Computational and Structural Biotechnology Journal 23 (2024) 2122–2131

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

WUREN: Whole-modal union representation for epitope prediction

Xiaodong Wang [a], Xiangrui Gao [a], Xuezhe Fan [a], Zhe Huai [a], Genwei Zhang [a], Mengcheng Yao [a], Tianyuan Wang [a], Xiaolu Huang [b], Lipeng Lai [a,*]

[a] XtalPi Innovation Center, Beijing, China
[b] XtalPi Investment, Suzhou, China

ARTICLE INFO

Keywords:
Deep learning
Multimodal model
B-cell epitope prediction

ABSTRACT

B-cell epitope identification plays a vital role in the development of vaccines, therapies, and diagnostic tools. Currently, molecular docking tools in B-cell epitope prediction are heavily influenced by empirical parameters and require significant computational resources, rendering a great challenge to meet large-scale prediction demands. When predicting epitopes from antigen-antibody complex, current artificial intelligence algorithms cannot accurately implement the prediction due to insufficient protein feature representations, indicating novel algorithm is desperately needed for efficient protein information extraction. In this paper, we introduce a multimodal model called WUREN (Whole-modal Union Representation for Epitope predictioN), which effectively combines sequence, graph, and structural features. It achieved AUC-PR scores of 0.213 and 0.193 on the solved structures and AlphaFold-generated structures, respectively, for the independent test proteins selected from DiscoTope3 benchmark. Our findings indicate that WUREN is an efficient feature extraction model for protein complexes, with the generalizable application potential in the development of protein-based drugs. Moreover, the streamlined framework of WUREN could be readily extended to model similar biomolecules, such as nucleic acids, carbohydrates, and lipids.

1. Introduction

Epitopes are specific amino acid regions on antigens and capable of antibody binding, and B-cell epitope predictions play a critical role in the development of antibody drugs [1]. Currently, a range of methods exist for identifying or predicting B-cell epitopes, both experimentally and computationally. For example, the structure of antigen-antibody complexes can be directly determined through cryo-electron microscopy, X-ray crystallography, and nuclear magnetic resonance (NMR) techniques. Nevertheless, these methods have their inherent limitations. Cryo-EM, on average, requires weeks to months time and costs over several thousands dollars to obtain a protein complex [2]. X-ray crystallography necessitates protein crystallization as the initial step, which sometimes can be experimentally challenging or infeasible [3]. While NMR is only suitable for small complexes with molecular weights below 40 kDa [4]. Of note, these methods often exhibit low success rates in resolving complex protein structures [5].

As a well-known computational approach, molecular docking tools are frequently employed for B-cell epitope prediction. Two commonly used molecular docking tools for this purpose are SnugDock and Attract [6,7]. However, these methods are susceptible to the influence of empirical parameters [8], and there is still room for improvement in terms of B-cell epitope prediction accuracy.

There is a wide range of machine learning tools utilized for B-cell epitope prediction, which can be broadly classified into two main categories: sequence-based and structure-based tools. Sequence-based tools predict B-cell epitope using the input antigen sequence. These tools utilize amino acid features extracted by tools like Biopython and/or representations generated by protein language models like ESM for constructing machine learning models [9,10]. Notable examples of such tools include BepiPred 3.0, CBTOPE, and SEPIa [11–13]. Among these sequence-based approaches, BepiPred 3.0 currently achieves state-of-the-art performance [11].

On the other hand, structure-based tools using the input antigen structure for B-cell epitope prediction. Examples of such tools include DiscoTope 3.0, SEMA, Epitope3D, PEPITO, EPCES, EPSVR, ElliPro and SEPPA 3.0 [14–20]. To assess the tools' applicability, some structure-based approaches are tested with non-crystal structures, often utilizing structures generated by AlphaFold. Among the structure-based tools, DiscoTope 3.0 currently achieves state-of-the-art performance

* Corresponding author.
E-mail address: lipeng.lai@xtalpi.com (L. Lai).

https://doi.org/10.1016/j.csbj.2024.05.023
Received 26 January 2024; Received in revised form 14 May 2024; Accepted 14 May 2024
Available online 16 May 2024

[15]. DiscoTope 3.0 employs the ESM-IF1 model to extract amino acid features and incorporates solved structures and generated structures in the training set to mitigate the impact of structural perturbations [10]. Additionally, the model employs a positive-unlabeled strategy to address the uncertainty of negative labels.

In addition to the work mentioned above, there are some other important research findings in the field of B-cell epitope prediction. CluSMOTE utilizes the Synthetic Minority Over-sampling Technique (SMOTE) algorithm to tackle the problem of imbalanced positive and negative samples in B-cell epitope prediction [21]. On the other hand, Mimotope divides antigen surfaces into overlapping patches and predicts epitopes by leveraging Amino Acid Pairs (AAPs) derived from mimotopes and the corresponding surface patch [22]. This approach introduces a fresh perspective on B-cell epitope prediction.

Recent advances in artificial intelligence and deep learning have illuminated alternative approaches for B-cell epitope prediction. AlphaFold2 demonstrates the abundance of information that can be extracted from protein sequences alone [23]. Subsequent work, such as AF2-Multimer and xTrimo-Multimer [24,25], has extended the application of protein structure prediction to protein complexes and multimers. PECAN predicts antigen-antibody binding surfaces using graph features [26]. ScanNet and PINet are two tools that utilize deep geometric learning algorithms for predicting protein-protein interactions [27,28]. Given that antigen-antibody interactions are a specific type of protein-protein interaction, ScanNet and PINet can also be utilized for B-cell epitope prediction. The PINet paper reports its testing results on the B-cell epitope prediction benchmark, EpiPred. These examples demonstrate that different features, such as sequence, graph, and structure features, can effectively characterize proteins from different aspects. But an important question to ask is whether these features encode the same information or if their combination can further improve model performance.

Theoretically, protein sequences contain all the information needed to determine protein structures, dynamics, and functions. However, when only limited training data is available, additional information may be helpful in enhancing the learning process and, consequently, improving model performance. As demonstrated in previous protein structure prediction methods like AlphaFold2, DCA-fold, and RaptorX [22,29,30], incorporating evolutionary information through Multiple Sequence Alignment (MSA) increases prediction accuracy. In this study, we show that graphic/topological (graph) and spatial (structure) information complement sequential features, and when they are either fully or partially merged, model performance achieves significant improvement in performing relevant tasks.

Implementing effective multimodal feature fusion is challenging. In most cases, the difficulties lie in assessing the confidence and correlation of each modality and realizing multimodal feature alignment and data registration. In this paper, we develop a method called WUREN (Whole-modal Union Representation for Epitope predictioN), named after a famous Chinese dessert made from a mix of various ingredients. This model is primarily applied to the characterization of protein complexes, such as the antigen-antibody complex (depiction in Fig. 1A). As illustrated in Fig. 1B, without loss of generality, Transformer [31], Graph Convolution Neural Network (GCN), and PointNet+ + are employed in WUREN to extract sequence, graph, and structure information, respectively [32,33]. In practical implementation, these methods can be replaced by other functionally similar methods, such as PointNet or Graph Attention Network (GAT) [34–37], within the WUREN framework, depending on the requirements of downstream tasks.

WUREN demonstrates a state-of-the-art (SOTA) AUC-PR (Area Under the Precision Recall Curve) result of 0.462 on the B-cell epitope prediction benchmark EpiPred and surpasses multimer methods on the recently released SAbDab dataset [38,40]. WUREN achieved AUC-PR values of 0.213, 0.217, and 0.193 on the solved structure, the structure after energy minimization using the Foldx [41], and the structure generated by AlphaFold, respectively. Surpassing other tools except for DiscoTope 3.0. Ablation experiments confirm that features from each dimension in the model play an indispensable role. Our findings demonstrate that the multimodal model WUREN proposed in this paper is a general and efficient model for protein representation, with potential applications in protein complex-related research and the development of protein-based drugs. Moreover, the general idea or framework of WUREN could potentially be applied to model other biomolecules, such as DNA and RNA.

2. Materials and methods

2.1. Problem statement

We chose the task of B-cell epitope prediction as an example to showcase WUREN's performance. The objective is to predict the region on the antigen surface that binds to the antibody, measured in amino acid units. For data processing, we employ various algorithms and tools to extract physical, chemical, and structural features of antigens and antibodies at the amino acid and point cloud levels, respectively. To determine the labels, we first calculate the distance between each point cloud of the antigen and the antibody, marking point clouds with a distance less than 2 Å as 1, indicating that the point cloud belongs to the

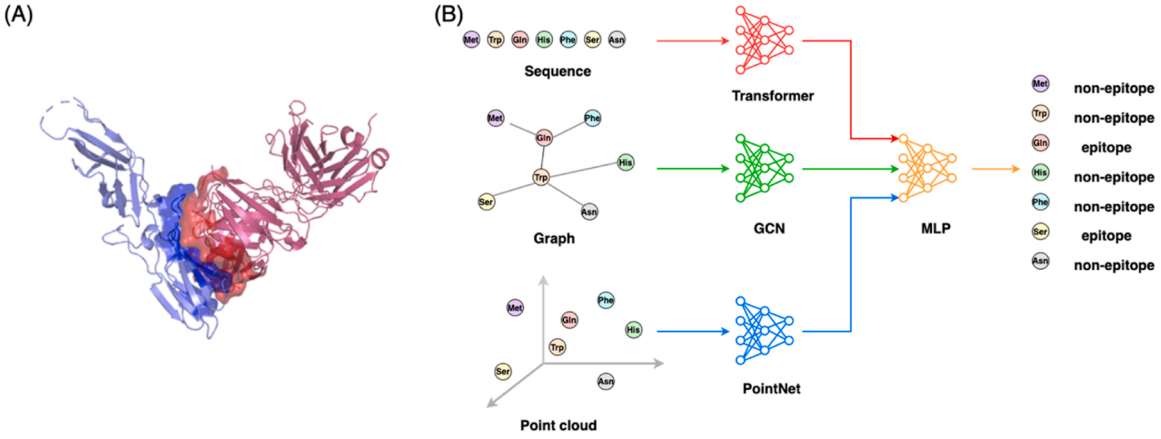

Fig. 1. Schematic diagram of the B-cell epitope prediction process. (A) Schematic representation of the antigen-antibody complex interface, which includes the epitope – the region on the antigen specifically bound by the antibody. The example diagram shows a complex (PDB ID: 1ahw) between the extracellular domain of human tissue factor (left) and the antibody (right). (B) Schematic representation of the WUREN model framework. WUREN consists of a transformer, GCN, and PointNet, effectively fusing sequential, topological, and structural features.

binding surface, while other point clouds are marked as 0. Similarly, we calculate the distance between each amino acid of the antigen and the antibody. Amino acids with a distance less than 4.5 Å are marked as 1, indicating that they belong to the epitope region, while the remaining amino acids are marked as 0. Finally, we use the point cloud data to train the deep learning model, then apply a 2 Å threshold to aggregate the obtained point cloud embeddings into amino acid features. We combine the prepared amino acid features to train the final model and predict the probability that each amino acid belongs to the epitope region, thus completing the B-cell epitope prediction.

2.2. Data collection

The EpiPred dataset [38], the BM dataset [39], and the SAbDab dataset are three commonly used antigen-antibody complex datasets [40]. The EpiPred dataset is a widely used benchmark for B-cell epitope prediction. In this study, we use the EpiPred dataset as a benchmark and conduct a parallel comparison between WUREN and similar methods such as Attract, EpiPred, PECAN, and PINet, as well as an ablation experiment for the feature module. The EpiPred dataset consists of 148 antigen-antibody complexes, with 118 used for training and 30 used for testing. The sequence similarity of antigens in the training and validation sets is less than 90 %. It should be noted that in this study, we use Blast as a sequence similarity calculation tool and identity as a similarity evaluation metric [42]. Supplementary Table.S1 lists the PDB codes for the 118 protein complexes used for training and Supplementary Table. S3 lists the PDB codes for the 30 protein complexes used for testing.

Since the EpiPred paper did not use a validation set, we use the BM dataset as an independent validation set. This dataset is composed of 44 antigen-antibody complexes, with sequence similarity of antigens in this dataset and the EpiPred test set both less than 25 %. Supplementary Table.S2 lists the PDB codes for the 44 protein complexes.

To provide a more comprehensive assessment of WUREN's performance, we collect 636 newly published antigen-antibody complex data from the SAbDab database between 2021 and 2022. To ensure reliable test results, we remove 288 data points with antigens being SARS-CoV-2 proteins (redundant epitopes), and further excluded data with antigen sequence similarity greater than 50 % to the training set. Finally, we select 77 entries as the test dataset, Supplementary Table.S5 lists the PDB codes for the 77 protein complexes, along with the antigen and antibody chain identifiers. On this dataset, we compared the B-cell epitope prediction performance of WUREN with protein complex structure prediction models AF2-Multimer and xTrimo-Multimer.

To facilitate a comprehensive comparison between WUREN and PINet [28], PECAN [26], DiscoTope3 [14], BepiPred3 [11], Epitope3D [16], EPCES [18], EPSVR [18], and ElliPro [19], we have gathered the DiscoTope3, SEDB, Discotope, Epitome, and SEMA datasets. The DiscoTope3 dataset comprises 24 data points originates from the DiscoTope3 paper, each categorized into one of three groups: structures obtained after energy minimization using Foldx, structures solved through experimental methods, or structures generated by AlphaFold. We have designated them as DiscoTope3_Foldx, DiscoTope3_Solved, and DiscoTope3_Af2, respectively. Supplementary Table.S7 presents the PDB codes of the 24 protein complexes in the DiscoTope3 dataset. Through a thorough comparison, we have confirmed that none of these 24 data points were included in the training or validation sets of our study. The SEDB dataset originates from the SEDB paper [43]. Initially, we selected 272 antigen-antibody complex data points and subsequently compared them with the training and validation sets of the aforementioned tools. We have excluded any data points that appeared in the training or validation sets of these tools, resulting in a final SEDB dataset consisting of 89 data points. Supplementary Table.S8 presents the PDB codes of the 89 protein complexes in the SEDB dataset. The Discotope dataset, sourced from the Discotope Supplementary Materials [44], encompasses 75 antigen-antibody complexes from 25 different protein families. Similar to the previous datasets, we compared them with the training and validation sets of the aforementioned tools and eliminated any overlapping data points. As a result, the final Discotope dataset comprises 56 data points. Supplementary Table.S9 presents the PDB codes of the 56 protein complexes in the Discotope dataset. The Epitome dataset, obtained from the Epitome paper [45], includes 140 antigen-antibody complexes from 35 different protein families. By comparing them with the training and validation sets of the aforementioned tools, we have excluded any data points that appeared in those sets. Consequently, the final Epitome dataset consists of 78 data points. Supplementary Table.S10 presents the PDB codes of the 78 protein complexes in the Epitome dataset. The SEMA dataset, derived from the SEMA paper [15], encompasses 103 antigen-antibody complex data points. Importantly, none of these data points were present in the training or validation sets of the aforementioned tools. Supplementary Table.S11 presents the PDB codes of the 103 protein complexes in the SEMA dataset.

2.3. Features

We extract point cloud and amino acid features of antigens and antibodies.

2.3.1. Point cloud features

First, we process the PDB file using PDB2PQR [46,47], removing solvent molecules and filling in missing atoms. Then, we extract the surface meshes of antigens and antibodies to obtain point cloud data. We process the point cloud data using the Adaptive Poisson-Boltzmann Solver (APBS) to obtain Poisson–Boltzmann electrostatics for each point cloud [48,49]. In this part, we obtain the point cloud features of antigens and antibodies separately. Each point cloud contains 3-dimensional spatial coordinate Gi information (x(i), y(i), z(i)) and 11-dimensional charge Ei information.

2.3.2. Amino acid features

We use various algorithms and tools to extract antigen and antibody amino acid features. (1) One-hot encoding: we perform one-hot encoding for all amino acids, classifying all uncommon amino acids into one class, and ultimately obtaining a feature of length 21 for each amino acid. (2) Neighbor composition: we calculate the frequency of 20 common amino acids within an 8 Å distance for each amino acid, providing a feature of length 20. (3) Absolute and relative solvent accessible surface area: we use PyRosetta [50], which is encapsulated by the powerful Rosetta platform [51], to calculate the absolute and relative solvent accessible surface area of each amino acid, forming a feature of length 2. (4) Position-Specific Scoring Matrix (PSSM): we use PSI-BLAST to calculate the PSSM of each amino acid, obtaining a feature of length 20 [52]. (5) Peptides: we use the R Package Peptides to extract amino acid features and obtain physicochemical features such as Cruciani Properties [53,54], forming a feature of length 66. (6) Others: we calculate residue depth, residue adjacency degree, average B-factor, isoelectric point, and molecular weight, forming a feature of length 6. Finally, we compute a feature of length 136 for each amino acid of the antigen and antibody.

2.4. Model framework of WUREN

We construct a deep learning model as shown in Fig. 2A, which consists of four modules: Cross Attention PointNet+ + (CAP), Cross Attention GCN (CAG), Cross Attention Transformer (CAT), and Multilayer Perceptron (MLP). First, the point cloud features of antigens and antibodies are input into the Cross Attention PointNet+ + (CAP) module, obtaining the corresponding point cloud embeddings. Next, the antigen and antibody point cloud embeddings pass through a Pooling layer, which aggregates the embeddings of each point cloud to the nearest amino acid, resulting in amino acid-level embeddings that carry spatial structural information. These embeddings are then concatenated

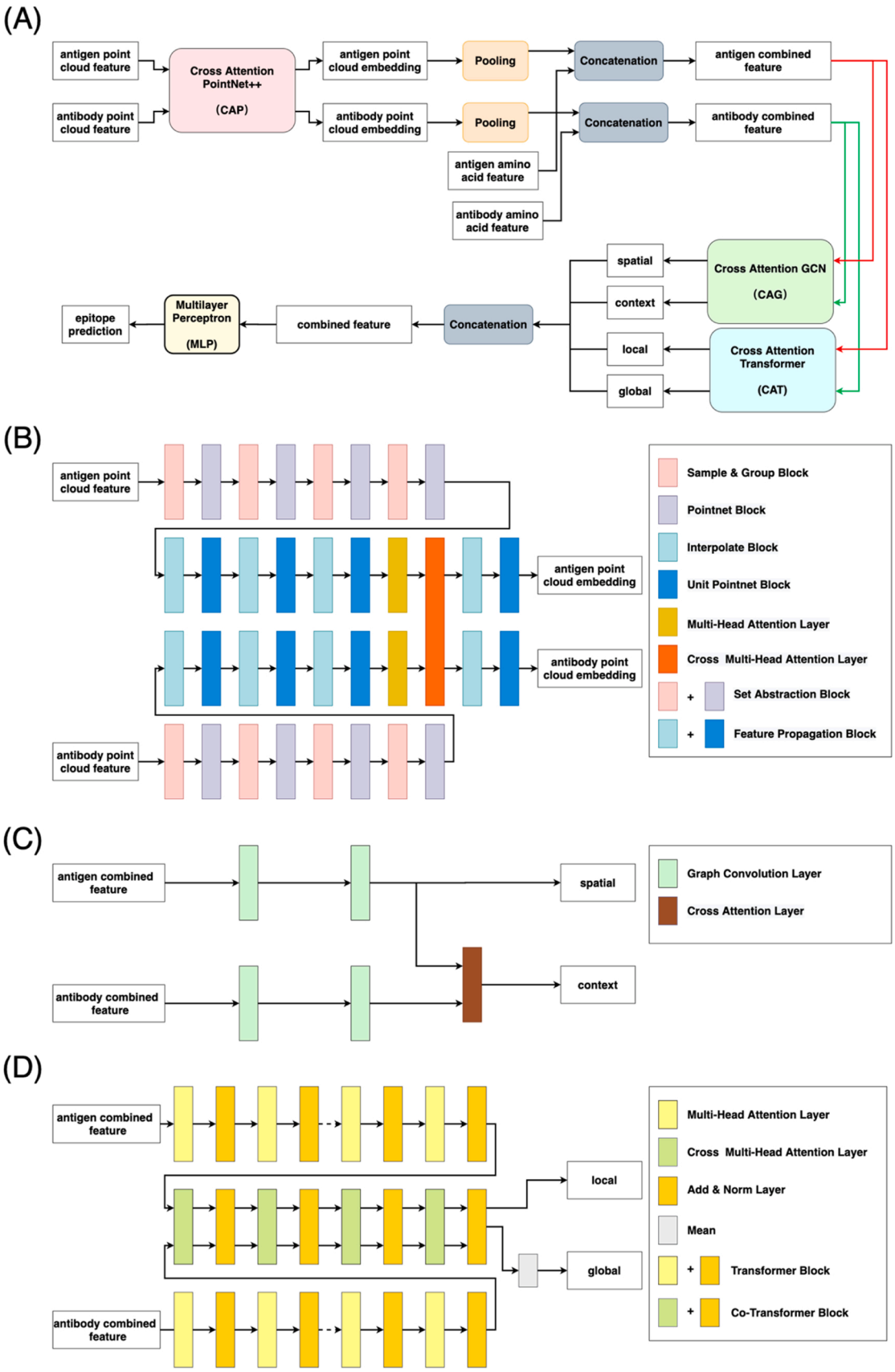

Fig. 2. WUREN. (A) Overall framework of the WUREN model. (B) Diagram of Cross Attention PointNet+ + (CAP) block. (C) Diagram of Cross Attention GCN (CAG) block. **(D)** Diagram of Cross Attention Transformer (CAT) block.

with the pre-extracted physicochemical features at the amino acid level, yielding the antigen combined feature and antibody combined feature. Subsequently, the antigen combined feature and antibody combined feature are input into the Cross Attention GCN (CAG) module, obtaining spatial features and context features. The antigen combined feature and antibody combined feature are fed into the Cross Attention Transformer (CAT) module, acquiring local features and global features. Finally, the spatial, context, local, and global features are concatenated to obtain the combined feature, which is processed through a final Multilayer Perceptron (MLP) to predict the B-cell epitope probability.

The structure of the CAP module is shown in Fig. 2B. Firstly, the point cloud features of antigens and antibodies are downsampled through four Set Abstraction (SA) blocks, each of which consists of a Sample & Group block and a Pointnet block. Secondly, the antigen and antibody features are upsampled through three Feature Propagation (FP) blocks, with each FP block comprising an Interpolate block and a Unit Pointnet block. Subsequently, the antigen and antibody features undergo self-attention feature extraction via a Multi-Head Attention layer. Then, the self-attention features of antigens and antibodies are processed through a Cross Multi-Head Attention layer to extract interactive attention features. Finally, the antigen and antibody features are upsampled through the last FP block, outputting the point cloud embeddings of antigens and antibodies. The Sample & Group, Pointnet, Interpolate, and Unit Pointnet blocks are consistent with the descriptions provided in the PointNet+ + [33].

The structure of the CAG module is shown in Fig. 2C. In this module, the antigen combined feature and antibody combined feature are extracted by two GCN layers, respectively. The graph feature of the antigen itself is spatial, and context information is obtained after calculating the cross-attention of the antigen and antibody.

The structure of the CAT module is shown in Fig. 2D. First, the antigen combined feature and antibody combined feature each pass through four Transformer blocks, with each Transformer block consisting of a Multi-Head Attention layer and an Add & Norm layer. Subsequently, the antigen and antibody features together go through four Co-Transformer blocks, each of which is composed of a Cross Multi-Head Attention layer and an Add & Norm layer [55]. In the Transformer block, in addition to using the absolute position information of each amino acid, we also incorporate the Text-to-Text Transfer Transformer (T5) relative position information, representing the distance between every two amino acids [56]. In the Co-Transformer block, the features of the antigen are used as Query, while the features of the antibody are used as Key and Value. The attention weight between each pair of amino acids in the.

antigen and antibody is calculated. After processing by this block, we obtain the local information of each amino acid and the global information after averaging all amino acid features.

With the model framework described above, to synergistically fuse spatial, topological, and sequential information, we perform the following training steps:

First, we train the CAP model using the features of the antigen and antibody point clouds. Second, with the CAP model parameters fixed, we obtain the point cloud embeddings of antigens and antibodies through the model and aggregate the point cloud embeddings into amino acid level embeddings. Third, we concatenate the embeddings obtained by the point cloud model with other amino acid level features and train the CAG and CAT models. As a result, we obtain spatial, context, local, and global information for each amino acid. Finally, we concatenate the spatial, context, local, and global features and train the MLP prediction model to obtain the prediction probability of whether each amino acid is in the epitope set.

Following the above process, the spatial information extracted by the point cloud CAP module, the topological information extracted by the graph CAG module, and the sequence information extracted by the sequence CAT module are effectively fused to achieve an efficient amino acid representation.

2.5. Implementation details

We train the model using a single A100 GPU with 15 G memory.

2.5.1. Point cloud CAP model training

We use the Adam optimizer with a learning rate of 0.0001 and weight decay of 0.001 [57]. The Cross Entropy loss function is employed [58], with the weight set to 0.075 to mitigate the impact of the imbalance between positive and negative samples on model training. The batch size is set to 5, meaning that 5 complete antigen and antibody point cloud pairs are processed each time. The dropout rate is set to 0.6, and the training epoch is 300.

2.5.2. Splicing point cloud embeddings and amino acid features

We fix the parameters of the point cloud CAP model and extract the antigen and antibody point cloud embeddings with a length of 128 through the CAP model. We associate the coordinates of the point cloud with the coordinates of the amino acids in the PDB file, extract all point clouds with a distance less than 2 Å for each amino acid, and average the embeddings of these point clouds to obtain a 128-length embedding for each amino acid. As a result, the point cloud embedding is converted into amino acid embedding. We splice the 128-length amino acid embedding containing spatial structure information obtained by the CAP module and the 136-length amino acid feature obtained in the data preparation stage, resulting in a 264-length feature for each amino acid.

2.5.3. CAG and CAT model training

We use Adam as the optimizer with a learning rate of 0.0001 and weight decay of 0.001. We employ Cross Entropy as the loss function, with the weight set to 0.04 to mitigate the impact of the imbalance between positive and negative samples on model training. The batch size is set to 32, meaning that 32 amino acids of the antigen are processed each time. The dropout rate is set to 0.6, and the training epoch is 100.

3. Result

3.1. B-cell epitope prediction performance on EpiPred

We use 118 training data and 30 test data consistent with the EpiPred paper [38]. To avoid model overfitting, we use the 44 complexes from the BM dataset as the validation that are not included in the training set [39]. To ensure that the point cloud data of antigens and antibodies meet the model's input requirements, we sample the antigen and antibody point clouds using the Farthest Point Sampling (FPS) method, acquiring 8192 antigen point clouds and 8192 antibody point clouds, respectively. This process improves the training efficiency of the point cloud model without sacrificing its performance. To make the point cloud model more robust, we perform data augmentation on the point cloud data, subjecting each data point to random translation and rotation transformations. Finally, we normalize both the point cloud and amino acid features to avoid the deterioration of the model performance due to the singular values of some features.

On the EpiPred dataset, we compare the performance of WUREN with tools Attract, EpiPred, PECAN, and PINet [26,28,38]. During the testing process of Attract, a rigid docking mode is utilized. For each structure, 100 conformations are generated, and the top 1 conformation with the lowest free energy is selected as the result. Area under the precision recall curve (AUC-PR), area under the receiver operating characteristics curve (AUC-ROC), precision, and recall are used as test metrics, and the corresponding functions in the scikit-learn (sklearn) metric module are used for calculating these metrics [59]. As shown in Fig. 3A, on the 30 test data of EpiPred, WUREN achieves significant improvements in AUC-PR (0.462), AUC-ROC (0.877), and F1 (0.462) compared to Attract, EpiPred, PECAN, and PINet, obtaining state-of-the-art results. Supplementary Table S4 provides the specific evaluation metrics. To achieve a more intuitive comprehension of the

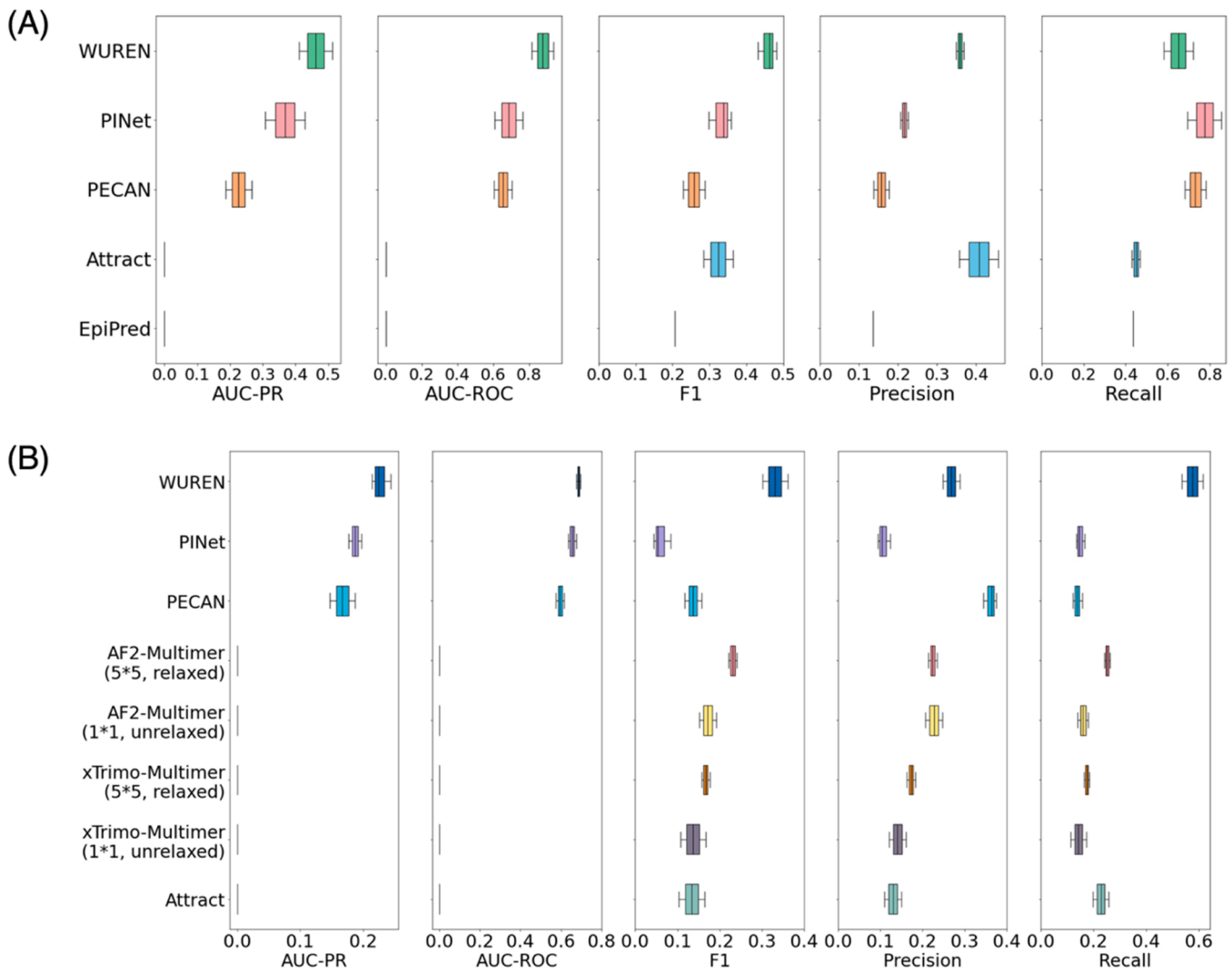

Fig. 3. Results of WUREN performance on EpiPred and SAbDab benchmarks. (A) WUREN achieved AUC-PR, AUC-ROC, F1, Precision, and Recall scores of 0.462, 0.877, 0.462, 0.359, and 0.647 respectively on the EpiPred Benchmark, outperforming the other models in comparison in terms of AUC-PR, AUC-ROC, and F1. (B) On the newly released data from SAbDab, WUREN reached AUC-PR, AUC-ROC, F1, Precision, and Recall scores of 0.224, 0.685, 0.331, 0.268, and 0.577 respectively, surpassing the other comparison models in terms of AUC-PR, AUC-ROC, F1 and Recall.

model's prediction outcomes, the 3D visualizations of the predictions for the 30 test samples are available in Supplementary Fig. S1, Supplementary Fig. 2, and Supplementary Fig. S3.

In the testing above, we have demonstrated that WUREN achieved state-of-the-art performance on the EpiPred benchmark. To further validate that WUREN has learned information about antigen-antibody interactions rather than any other biases, we calculated the correlation between the attention weights and the predicted probabilities on 30 test samples. The pearson correlation coefficient between the two was found to be 0.688, providing evidence that the model accurately predicts B-cell epitopes based on learned antigen-antibody interaction information. The detail result can be found at Supplementary Fig. S4.

3.2. B-cell epitope prediction performance on SAbDab

We conduct a more extensive test on a newly released dataset of 77 antigen-antibody complexes from SAbDab [40]. Besides comparing WUREN with Attract, PINet and PECAN, we also included comparisons with protein complex generation models such as AF2-Multimer and xTrimo-Multimer. F1 score, precision, and recall are used as test metrics.

In the practical implementation, two setups of AF2-Multimer and xTrimo-Multimer are used for comparison. One setup uses a single model to predict a single structure, while the other employs 5 models, each predicting 5 structures. In the latter configuration, the predicted structure is chosen as the best of the 25 structures after relaxation, based on prediction confidence. After obtaining the predicted complex structure, we calculate the pairwise amino acid distance between the antigen and the antibody to obtain the predicted epitope with a threshold of 4.5 Å.

As shown in Fig. 3B, WUREN generally achieves the best results on SAbDab dataset. Supplementary Table S6 provides the specific evaluation metrics. Additionally, in the process of testing the Multimer method, we find that the results derived from multiple models predicting multiple structures surpassed those from a single model predicting a single structure. Outcomes from relaxing the structure exceeded those where no relaxation was applied. In cases with the same model, structure prediction, and relaxation settings, AF2-Multimer yielded superior results compared to xTrimo_multimer.

3.3. B-cell epitope prediction performance on other external test set

To comprehensively evaluate the performance of WUREN, we compare it with eight other tools on nine antibody-antigen complex datasets using AUC-PR as the evaluation metric. The datasets included

EpiPred, SAbDab, DiscoTope3_Foldx, DiscoTope3_Solved, DiscoTope3_Af2, SEDB, DiscoTope, Epitome, and SEMA. The eight tools consisted of PINet, PECAN, DiscoTope 3.0, BepiPred 3.0, Epitope3D, EPCES, EPSVR, and ElliPro. We obtain the test results for each tool on the test dataset by utilizing the respective webservers. The webserver links for DiscoTope 3.0, BepiPred 3.0, Epitope3D, EPCES, EPSVR, and ElliPro can be found in their articles.

From Fig. 4, it can be observed that WUREN achieved the best performance on the EpiPred, SEDB, DiscoTope, and SEMA datasets. Additionally, WUREN performed second best on the SAbDab, DiscoTope3_Solved, DiscoTope3_Af2, and Epitome datasets, slightly behind DiscoTope 3.0. Supplementary Table S12 provides specific evaluation metrics. These results demonstrate that WUREN exhibits state-of-the-art B-cell epitope prediction performance.

It is worth noting that WUREN achieved AUC-PR values of 0.217, 0.213, and 0.193 on the DiscoTope3_Foldx, DiscoTope3_Solved, and DiscoTope3_Af2 datasets, respectively. WUREN exhibited enhanced performance on the Foldx energy-optimized structure in comparison to the solved structure, implying that structural accuracy influences WUREN's performance. Despite a decline in performance on the AlphaFold-generated structure, WUREN still surpassed other comparative tools, with the exception of DiscoTope 3.0. Since WUREN is a framework that integrates sequence, graph, and structural features, which differs from the DiscoTope 3.0 framework, the differentiation within WUREN framework may be explored for further improvement.

3.4. Model ablation experiment results

An ablation test was conducted to assess the individual contributions of the three modules responsible for extracting sequential, topological, and spatial information. As depicted in Fig. 5A, Fig. 5B, and Fig. 5C, ablation experiments were performed on the WUREN model using EpiPred's 30 test data samples. Supplementary Table S13 provides the specific evaluation metrics. In general, the model's performance improved as its complexity increased with the fusion of more representations.

4. Discussion

The structure of antigen-antibody complexes reveal the interactions between antigens and antibodies at atomic level. These structures are primarily determined by the unique sequences of each protein and the interactions among amino acids. Computational approaches offer an efficient means to understand protein functions based on their sequence compositions. While we can extract substantial information from amino acid sequences or structures alone, a practical model should take full advantage of the available data by utilizing suitable features from different representations. In this work, we introduce a multimodal deep learning model, WUREN, designed to efficiently merge sequence, graph, and structural features. Using WUREN, we significantly improve the B-cell epitope prediction in comparison with the SOTA methods.

The training, validation, and testing of WUREN involved the use of complex structure data. However, generated structures can be utilized for B-cell epitope prediction in application scenarios. Recognizing the impact of structural accuracy on WUREN's performance, we conducted tests on the external test data from DiscoTope 3.0. The results demonstrated that WUREN achieved higher AUC-PR on AlphaFold-generated structures compared to other tools, excluding DiscoTope 3.0, indicating its practical usability.

Beyond the improvements introduced by WUREN, our study finds that selecting and incorporating more features associated with the physicochemical properties of amino acid significantly enhance the model performance. When only the CAG module is utilized for modeling, increasing the dimension of amino acid representation from 63 (as in PECAN) to 136 (as in this study) raises the AUC-PR from 0.226 to 0.338. The 136 features, which contribute more to the B-cell epitope prediction task than others, are selected from 183 commonly used amino acid features (such as amino acid type, absolute solvent accessible surface area, etc.) using the random forest algorithm. We think the same procedure can be applied to improve model performances for other similar tasks.

Moreover, we discover that post-processing of the initial results also enhances the accuracy of epitope predictions. By clustering the predicted epitopes, selecting the cluster central points, and performing nearest neighbor sampling based on these central points, the F1 score on the test set can be improved by approximately 10 %.

It is worth mentioning that the quantity of training and test data utilized in this study is considered small within the realm of deep learning research. Increasing the volume of training/test data is expected to enhance the performance of the models and enable a more rigorous evaluation.

There is still room to further improve the WUREN's performance. Firstly, to mitigate the impact of precision in protein structures, we are considering incorporating AlphaFold-generated structures into the model training and testing process. Secondly, inspired by DiscoTope 3.0, integrating the positive-unlabeled (PU) learning strategy into the training process has the potential to enhance WUREN's performance. Lastly, since the process of aggregating point cloud embeddings into amino acid embeddings is non-differentiable, WUREN requires two-stage training currently, which increases training complexities and potentially also reduce the model's performance. If we can design the embedding aggregation process as a differentiable process, it would enable us to construct an end-to-end model to further improve the performance.

5. Conclusion

In this paper, we propose a multimodal framework called WUREN, which fully integrates one-dimensional sequence features, two-dimensional graph features, and three-dimensional structural features for improved B-cell epitope prediction. To demonstrate WUREN's advantages, we implemented the model with three specific modules—CAT, CAG, and CAP—to extract sequence features, graph features, and structural features, respectively. In the B-cell epitope prediction task, our model achieves SOTA results in AUC-PR, AUC-ROC, and precision on the EpiPred dataset, as well as improves model performance on 77 newly released SAbDab data samples compared to the advanced multimer methods. On the DiscoTope 3.0 external test data, WUREN obtained an AUC-PR of 0.193 on AlphaFold-generated structure, surpassing other tools except for DiscoTope 3.0. Ablation experiments reveal that sequence features, graph features, and structural features all play crucial roles.

Taken together, we present a new multimodal deep learning model, WUREN, for efficient B-cell epitope prediction from antigen-antibody complex, in which the model demonstrates surpassed SOTA performance. Importantly, we believe that our model design and implementation can be adapted to protein drug design or other biologic studies given the sequence, graph and structure information are available. We envisage that further investigation of this model will help us to better understand the relative importance of sequence, graph and structure and eventually aid the design of new therapeutics.

CRediT authorship contribution statement

Lipeng Lai: Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Xiaodong Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xuezhe Fan:** Data curation. **Xiangrui Gao:** Methodology. **Zhe Huai:** Data curation. **Mengcheng Yao:** Data curation. **Genwei Zhang:** Writing – review & editing. **Xiaolu Huang:** Supervision, Funding acquisition. **Tianyuan Wang:** Project administration, Data curation.

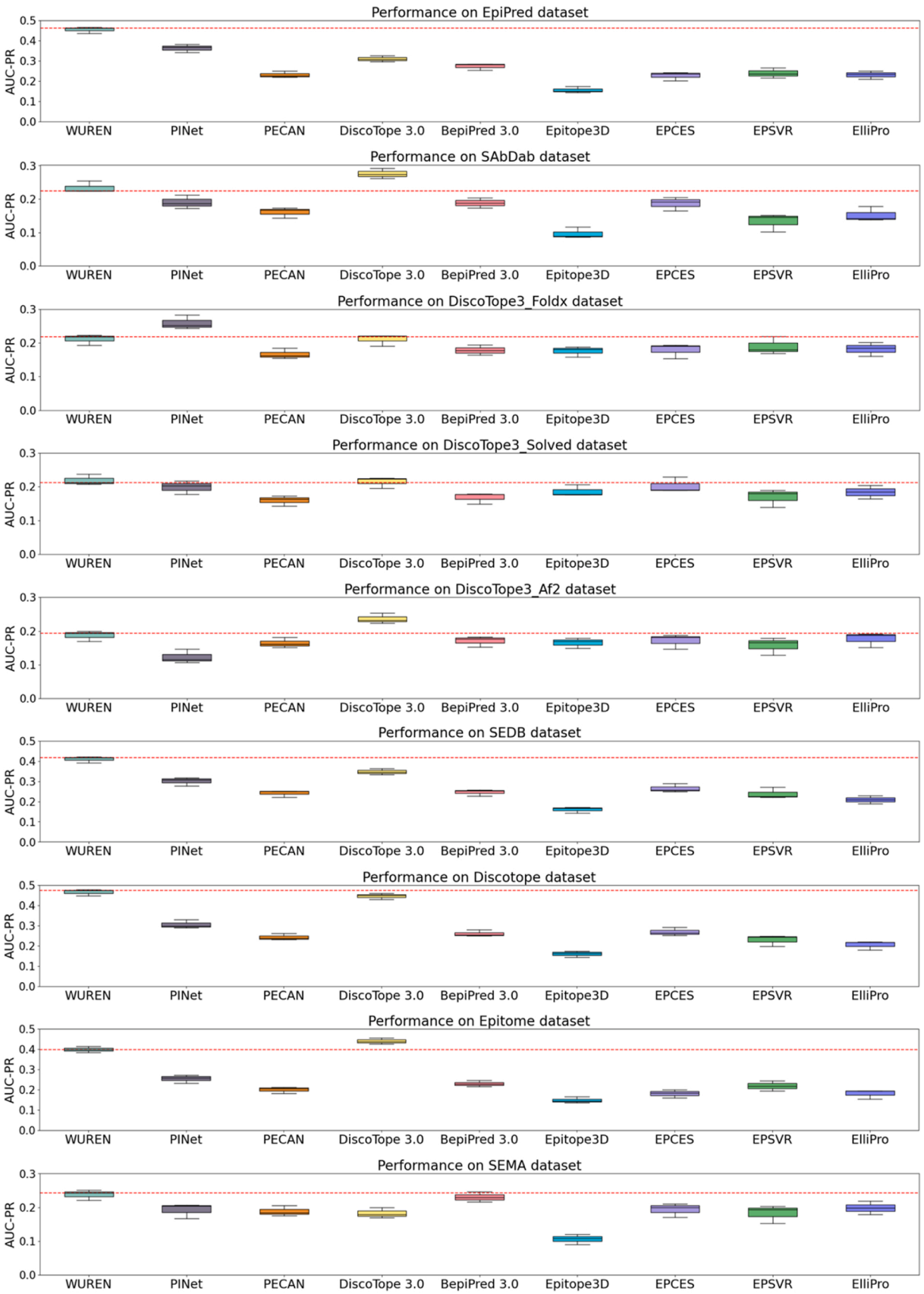

Fig. 4. Results of WUREN performance on nine external test data. WUREN exhibited state-of-the-art performance with AUC-PR values of 0.462, 0.418, 0.473, and 0.243 on the EpiPred, SEDB, Discotope, and SEMA datasets, respectively. It achieved the second-best performance on the SAbDab, DiscoTope3_Solved, DiscoTope3_Af2, and Epitome datasets, with AUC-PR values of 0.224, 0.213, 0.193, and 0.399, respectively.

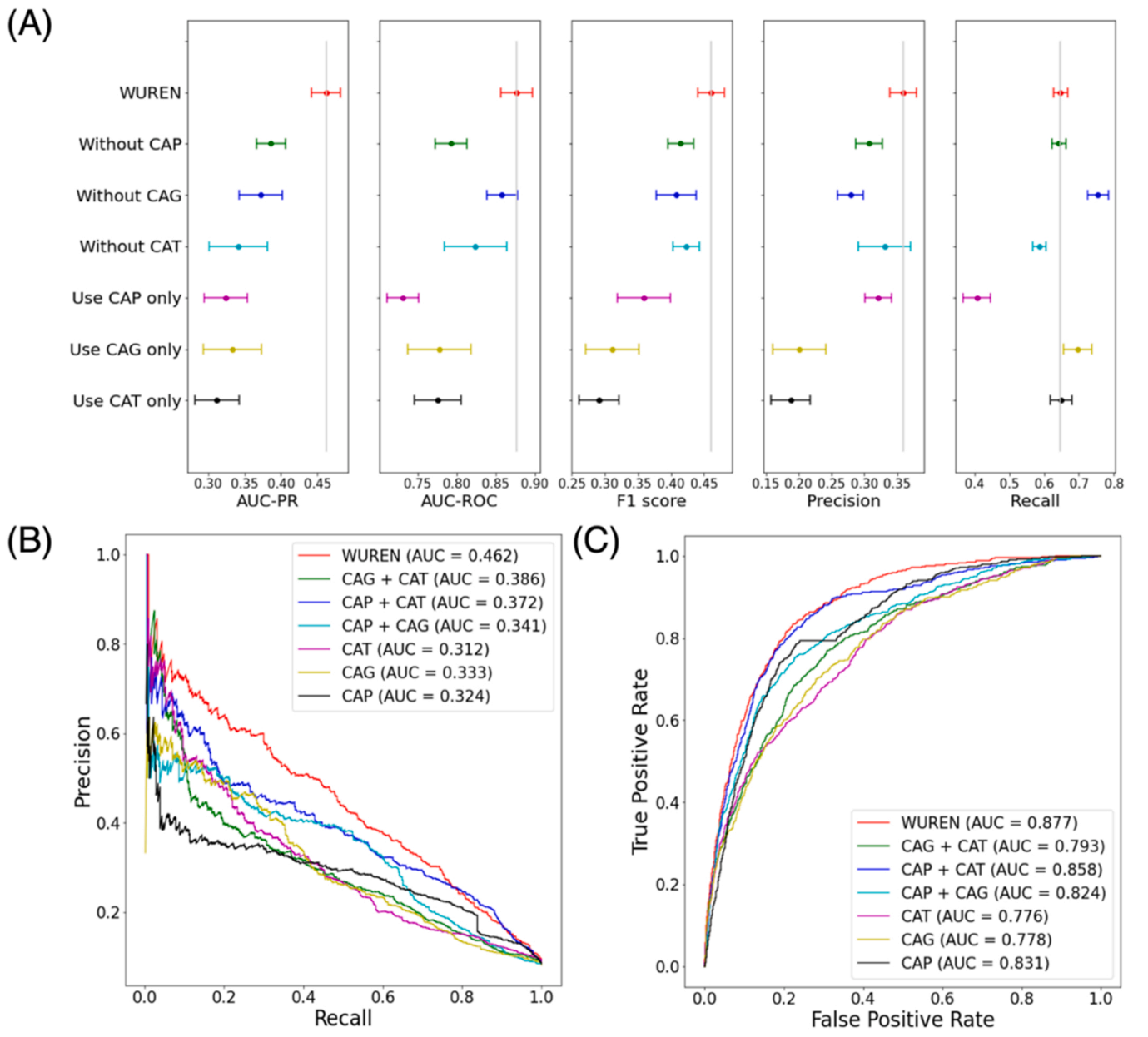

Fig. 5. Model ablation test results. (A) Ablation outcomes on the EpiPred benchmark show a significant decrease in AUC-PR, AUC-ROC, F1 score, and Precision for WUREN as the number of constituent modules is reduced. (B) The complete WUREN model achieved an AUC-PR of 0.462. After removing the CAP, CAG, or CAT modules, the AUC-PR decreased to 0.386, 0.372, and 0.341 respectively. When using the CAP, CAG, or CAT modules individually, the AUC-PR further decreased to 0.324, 0.333, and 0.312 respectively. (C) The complete WUREN model reached an AUC-ROC of 0.877. After removing the CAP, CAG, or CAT modules, the AUC-ROC dropped to 0.793, 0.858, and 0.824 respectively. When using the CAP, CAG, or CAT modules individually, the AUC-ROC further decreased to 0.831, 0.778, and 0.776 respectively.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to express their gratitude to Chaohui Gong for helpful comments, as well as Ruyu Wang for proofreading.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.023.

References

[1] Bashford-Rogers RJM, Bergamaschi L, McKinney EF, et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature 2019;574(7776):122–6. https://doi.org/10.1038/s41586-019-1595-3.
[2] Nakane T, Kotecha A, Sente A, et al. Single-particle cryo-EM at atomic resolution. Nature 2020;587(7832):152–6. https://doi.org/10.1038/s41586-020-2829-0.
[3] Maveyraud L, Mourey L. Protein X-ray crystallography and drug discovery. Molecules 2020;25(5):1030. https://doi.org/10.3390/molecules25051030.
[4] Fowler NJ, Sljoka A, Williamson MP. A method for validating the accuracy of NMR protein structures. Nat Commun 2020;11(1):6321. https://doi.org/10.1038/s41467-020-20177-1.
[5] Lyumkis D. Challenges and opportunities in cryo-EM single-particle analysis. J Biol Chem 2019;294(13):5181–97. https://doi.org/10.1074/jbc.REV118.005602.

[6] Sircar A, Gray JJ. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. PLoS Comput Biol 2010;6(1):e1000644. https://doi.org/10.1371/journal.pcbi.1000644.

[7] Zacharias M. ATTRACT: protein-protein docking in CAPRI using a reduced protein model. Proteins 2005;60(2):252–6. https://doi.org/10.1002/prot.20566.

[8] Stewart JJ. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J Mol Model 2013;19(1):1–32. https://doi.org/10.1007/s00894-012-1667-x.

[9] Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11): 1422–3. https://doi.org/10.1093/bioinformatics/btp163.

[10] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv 2022. https://doi.org/10.1101/2022.07.20.500902.

[11] Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: improved B-cell epitope prediction using protein language models. Protein Sci 2022;31(12):e4497. https://doi.org/10.1002/pro.4497.

[12] Ren J, Song J, Ellis J, Li J. Staged heterogeneity learning to identify conformational B-cell epitopes from antigen sequences. BMC Genom 2017;18(Suppl 2):113. https://doi.org/10.1186/s12864-017-3493-0.

[13] Dalkas GA, Rooman M. SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence. BMC Bioinform 2017;18(1):95. https://doi.org/10.1186/s12859-017-1528-9.

[14] Høie MH, Gade FS, Johansen JM, et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol 2024;15: 1322712. https://doi.org/10.3389/fimmu.2024.1322712.

[15] Shashkova TI, Umerenkov D, Salnikov M, et al. SEMA: antigen B-cell conformational epitope prediction using deep transfer learning. Front Immunol 2022;13:960985. https://doi.org/10.3389/fimmu.2022.960985.

[16] da Silva BM, Myung Y, Ascher DB, Pires DEV. epitope3D: a machine learning method for conformational B-cell epitope prediction. Brief Bioinform 2022;23(1): bbab423. https://doi.org/10.1093/bib/bbab423.

[17] Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics 2008; 24(12):1459–60. https://doi.org/10.1093/bioinformatics/btn199.

[18] Liang S, Zheng D, Yao B, Zhang C. EPCES and EPSVR: prediction of B-Cell antigenic epitopes on protein surfaces with conformational information. Methods Mol Biol 2020;2131:289–97. https://doi.org/10.1007/978-1-0716-0389-5_16.

[19] Ponomarenko J, Bui HH, Li W, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinforma 2008;9:514. https://doi.org/10.1186/1471-2105-9-514.

[20] Zhou C, Chen Z, Zhang L, et al. SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens. Nucleic Acids Res 2019;47(W1):W388–94. https://doi.org/10.1093/nar/gkz413.

[21] Solihah B, Azhari A, Musdholifah A. Enhancement of conformational B-cell epitope prediction using CluSMOTE. PeerJ Comput Sci 2020;6:e275. https://doi.org/10.7717/peerj-cs.275.

[22] Sun P, Qi J, Zhao Y, et al. A novel conformational B-cell epitope prediction method based on mimotope and patch analysis. J Theor Biol 2016;394:102–8. https://doi.org/10.1016/j.jtbi.2016.01.021.

[23] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2.

[24] Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv 2022. https://doi.org/10.1101/2021.10.04.463034.

[25] Chen B, Cheng X, Geng Y, et al. xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein. bioRxiv 2023. https://doi.org/10.1101/2023.07.05.547496.

[26] Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. Bioinformatics 2020;36(13): 3996–4003. https://doi.org/10.1093/bioinformatics/btaa263.

[27] Tubiana J, Schneidman-Duhovny D, Wolfson HJ. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods 2022;19(6):730–9. https://doi.org/10.1038/s41592-022-01490-7.

[28] Dai B, Bailey-Kellogg C. Protein interaction interface region prediction by geometric deep learning. Bioinformatics 2021;37(17):2580–8. https://doi.org/10.1093/bioinformatics/btab154.

[29] Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci USA 2012;109(26):10340–5. https://doi.org/10.1073/pnas.1207864109.

[30] Källberg M, Wang H, Wang S, et al. Template-based protein structure modeling using the RaptorX web server. Nat Protoc 2012;7(8):1511–22. https://doi.org/10.1038/nprot.2012.085.

[31] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Neural Inf Process Syst (NIPS) 2017;volume30. https://doi.org/10.48550/arXiv.1706.03762.

[32] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review. Comput Soc Netw 2019;6:11. https://doi.org/10.1186/s40649-019-0069-y.

[33] R Qi C , Yi L , Su H , et al. PointNet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the thirty first international conference on neural information processing systems; 2017. pp. 5105–14. Available from: doi:10.48550/arXiv.1706.02413.

[34] R. Qi C., Su H., Mo K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. pp. 77–85. Available from: doi:10.48550/arXiv.1612.00593.

[35] Veličković P., Cucurull G., Casanova A., et al. Graph attention networks. ICLR; 2018. Available from: doi:10.48550/arXiv.1710.10903.

[36] Guo Y, Wang H, Hu Q, Liu H, Liu L, Bennamoun M. Deep learning for 3D point clouds: a survey. IEEE Trans Pattern Anal Mach Intell 2021;43(12):4338–64. https://doi.org/10.1109/TPAMI.2020.3005434.

[37] Zhou J, Cui G, Hu S, et al. Graph neural networks: a review of methods and applications. AI Open 2020;1:57–81. https://doi.org/10.48550/arXiv.1812.08434.

[38] 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):2288–94. https://doi.org/10.1093/bioinformatics/btu190.

[39] Guest JD, Vreven T, Zhou J, et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. Structure 2021;29(6):606–621.e5. https://doi.org/10.1016/j.str.2021.01.005.

[40] Dunbar J, Krawczyk K, Leem J, et al. SAbDab: the structural antibody database. Nucleic Acids Res 2014;42(Database issue):D1140–6. https://doi.org/10.1093/nar/gkt1043.

[41] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. W382-W388 Nucleic Acids Res 2005;33(Web Server issue). https://doi.org/10.1093/nar/gki387.

[42] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004;32(Web Server issue):W20–5. https://doi.org/10.1093/nar/gkh435.

[43] Sharma Om Parkash, Das Arindam Atanu, et al. Structural Epitope Database (SEDB): a web-based database for the epitope, and its intermolecular interaction along with the tertiary structure information. J Proteom Bioinforma 2012;5(3): 84–9. https://doi.org/10.4172/jpd.1000217.

[44] Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 2006;15(11):2558–67. https://doi.org/10.1110/ps.062405906.

[45] Schlessinger A, Ofran Y, Yachdav G, Rost B. Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res 2006;34(Database issue):D777–80. https://doi.org/10.1093/nar/gkj053.

[46] Berman HM, Battistuz T, Bhat TN, et al. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58(Pt 6 No 1):899–907. https://doi.org/10.1107/s0907444902003451.

[47] Dolinsky TJ, Czodrowski P, Li H, et al. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 2007;35(Web Server issue):W522–5. https://doi.org/10.1093/nar/gkm276.

[48] Jurrus E, Engel D, Star K, et al. Improvements to the APBS biomolecular solvation software suite. Protein Sci 2018;27(1):112–28. https://doi.org/10.1002/pro.3280.

[49] Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 2001;98(18):10037–41. https://doi.org/10.1073/pnas.181342398.

[50] Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 2010; 26(5):689–91. https://doi.org/10.1093/bioinformatics/btq007.

[51] Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzym 2004;383:66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.

[52] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17): 3389–402. https://doi.org/10.1093/nar/25.17.3389.

[53] Osorio D, Rondon-Villarreal P, Torres R. Peptides: a package for data mining of antimicrobial peptides. R J 2015;7(1):4–14. https://doi.org/10.32614/RJ-2015-001.

[54] Cruciani G, Baroni M, Carosati E, et al. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. J Chemom 2004;18: 146–55. https://doi.org/10.1002/cem.856.

[55] Qin L., Liu T., Che W., et al. A co-interactive transformer for joint slot filling and intent detection. In: Proceedings of the ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2020; 8193–8197. Available from: doi: 10.1109/ICASSP39728.2021.9414110.

[56] Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J Mach Learn Res (JMLR) 2020;21(140):1–67. https://doi.org/10.48550/arXiv.1910.10683.

[57] Diederik P.Kingma ,Jimmy Ba. Adam: a method for stochastic optimization. international conference on learning representations (ICLR); 2015. doi:10.48550/arXiv.1412.6980.

[58] Zhilu Zhang ,Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In: Proceedings of the thirty second conference on neural information processing systems (NeurIPS); 2018p. 8792–802. doi:10.48550/arXiv.1805.07836.

[59] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res (JMLR) 2011;12:2825–30. https://doi.org/10.48550/arXiv.1201.0490.