609 (2013).
34. Angeletti, D. et al. Defining B cell immunodominance to viruses. *Nat. Immunol.* **18**, 456–463 (2017).
35. Li, Q., Han, Z. & Wu, X. M. in *Proceedings of the AAAI conference on artificial intelligence.*
36. Hsu, C. et al. in *International Conference on Machine Learning.* 8946–8970 (PMLR).
37. Rost, B. & Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Bioinform.* **20**, 216–226 (1994).
38. da Silva, B. M., Myung, Y., Ascher, D. B. & Pires, D. E. epitope3D: a machine learning method for conformational B-cell epitope prediction. *Brief. Bioinform.* **23**, bbab423 (2022).
39. Pei, H., Wei, B., Chang, K. C. C., Lei, Y. & Yang, B. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:05287* (2020). (2002).
40. Schrodinger, L. The PyMOL molecular graphics system. *Version*. **1**, 0 (2010).
41. Qiu, T. et al. SEPPA-mAb: spatial epitope prediction of protein antigens for mAbs. *Nucleic Acids Res.*, gkad427 (2023).
42. Li, F. et al. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Brief. Bioinform.* **23**, bbab461 (2022).
43. Chakravarty, S. & Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*. **7**, 723–732 (1999).
44. Sunita, Sajid, A., Singh, Y. & Shukla, P. Computational tools for modern vaccine development. *Hum. Vaccines Immunotherapeutics*. **16**, 723–735 (2020).
45. Ren, J., Liu, Q., Ellis, J. & Li, J. Positive-unlabeled learning for the prediction of conformational B-cell epitopes. *BMC Bioinform.* **16**, 1–15 (2015).
46. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
47. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Res. Biomolecules*. **22**, 2577–2637 (1983).
48. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150–3152 (2012).
49. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 1–13 (2020).

Acknowledgements

We would like to thank Woosung Jeon for providing helpful comments on building a web server.

Author contributions

Sungjin Choi; Methodology, Investigation, Visualization Writing-Original Draft, Visualization Dongsup Kim; Writing-Review and editing, Validation, Suvervision.

Funding

This work was supported by Korea Advanced Center of Vaccine Development (KAVAD) grant funded by Korean Government N06240065 and National Research Foundation of Korea (NRF) grants funded by Korean Government N01240404 and N01240859.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78506-z.

Correspondence and requests for materials should be addressed to S.C. or D.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.