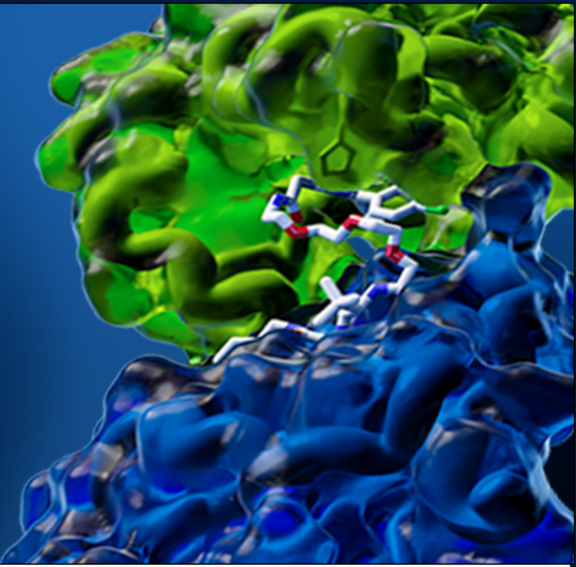
bio-techne®

Analytical Exchange Webinar

Accelerating Protein Analysis Throughput at Boehringer Ingelheim

October 3, 2023
3pm BST | 4pm CEST | 10am EDT | 7am PDT

Register



Advancing Drug Discovery with Targeted Protein Degradation

Unlike traditional drug discovery, which focuses on inhibiting or activating proteins, targeted protein degradation (TPD) offers a more precise and efficient way to alter cellular pathways. However, TPD faces several challenges that need to be overcome to reach its full potential, including reducing the bottleneck of quantitative protein analysis.

In this upcoming webinar, the team at Boehringer Ingelheim present their recent real- world TPD project examples and in-depth analysis, covering insights on device and consumable management and optimized assay setups to increase protein analysis throughput.

Speakers

Andrea Stingu
Lab Scientist
Boehringer Ingelheim

Johannes Wachter
Lab Scientist
Boehringer Ingelheim

Teressa Puchner
Lab Scientist
Boehringer Ingelheim

Register Now

Connect with us: • Search for Products • Family of Brands • Distributors • Contact

Bio-Techne
614 McKinley Place NE, Minneapolis, MN USA 55413

Received: 29 July 2022 | Revised: 27 September 2022 | Accepted: 28 September 2022

DOI: 10.1002/pro.4462

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

BindWeb: A web server for ligand binding residue and pocket prediction from protein structures

Ying Xia | Chunqiu Xia | Xiaoyong Pan | Hong-Bin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China

Correspondence
Xiaoyong Pan and Hong-Bin Shen, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China.
Email: 2008xypan@sjtu.edu.cn and hbshen@sjtu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61725302, 62073219, 61903248; Science and Technology Commission of Shanghai Municipality, Grant/Award Numbers: 20S11902100, 22511104100

Review Editor: Nir Ben-Tal

Abstract

Knowledge of protein–ligand interactions is beneficial for biological process analysis and drug design. Given the complexity of the interactions and the inadequacy of experimental data, accurate ligand binding residue and pocket prediction remains challenging. In this study, we introduce an easy-to-use web server BindWeb for ligand-specific and ligand-general binding residue and pocket prediction from protein structures. BindWeb integrates a graph neural network GraphBind with a hybrid convolutional neural network and bidirectional long short-term memory network DELIA to identify binding residues. Furthermore, BindWeb clusters the predicted binding residues to binding pockets with mean shift clustering. The experimental results and case study demonstrate that BindWeb benefits from the complementarity of two base methods. BindWeb is freely available for academic use at http://www.csbio.sjtu.edu.cn/bioinf/BindWeb/.

KEYWORDS

bioinformatics, deep learning, protein-ligand interaction, binding residues, binding pockets

1 | INTRODUCTION

Protein–ligand interactions are indispensable for many biological processes, such as gene expression, signal transduction, and antigen–antibody interaction.[1–7] To explore the interaction mechanisms, experimental methods are applied to resolve complex structures, such as X-ray, nuclear magnetic resonance spectroscopy, and laser Raman spectroscopy. However, considering the time-consuming and high cost of experimental methods, developing efficient computational methods for binding pocket prediction has become an essential topic in structural bioinformatics. We roughly categorize existing computational methods according to input features, computational algorithms, and ligand types.

According to the input features, existing methods can be generally divided into sequence-based methods and structure-based methods. Due to the lack of resolved protein tertiary structures, some methods infer binding residues from protein primary sequences.[8–11] For example, DRNApred is a sequence-based nucleic-acid-binding residue predictor, which utilizes sequence-derived features from the input protein sequence, including evolutionary profiles, putative intrinsic disorder, secondary structures, solvent accessibility, and a variety of physicochemical and biochemical properties.[9] Although a binding pocket generally performs a conservative region on a protein tertiary structure, the contained binding residues are not necessarily sequential in the protein sequence, which raises challenges for sequence-based methods.

Therefore, structure-based methods are proposed to discover binding patterns from the view of protein tertiary structures.[12–16] For example, structure-based NucleicNet calculates the surrounding physicochemical environment of grid points on protein surfaces with the FEATURE program[17] for RNA-binding preference prediction.[18]

Based on the computational algorithms, existing methods can be generally grouped into template-based methods and templated-free methods. A template-based method usually applies sequence or structure alignment as a search engine to search templates for a query protein from an extensive template library of known protein–ligand complexes, and maps the binding residues from templates to the query protein.[19,20] Differently, template-free methods are designed to embed the binding patterns that rely on intrinsic local properties of protein sequences or structures with machine learning methods. For example, DNAPred ensembles hyperplane-distance-based support vector machines (SVMs) for DNA-binding residue prediction from protein sequences.[21] DeepPocket proposes a 3-dimensional convolutional neural network (3DCNN) for ligand-general binding pocket prediction from protein structures.[15] Template-based methods can achieve high confidence predictions when reliable templates are found, and could be easier interpreted on the case of having good templates. On the other hand, template-free methods can potentially identify novel binding pockets, while template-based methods may fail due to the lack of similar proteins in the template library.

Based on the ligand types, existing methods can be roughly divided into ligand-general methods and ligand-specific methods. Ligand-general methods collect proteins which interact with various ligands as a united dataset and predict binding residues without distinguishing ligand types. These methods are useful when the ligand type is unknown, or the number of binding proteins of a specific ligand is too few to train a model.[13,15,16,22] However, their performance may degrade for a particular ligand type due to the binding diversity of different ligands. For example, metal complexations are important for interactions between metal ions and proteins, while the topological features of proteins are more critical for binding to nucleic acids. Therefore, some methods focus on predicting binding residues for a specific ligand type which interacts with enough proteins for constructing a ligand-specific dataset to train a ligand-specific predictor. For example, DNAPred and ATPbind are binding residue predictors for nucleic acids and ATP, respectively.[21,23]

In fact, the differences in input features, computational algorithms and ligand types make existing methods complement each other. Thus, some consensus methods integrate various base methods to boost the identification success rates.[14,19,24] From the user's perspective, it is imperative to make these computational methods accessible by developing easy-to-use web servers.[13–16,18,19]

In this study, we report a user-friendly web server BindWeb for ligand binding residue and pocket prediction from protein structures. BindWeb has two functional modes, a ligand-specific mode for seven specific ligand types (i.e., DNA, RNA, Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME) and a ligand-general mode for any ligand type. BindWeb is a structure-based consensus method and integrates two deep-learning-based methods, a graph neural network (GNN)-based method GraphBind,[12] and a hybrid convolutional neural network (CNN) and bidirectional long short-term memory network (biLSTM)-based method DELIA.[25] Experimental results demonstrate that BindWeb benefits from the complementarity of the two base methods and appears more competitive in binding residue prediction. Furthermore, a spatial clustering module is designed for assigning the predicted binding residues into binding pocket(s) according to their spatial positions. The binding pocket inference helps discover how many pockets are in the query structure and provides a clue for further studying the locally matched geometry between the protein structures and ligands.

2 | MATERIALS AND METHODS

BindWeb is a user-friendly web server for predicting ligand binding residues and pockets. It consists of two deep-learning-based methods, GraphBind and DELIA, which achieve promising performance on benchmark datasets.[12,25] Considering the complementarity of the two base methods, we integrate the prediction results of the two base methods to generate more reliable predictions. Besides, mean shift clustering is used for binding pocket prediction.

2.1 | Benchmark datasets

GraphBind and DELIA construct the ligand-specific datasets for two types of nucleic acids (DNA and RNA) and four types of small ligands (Ca^{2+}, Mn^{2+}, Mg^{2+}, and HEME) from the BioLip database,[26] respectively. The BioLip database semi-manually collects the structures and interactions of the biologically relevant ligand–protein complexes from the Protein Data Bank (PDB)[27] and defines the binding residues based on the atomic distance between the residue and the ligand. For each ligand, proteins are assigned to the training and test sets based on their released dates. To evaluate the generalization of the methods on novel proteins, the redundant proteins are removed to make sure the sequence identity of a training (test) set and the sequence identity between a pair of training and test sets are <30% with CD-HIT.[28] The ATP dataset is released in ATPbind[23] with sequence identity <40%. The above datasets are used for training and evaluating the ligand-specific GraphBind and DELIA.

To expand the method for unseen ligand binding residue prediction, three datasets are used to train and evaluate the ligand-general GraphBind-G, including a training set CHEN11, a validation set JOINED, and a test set COACH420.[12,13] Non-redundant CHEN11 contains 251 proteins that bind to 476 ligands and belong to different SCOP families.[29] JOINED and COACH420 respectively consist of 541 and 420 proteins binding to various drugs and natural ligands. There are no shared proteins between the test set and the training/validation set.

2.2 | GraphBind

The GNN-based GraphBind consists of two modules: the structural-context-based graph construction and the hierarchical GNN.[12] In graph construction, for each target residue, a structural context composed of adjacent residues of the target residue within a fixed distance is extracted. Then, we represent the structural context as a graph $G=(V,E,u,A)$, where V, E, u, and A denote node, edge and graph features, and the adjacency matrix, respectively. In the graph, a residue is denoted as a node, which is represented by a feature vector consisting of relative position encoding, component characteristics, secondary structure features, and evolutionary conversation features. An edge is defined as the positional relationship between two nodes. The hierarchical GNN consists of a GNN encoder, the gated-recurrent-unit-based GNN blocks, and a multilayer perceptron classifier. It progressively updates the edge, node, and graph features, and further learns high-level features for classifying the binding residues. GraphBind is trained independently on seven ligand-specific datasets of DNA, RNA, Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME for ligand-specific binding residue prediction. In addition, GraphBind is trained with ligand-general datasets as GraphBind-G for any ligands.

2.3 | DELIA

DELIA is a biLSTM-CNN-based predictor,[25] which integrates biLSTM models and CNN models to integrate heterogeneous protein information. The biLSTM model is designed for making predictions based on sequence-derived features, including two evolutionary conversation profiles calculated with sequence alignment tools PSI-BLAST[30] and HHblits,[31] the secondary structure profile and relative solvent accessibility predicted with SCRATCH-1D,[32] and the binding propensity predicted by template-based S-SITE.[19] Since the residues closing to each other in a protein sequence may be far from each other from a protein tertiary structure view, distance matrices are used to represent the spatial correlation of residues. The CNN model is applied to learning binding patterns from the distance matrices. In addition, a random undersampling-based ensemble strategy is designed to deal with the imbalanced data and combine prediction results of multiple classifiers trained with different sub-datasets. Finally, the predicted binding scores of classifiers are concatenated and processed with a logistic-regression-based stacked ensemble strategy to generate the final propensity of being a binding residue. Here, DELIA trains five ligand-specific binding residue predictors for Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME.

2.4 | BindWeb pipeline

2.4.1 | Consensus of GraphBind and DELIA predictions

GraphBind and DELIA are constructed based on different statistical machine learning models. Specifically, GraphBind is an end-to-end GNN-based method for embedding local structural and biophysicochemical patterns from graph representations of residue structural contexts. In comparison, DELIA integrates Euclidean-space-based biLSTMs and CNNs to learn local patterns from residue sequence contexts and the distance matrices, respectively. Given the complementarity in protein representations and model architectures of the two base methods, BindWeb integrates them for ligand binding residue prediction. Specifically, for the shared five ligands (Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME) of GraphBind and DELIA, BindWeb provides two sets of predictions, high confidence predictions and medium confidence predictions, which denote the outputs predicted by averaging and pooling the results of the predictors, respectively.

- High confidence: For each residue, the average of predicted binding scores of GraphBind and DELIA is defined as the final binding score, which is used for binding/non-binding residue classification based on the averaged thresholds of two base methods.

- Median confidence: The pooled binding residues predicted by either GraphBind or DELIA are defined as binding residues of BindWeb.

Then, the predicted binding residues with high and medium confidence are clustered into binding pocket(s) based on their spatial information. Figure 1 shows the overall pipeline of BindWeb.

2.4.2 | Clustering binding residues to binding pocket(s) with mean shift clustering

Given that residues located in protein–ligand binding interfaces tend to form spatial clusters,[33] we apply mean shift clustering[34] with the spatial positions of residues to further identify which of the predicted binding residues may potentially form the binding pocket(s). Mean shift clustering is a centroid-based algorithm for locating the maxima of a density function. Compared with the other clustering algorithms, mean shift clustering does not require a predefined number of clusters, which is essential since the number of binding pocket(s) in a protein is uncertain. In addition, mean shift clustering is effective for the case of a few sample-clustering problem in the Euclidean space, which is useful in our study as the predicted binding residues in proteins are generally few. It randomly initializes a set of candidate centroids and iteratively updates candidates for centroids to be the mean of the data within a given region and pushes the neighboring candidates to form the final centroids until the centroids are basically stable. As shown in Figure 1, for the iteration t, the algorithm updates the candidate centroid x_i^{t+1} as the mean of x_i^t's neighborhoods by:

$$x_i^{t+1} = m\left(x_i^t\right) \tag{1}$$

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K\left(x_j - x_i\right) x_j}{\sum_{x_j \in N(x_i)} K\left(x_j - x_i\right)} \tag{2}$$

where $m(x_i)$ is the mean shift vector for moving the centroid toward a region of the maximum increase in the density of data. $K\left(x_j - x_i\right) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{\left\|x_j - x_i\right\|^2}{2h^2}}$ is a Gaussian kernel function with a bandwidth h. $N(x_i)$ is the set of neighbors around x_i with the given bandwidth h.

Here, mean shift clustering takes the spatial positions of the predicted binding residues as the inputs, iteratively searches for the centroids of binding pocket(s), and outputs the binding pocket(s). Due to the specificity of different ligands, the parameter bandwidth h is optimized by maximizing the Fowlkes-Mallows index (FMI) of the training set for each ligand. In particular, for DNA, RNA, Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, HEME, and general ligands, the values of h are 49.5, 49.5, 9.5, 11.5, 11.5, 13.5, 11.5, and 15 Å, respectively.

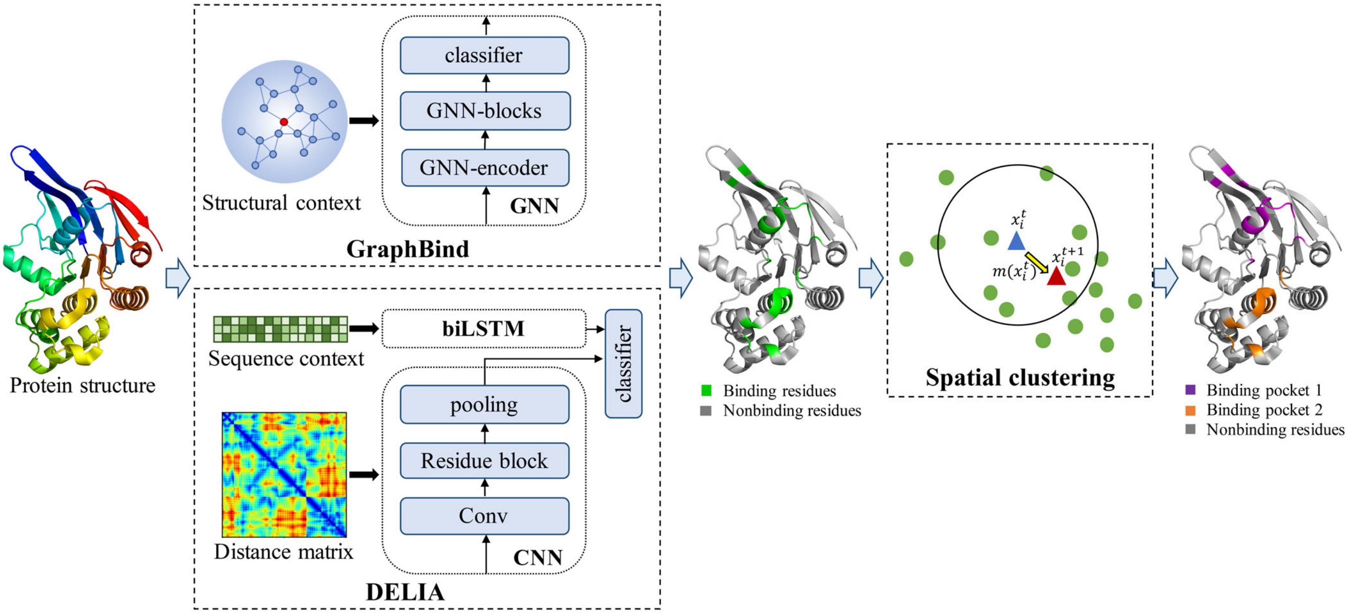


FIGURE 1 The overall pipeline of BindWeb. The binding scores predicted by GraphBind and DELIA are integrated to produce the predicted binding residues. Mean shift clustering is applied for clustering the binding residues to binding pocket(s). In the diagram of spatial clustering, the green points represent the predicted binding residues, the blue and red triangles stand for the position of a centroid in iteration t and t+1, the black circles show the neighbors of the centroid in iteration t, and the yellow arrow stands for the mean shift vector

3 | USAGE OF BINDWEB

3.1 | Web server interface

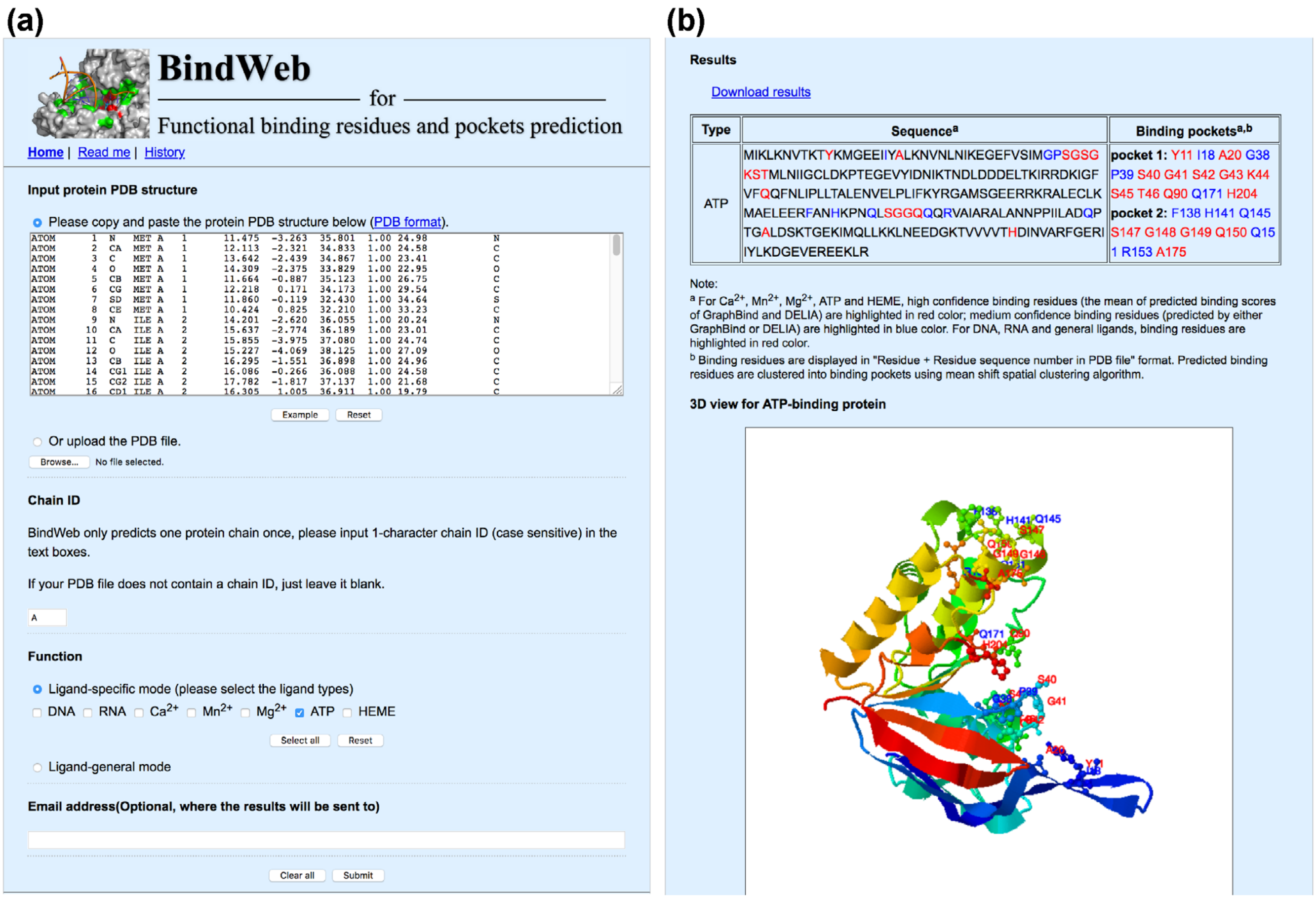
The interface of BindWeb is shown in Figure 2a. The top panel of the web server provides three links to individual pages. Users can submit a new job for ligand-binding residue and pocket prediction with a selected function in homepage. The readme page gives a short description of the methods. To illustrate the usage of BindWeb, we introduce the details of the inputs and outputs with an example input, protein structure 1L2T_A.[35]

3.2 | Inputs

BindWeb accepts a protein structure in PDB format[27] and a unique chain ID of the PDB structure as the input. If users want to predict binding residues for one of the following seven ligands, then the ligand-specific mode should be chosen; otherwise, the ligand-general mode can be a choice. In the ligand-specific mode, users can select one or multiple ligands simultaneously. For DNA and RNA, GraphBind is applied for binding residue prediction. For the other five ligands, Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME, BindWeb uses GraphBind and DELIA as the prediction engines and displays the integrated results. Finally, users can optionally provide an email address to receive the prediction results.

3.3 | Outputs

After a job is submitted on the frontend, the uploaded protein structure and other inputs are delivered to the backend program. When a new job is detected, the backend program can automatically infer the ligand binding residues and binding pocket(s). The backend program extracts the sequence and structure features of the protein, predicts binding residues with the optimized model parameters and clusters the predicted binding residues into binding pocket(s) for each selected ligand. When the job is finished, the results will be shown on the result page, and the web server will automatically send the results to the provided email address.



Input protein PDB structure

● Please copy and paste the protein PDB structure below (PDB format).

```
ATOM      1  N   MET A   1      11.475  -3.263  35.801  1.00 24.98           N
ATOM      2  CA  MET A   1      12.113  -2.321  34.833  1.00 24.58           C
ATOM      3  C   MET A   1      13.642  -2.439  34.867  1.00 23.41           C
ATOM      4  O   MET A   1      14.309  -2.375  33.829  1.00 22.95           O
ATOM      5  CB  MET A   1      11.664  -0.887  35.123  1.00 26.75           C
ATOM      6  CG  MET A   1      12.218   0.171  34.173  1.00 29.54           C
ATOM      7  SD  MET A   1      11.860  -0.119  32.430  1.00 34.64           S
ATOM      8  CE  MET A   1      10.424   0.825  32.210  1.00 33.23           C
ATOM      9  N   ILE A   2      14.201  -2.620  36.055  1.00 20.24           N
ATOM     10  CA  ILE A   2      15.637  -2.774  36.189  1.00 23.01           C
ATOM     11  C   ILE A   2      15.855  -3.975  37.080  1.00 24.74           C
ATOM     12  O   ILE A   2      15.227  -4.069  38.125  1.00 27.09           O
ATOM     13  CB  ILE A   2      16.295  -1.551  36.898  1.00 24.96           C
ATOM     14  CG1 ILE A   2      16.086  -0.266  36.088  1.00 24.58           C
ATOM     15  CG2 ILE A   2      17.782  -1.817  37.137  1.00 21.68           C
ATOM     16  CD1 ILE A   2      16.305   1.005  36.911  1.00 19.79           C
```

○ Or upload the PDB file.

No file selected.

Chain ID

BindWeb only predicts one protein chain once, please input 1-character chain ID (case sensitive) in the text boxes.

If your PDB file does not contain a chain ID, just leave it blank.

A

Function

● Ligand-specific mode (please select the ligand types)

□ DNA □ RNA □ Ca^{2+} □ Mn^{2+} □ Mg^{2+} ☑ ATP □ HEME

○ Ligand-general mode

Email address(Optional, where the results will be sent to)

Results

Type	Sequence[a]	Binding pockets[a,b]
ATP	MIKLKNVTKTYKMGEEIIYALKNVNLNIKEGEFVSIMGPSGSGKSTMLNIIGCLDKPTEGEVYIDNIKTNDLDDDELTKIRRDKIGFVFQQFNLIPLLTALENVELPLIFKYRGAMSGEERRKRALECLKMAELEERFANHKPNQLSGGQQQRVAIARALANNPPIILADQPTGALDSKTGEKIMQLLKKLNEEDGKTVVVVTHDINVARFGERIIYLKDGEVEREEKLR	**pocket 1:** Y11 I18 A20 G38 P39 S40 G41 S42 G43 K44 S45 T46 Q90 Q171 H204 **pocket 2:** F138 H141 Q145 S147 G148 G149 Q150 Q151 R153 A175

Note:

[a] For Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP and HEME, high confidence binding residues (the mean of predicted binding scores of GraphBind and DELIA) are highlighted in red color; medium confidence binding residues (predicted by either GraphBind or DELIA) are highlighted in blue color. For DNA, RNA and general ligands, binding residues are highlighted in red color.

[b] Binding residues are displayed in "Residue + Residue sequence number in PDB file" format. Predicted binding residues are clustered into binding pockets using mean shift spatial clustering algorithm.

3D view for ATP-binding protein

FIGURE 2 The input page (a) and output page (b) of BindWeb for predicting ATP-binding residues and pockets in the structure 1L2T_A

Figure 2b shows an example result page, which includes a "Download results" link, a result table, and the visualization of binding residues on the protein structure. In the downloaded result file, for each row, the columns stand for the residue number in the sequence, residue name, binding residue predictons of GraphBind, DELIA and BindWeb, and binding pocket predictions. As shown in the result table, the first column stands for the ligand type. The protein sequence is given in the second column with the highlighted predicted binding residues. For Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME, high confidence binding residues are highlighted in red, and medium confidence binding residues are highlighted in blue. For DNA, RNA, and general ligands, the predicted binding residues of GraphBind or GraphBind-G are shown in red. The third column shows the predicted binding pockets and the contained residues. Figure 2b shows two predicted ATP-binding pockets consisting of 15 and 10 residues in protein structure 1L2T_A.[35] For the sake of distinction, the binding and nonbinding residues are displayed in cartoon and atom-bond formats, respectively.

4 | RESULTS

4.1 | The integration of GraphBind and DELIA boosts the performance of BindWeb

To investigate the effectiveness of the integration, BindWeb is compared with its base methods GraphBind and DELIA. For the shared five ligands of GraphBind and DELIA, five metrics are employed for performance comparison, including recall (Rec), precision (Pre), F1-score (F1), Matthews correlation coefficient (MCC), and the area under the receiver operator characteristic curve (AUC).[12] Here, BindWeb (high) and BindWeb (medium) stand for the predictions with high and medium confidence, respectively. As shown in Table 1, the AUCs (MCCs) of BindWeb with high confidence are 0.009–0.086 (0.009–0.125) and 0.002–0.018 (0.018–0.078) higher than those of DELIA and GraphBind for the five ligands, respectively. The results demonstrate that integrating the diverse deep learning methods can improve the prediction performance for binding residue prediction. In addition, BindWeb with medium confidence achieves the

TABLE 1 Performance comparison of BindWeb and the two base methods

Dataset	Method	Rec	Pre	F1	MCC	AUC
Ca^{2+}	DELIA[a]	0.182	0.556	0.274	0.313	0.795
	GraphBind[a]	0.325	0.563	0.410	0.420	0.863
	BindWeb (high)[b]	0.321	**0.615**	0.422	**0.438**	**0.881**
	BindWeb (medium)[c]	**0.383**	0.515	**0.439**	0.436	N/A[d]
Mn^{2+}	DELIA	0.513	0.632	0.566	0.565	0.903
	GraphBind	0.563	0.626	0.591	0.588	0.951
	BindWeb (high)	0.516	**0.721**	0.602	0.606	**0.953**
	BindWeb (medium)	**0.645**	0.585	**0.614**	**0.610**	N/A
Mg^{2+}	DELIA	0.143	**0.562**	0.228	0.280	0.780
	GraphBind	0.259	0.410	0.317	0.320	0.827
	BindWeb (high)	0.227	0.547	0.321	**0.349**	**0.841**
	BindWeb (medium)	**0.287**	0.394	**0.332**	0.331	N/A
ATP	DELIA	0.642	**0.758**	0.695	0.685	0.947
	GraphBind	0.603	0.666	0.631	0.616	0.939
	BindWeb (high)	0.705	0.711	**0.708**	**0.694**	**0.956**
	BindWeb (medium)	**0.786**	0.583	0.670	0.660	N/A
HEME	DELIA	0.648	**0.660**	0.654	0.628	0.951
	GraphBind	0.775	0.610	0.682	0.661	0.962
	BindWeb (high)	0.787	**0.660**	**0.718**	**0.698**	**0.973**
	BindWeb (medium)	**0.852**	0.554	0.672	0.659	N/A

Note: Bold face indicates the method yields the best result across the compared methods.

[a] Results of DELIA and GraphBind are directly from the original studies.[12,25]

[b] The BindWeb predicts binding residues with high confidence.

[c] The BindWeb predicts binding residues with medium confidence.

[d] Given that the BindWeb (medium) directly pools binding residues predicted by either GraphBind or DELIA to generate the binary results, it does not provide predicted binding probability for AUC calculation.

highest recall for all the five ligands, since BindWeb with medium confidence collects the predicted binding residues of GraphBind and DELIA and results in more binding residues.

4.2 | Case study

In this section, we present a HEME-binding protein structure 6BME[36] as a case study to demonstrate the usage and characteristics of BindWeb. We download the structure 6BME from the PDB,[27] upload it to BindWeb, and select chain A. Then, we choose the ligand-specific mode and select the ligand HEME. Finally, we fill in the email address for receiving results. After the job is submitted, the web page reloads every 10 seconds until it automatically redirects to the result page once the job is finished. Users can bookmark this page to check the results later.

Figure 3a–d visualize the predictions of the chain A in the structure 6BME by DELIA, GraphBind, and BindWeb with high and medium confidence, respectively. The chain A of the structure 6BME has 127 residues, including a HEME-binding pocket with 17 binding residues. Although the binding pockets predicted by the four methods overlap with the native binding pocket, there are still some differences. Specifically, the base methods DELIA and GraphBind predict 22 and 19 binding residues, including 13 and 12 true positive residues as well as 9 and 7 false positive residues, respectively. As Figure 3d shows, BindWeb with median confidence predicts 14 true positive residues of the pooled 24 predicted binding residues from DELIA and GraphBind, increasing the recall from 0.765 and 0.706 of DELIA and GraphBind to 0.824. Furthermore, BindWeb with high confidence reduces 24 predicted binding residues with medium confidence to 21 predicted binding residues with high confidence. Of them, 14 are true positive residues and 7 are false positive residues. As shown in Figure 3, compared to other three methods, BindWeb with high confidence yields an improvement of 0.059–0.083 on MCC and 0.014–0.059 on AUC.

4.3 | Runtime of BindWeb

BindWeb's runtime is affected by its two base predictors. For ligand-general mode and ligand-specific mode for nucleic acids, the runtime of BindWeb is determined by GraphBind. For the other five ligands, the runtime is determined by DELIA. For a protein with 100 residues, GraphBind and DELIA take approximately 10 minutes and an hour, respectively. The runtime of GraphBind is mainly spent on calculating the evolutionary conversation profiles with sequence alignment tools PSI-BLAST and HHblits. Although time-consuming, the experiments

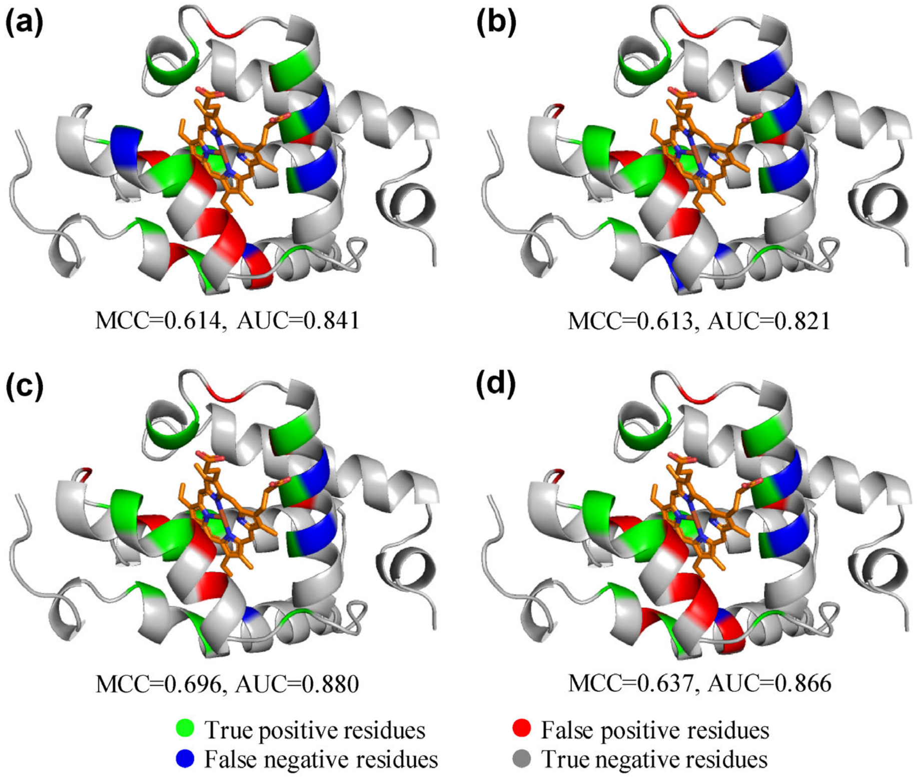


FIGURE 3 (a)–(d) shows the ligand binding residue predictions of the chain A in protein structure 6BME by DELIA, GraphBind, BindWeb with high and medium confidence, respectively.

prove that the evolutionary conversation profiles boost the performance of GraphBind.[12]In addition to PSI-BLAST and HHblits, DELIA uses sequence-template-based S-SITE for binding propensity prediction, resulting in a longer runtime. A template-based method typically takes a long time to search the query protein against templates. For example, the consensus method COACH combines five predictors, including a sequence profile alignment S-SITE, three structural-templated-based predictors TM-SITE, FINDSITE, and COFACTOR, and an ab initio predictor ConCavity. Due to the homologous template alignments, the web server of COACH takes several hours for a single protein.[19]

5 | THE ADVANTAGES AND LIMITATIONS OF BINDWEB

BindWeb combines GNN-based GraphBind and biLSTM-CNN-based DELIA for ligand binding residue prediction with two integration strategies, and the experimental results demonstrate its superiority. By averaging the predictions of the two base methods, BindWeb with high confidence yields higher MCCs and AUCs than the base methods. Meanwhile, recall of BindWeb with medium confidence is increased by pooling the predicted binding residues of base methods. In addition, BindWeb applies mean shift clustering to identify binding pocket(s) based on spatial coordinates of predicted binding residues.

In the future, BindWeb will be continuously upgraded to address its limitations. Given that most ligands do not have a sufficient number of binding proteins for building a ligand-specific method, BindWeb only covers seven specific ligands. We will investigate the few-shot learning or transfer learning for ligand-specific methods to cover more ligands with a limited number of binding proteins. In addition, since many proteins only have primary sequences but no experimental structures, we expect to figure out new algorithms for binding residue prediction from protein structures predicted with the structure prediction algorithms, such as AlphaFold2[37] and RoseTTAFold.[38] Besides, the prediction speed of BindWeb is still relatively slow, and we expect to accelerate it in the future.

6 | CONCLUSION

BindWeb is a user-friendly web server for structure-based ligand binding residue and pocket prediction. It supports two functional modes: ligand-specific binding residue prediction for seven specific ligands (i.e., DNA, RNA, Ca^{2+}, Mn^{2+}, Mg^{2+}, ATP, and HEME) and ligand-general binding residue prediction. BindWeb integrates GNN-based GraphBind and biLSTM-CNN-based DELIA for binding residue prediction and provides a new function for clustering predicted binding residues into binding pocket(s). The experimental results verify that the combination of complementary base methods improves the prediction performance of BindWeb. In the future, we will continuously upgrade the BindWeb from datasets, functions, and algorithms.

AUTHOR CONTRIBUTIONS

Ying Xia: Data curation (equal); investigation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (lead). **Chunqiu Xia:** Data curation (equal); investigation (equal); methodology (equal); software (equal); validation (equal); writing – review and editing (equal). **Xiaoyong Pan:** Conceptualization (equal); formal analysis (equal); funding acquisition (equal); project administration (equal); resources (equal); supervision (equal); writing – review and editing (equal). **Hong-Bin Shen:** Conceptualization (equal); formal analysis (equal); funding acquisition (equal); project administration (equal); resources (equal); supervision (equal); writing – review and editing (equal).

FUNDING INFORMATION

This work was supported by the National Natural Science Foundation of China (61725302, 62073219, and 61903248), and the Science and Technology Commission of Shanghai Municipality (20S11902100 and 22511104100).

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

http://www.csbio.sjtu.edu.cn/bioinf/BindWeb/

ORCID

Ying Xia https://orcid.org/0000-0001-8437-0604
Hong-Bin Shen https://orcid.org/0000-0002-4029-3325

REFERENCES

1. Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, Ohta K. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature. 2008;456(7218):130–134.
2. Asselah T, Durantel D, Pasmant E, Lau G, Schinazi RF. COVID-19: Discovery, diagnostics and drug development. J Hepatol. 2021;74(1):168–184.
3. Scrofani SB, Wright PE, Dyson HJ. The identification of metal-binding ligand residues in metalloproteins using nuclear magnetic resonance spectroscopy. Protein Sci. 1998;7(11):2476–2479.

4. Qian Y, Li X, Wu J, Zhou A, Xu Z, Zhang Q. Picture-word order compound protein interaction: Predicting compound-protein interaction using structural images of compounds. J Comput Chem. 2022;43(4):255–264.
5. Pan X, Yang Y, Xia CQ, Mirza AH, Shen HB. Recent methodology progress of deep learning for RNA–protein interaction prediction. Wiley Interdiscip Rev RNA. 2019;10(6):e1544.
6. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. Protein Sci. 1997;6(3):524–533.
7. Ertekin A, Nussinov R, Haliloglu T. Association of putative concave protein-binding sites with the fluctuation behavior of residues. Protein Sci. 2006;15(10):2265–2277.
8. Yu D-J, Hu J, Yang J, Shen HB, Tang J, Yang JY. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(4):994–1008.
9. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. Nucleic Acids Res. 2017;45(10):e84.
10. Zhang J, Kurgan L. SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019;35(14):i343–i353.
11. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. Brief Bioinform. 2018;19(5):821–837.
12. Xia Y, Xia CQ, Pan X, Shen HB. GraphBind: Protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Res. 2021;49(9):e51.
13. Kriválk R, Hoksza D. P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018;10(1):1–12.
14. Su H, Liu M, Sun S, Peng Z, Yang J. Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. Bioinformatics. 2019;35(6):930–936.
15. Aggarwal R, Gupta A, Chelur V, Jawahar CV, Priyakumar UD. Deeppocket: Ligand binding site detection and segmentation using 3d convolutional neural networks. J Chem Inf Model. 2021. https://doi.org/10.1021/acs.jcim.1c00799.
16. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, de Fabritiis G. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33(19):3036–3042.
17. Halperin I, Glazer DS, Wu S, Altman RB. The FEATURE framework for protein function annotation: Modeling new functions, improving performance, and extending to novel applications. BMC Genomics. 2008;9(2):1–14.
18. Lam JH, Li Y, Zhu L, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun. 2019;10(1):1–13.
19. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics. 2013;29(20):2588–2595.
20. Skolnick J, Brylinski M. FINDSITE: A combined evolution/-structure-based approach to protein function prediction. Brief Bioinform. 2009;10(4):378–391.
21. Zhu Y-H, Hu J, Song XN, Yu DJ. DNAPred: Accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. J Chem Inf Model. 2019;59(6):3057–3071.
22. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10(1):1–11.
23. Hu J, Li Y, Zhang Y, Yu DJ. ATPbind: Accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons. J Chem Inf Model. 2018;58(2):501–510.
24. Hu X, Dong Q, Yang J, Zhang Y. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. Bioinformatics. 2016;32(21):3260–3269.
25. Xia C-Q, Pan X, Shen H-B. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. Bioinformatics. 2020;36(10):3018–3027.
26. Dou Y, Wang J, Yang J, Zhang C. L1pred: A sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. PLoS One. 2012;7(4):e35666.
27. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–242.
28. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: A web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680–682.
29. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–540.
30. 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):3389–3402.
31. Remmert M, Biegert A, Hauser A, Söding J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012;9(2):173–175.
32. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: A protein structure and structural feature prediction server. Nucleic Acids Res. 2005;33(suppl_2):W72–W76.
33. Schueler-Furman O, Baker D. Conserved residue clustering and protein structure prediction. Proteins. 2003;52(2):225–235.
34. Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell. 1995;17(8):790–799.
35. Smith PC, Karpowich N, Millen L, et al. ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. Mol Cell. 2002;10(1):139–149.
36. Johnson EA, Russo MM, Nye DB, Schlessman JL, Lecomte JTJ. Lysine as a heme iron ligand: A property common to three truncated hemoglobins from Chlamydomonas reinhardtii. Biochim Biophys Acta Gen Subj. 2018;1862(12):2660–2673.
37. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
38. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–876.

How to cite this article: Xia Y, Xia C, Pan X, Shen H-B. BindWeb: A web server for ligand binding residue and pocket prediction from protein structures. Protein Science. 2022;31(12):e4462. https://doi.org/10.1002/pro.4462