Analytical Biochemistry 705 (2025) 115929

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Improving protein-protein interaction site prediction using graph neural network and structure profiles

Qing Zhang [a], You-Hang Hu [a], Yu Zhou [a], Jun Hu [b,*], Xiao-Gen Zhou [a,**], Biao Zhang [a,b,***]

[a] College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China
[b] Center for AI and Computational Biology, Suzhou Institute of Systems Medicine, Suzhou, 215123, China

ARTICLE INFO

Keywords:
Protein-protein interaction site prediction
Deep learning
Ensemble strategy

ABSTRACT

Protein-protein interactions (PPIs) play a pivotal role in numerous biological processes. Accurate identification of the amino acid residues involved in these interactions is essential for understanding the functional mechanisms of proteins. To effectively integrate both structure and sequence information, we propose a new interaction site predictor, TargetPPI, which leverages bidirectional long short-term memory networks (Bi-LSTM), convolutional neural networks (CNN), and Edge Aggregation through Graph Attention layers with Node Similarity (EGR-NS) neural networks. In TargetPPI, CNN and Bi-LSTM are first employed to extract the global and local feature information, respectively. The combination of global and local features is then used as node embeddings in the graph derived from the protein structure. We have also extracted six discriminative structural features as edge features in the graph. Additionally, a mean ensemble strategy is used to integrate multiple prediction models with diverse model parameters into the final model, resulting in more accurate PPIs prediction performance. Benchmarked results on seven independent testing datasets demonstrate that, compared to most of the state-of-the-art methods, TargetPPI achieves higher accuracy, precision, and Matthews Correlation Coefficient (MCC) values on average, specifically, 84.3 %, 57.6 %, and 0.383, respectively. The source code of TargetPPI is freely available at https://github.com/bukkeshuo/TargetPPI.

1. Introduction

Proteins serve as fundamental building blocks of life and essential organic components of cells, playing critical roles in processes such as DNA replication, gene expression, protein synthesis, intercellular signaling, and material transport [1]. Their functions are primarily mediated through protein-protein interactions (PPIs), which involve highly specific physical contacts between protein molecules at their interface residues through chemical bonds [2,3]. Accurate identification of PPIs sites is pivotal for analyzing protein functions and facilitating drug design [4,5]. The rapid advancement of biological experiments has resulted in the accumulation of extensive PPIs data in the Protein Data Bank(PDB) database [2]. However, due to the high cost and time-consuming nature of wet-lab techniques, they are insufficient to meet the growing demand for PPIs site prediction. Consequently, the development of efficient and automated computational methods for PPIs site prediction has emerged as a critical research direction in bioinformatics [5,6].

Over the past two decades, significant advancements have been made in the prediction of protein-protein interaction sites, which can primarily be categorized into sequence-based and structure-based methods [6,7]. In recent years, sequence-based PPIs site prediction methods have made remarkable progress in both prediction accuracy and generalization capability, particularly with increasing application of sequence-derived features and advancements in deep learning techniques. Sequence-derived features have been widely used as embeddings for PPIs site prediction. One of the most commonly used features is Position-Specific Scoring Matrix (PSSM), which is employed in DLPred [8], DELPHI [6], ProB-Site [9], HN-PPISP [10], D-PPIsite [5], and SENSDeep [11]. In addition to PSSM, other features generated through prediction, such as secondary structure and Relative Solvent Accessibility, have also been applied in PPIs site prediction. In DeepPPISP [4], ProB-site [9], and HN-PPISP [10], the secondary structure predicted by DSSP [12] program is utilized as an input feature for the models. In

* Corresponding author.
** Corresponding author.
*** Corresponding author. College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China.
E-mail addresses: junh_cs@126.com (J. Hu), zxg@zjut.edu.cn (X.-G. Zhou), zhangb@ism.cams.cn (B. Zhang).

https://doi.org/10.1016/j.ab.2025.115929
Received 25 January 2025; Received in revised form 18 June 2025; Accepted 27 June 2025
Available online 28 June 2025

contrast, DELPHI [6] and SENSDeep [11] employ the ASAquick [13] program to predict the relative solvent accessibility of protein residues, which is then used as an input feature. In recent years, with the rise of natural language processing, protein language models have garnered widespread attention. The embeddings generated by language models have also been applied to PPIs site prediction. For example, PITHIA [14] and Seq-Insite [15] leverage MSA Transformer [16] to generate 768-dimensional embeddings as input features. Additionally, EnsemPPIS [17] and Seq-Insite [15] also adopt 1024-dimensional embeddings generated by ProtT5-XL [18] as input features. The embeddings generated by protein language models have further improved the predictive accuracy of these models. Although sequence-derived features have made substantial progress in PPIs site prediction, these methods primarily rely on sequence information, overlooking spatial information. For instance, due to protein folding, residues that are adjacent in the sequence may be spatially distant in the three-dimensional structure, which can affect prediction accuracy.

To fully leverage structural information for predicting PPIs sites, graph neural networks are widely applied to explore spatial information from protein structures [19]. For instance, GraphPPIS [20] proposes a GCN framework that combines protein sequence features (e.g., PSSM and HMM) with structural features (e.g., DSSP-derived attributes from PDB files) as node attributes, and constructs protein graphs by generating an adjacency matrix based on the $C\alpha$ distance threshold between residues in PDB files. RGN [21] introduces residual-based graph attention and convolutional networks, using PSSM, HMM, hydrogen bond estimation, and ProtT5-XL [18] embeddings as node features, while constructing the protein graph based on the spatial distances between amino acid residues calculated from the PDB. EGRET [19] extends the Graph Attention Network (GAT) by incorporating edge information, proposing the Edge-Aggregated Graph Attention Network. This method integrates edge importance into the graph attention mechanism, highlighting the crucial role of edge information in the learning process of the model. Node features are represented embeddings generated by ProtT5-XL [18], while edge features are defined by distance and angle matrices calculated from PDB file. Similar to previous models, the protein graph is constructed based on spatial distances between amino acid residues. In graph neural network-based methods, both node and edge information are essential [19]. The edge information is derived from real structural data, while node features primarily come from sequence-derived data and may contain more noise. However, in the EGRET method, treating edge information and node information equivalently can lead to the model overly emphasis on edge information, while overlooking the noise component in the node features, thereby compromising the prediction accuracy of the model. Furthermore, the extraction of node features does not fully account for the impact of local contextual information on residue features, thereby limiting improvements in prediction performance. Currently, most structure-based PPIs prediction methods rely on high-resolution experimental protein structures from the PDB database. However, these structures are inadequate for fulfilling the requirements of PPIs site prediction and fail to fully exploit recent advancements in the field of protein structure prediction, such as: trRosetta [22], I-TASSER [23], and AlphaFold3 [24]. This limitation hinders further enhancement of PPIs site prediction performance.

Inspired by EGRET, we propose a new graph neural network-based method for predicting PPIs sites, named TargetPPI. Specifically, we introduce the Edge Aggregated Graph Attention Network with Node Similarity [25] (EGR-NS). EGR-NS enhances node associations by calculating node similarity, thereby improving the extraction of informative node features. To fully exploit the learning potential of EGN-NS, we incorporate local sequence context information into the residue features to enhance node representations. Additionally, new distance and angle information are integrated into the edge features, as detailed in Section 2.4. Given the limited availability of experimental structural data, we supplement the training dataset with protein structures predicted by AlphaFold2 [26], which enhances the structural learning capacity of the model. To further refine model performance, we employ a mean ensemble strategy, integrating nine deep learning models to generate the final predictive model. The model inputs include ProtT5-XL [18] generated embeddings and six sequence-derived features, including PSSM, physicochemical properties, physical characteristics, hydropathy, PKX [8] and positional information. These features are processed and integrated into node representations. The protein graph is constructed based on residue distance information obtained from PDB files, with residue distance and angle matrices serving as edge features. EGR-NS processes these node and edge features to predict the probability of each residue being a PPIs site. Benchmark tests demonstrate that TargetPPI outperforms most existing PPIs site prediction methods in terms of performance.

Table 1
The information of the training and test datasets, including the number of proteins, residues, binding sites, and non-binding sites.

Dataset	Proteins sequences	Total residues	Binding sites	Non-binding sites	% Rate of Binding sites
Dset_9982	9982	4 254 198	427 687	3 826 511	10.1 %
Dset_186	186	36 219	5517	30 702	15.2 %
Dset_164	164	33 681	6096	27 585	18.1 %
Dset_72	72	18 140	1923	16 217	10.6 %
Dset_60	60	13 144	2075	11 069	15.78 %
Dset_70	70	11 791	2332	9459	19.78 %
Dset_448	448	116 500	15 810	100 690	13.6 %
Dset_355	355	95 940	11 467	84 473	12.0 %

2. Materials and methods

2.1. Benchmark data sets

In this study, we evaluate the prediction performance of TargetPPI using seven independent test datasets: Dset_186 [27],Dset_164 [28], Dset_72 [27], Dset_60 [20], Dset_70 [4], Dset_448 [29], and Dset_355 [6]. Dset_186 consists of 186 protein sequences derived from 105 heterodimeric protein complexes, all exhibiting a sequence identity of less than 25 %. Dset_72, comprising 72 protein sequences, is curated by applying BLASTClust [30] to remove sequences with sequence identity of $\geq$25 % to those in Dset_186. Dset_164 includes 164 protein sequences obtained from newly annotated protein entries in the Protein Data Bank between June 2010 and November 2013. The same stringent filtering criteria applied to Dset_186 and Dset_72 are also employed in the construction of Dset_164. To further enhance dataset diversity and coverage, GraphPPIS [20] integrates Dset_186, Dset_72, and Dset_164 into a unified dataset. BLASTClust [30] is subsequently used to remove redundant proteins with a sequence identity exceeding 25 % or overlapping regions greater than 90 %, resulting in 395 non-redundant protein chains. Of these, 335 are designated for the training set, while the remaining 60 protein chains are reserved for the independent test set, Dset_60. Similarly, DeepPPISP [4] combines Dset_186, Dset_72, and Dset_164 into a comprehensive dataset, which is then randomly splits into training and testing subsets with an 83.3 % training ratio. The resulting test set, consisting of 70 protein chains, is designated as Dset_70. Dset_448, which consists of 448 protein sequences, is sourced from the BioLiP database [31]. The sequence identity between any two sequences in Dset_448 is less than 25 %. To reduce redundancy, Dset_355 is derived from Dset_448 by removing 93 proteins that exhibit over 40 % sequence identity with the training dataset of DLPred [8], a process conducted within DELPHI [6].

To train the prediction model, we use a dataset named Dset_9982, originally collected in DELPHI [6]. Dset_9982 comprises 9982 non-redundant protein sequences, including 427 687 binding residues

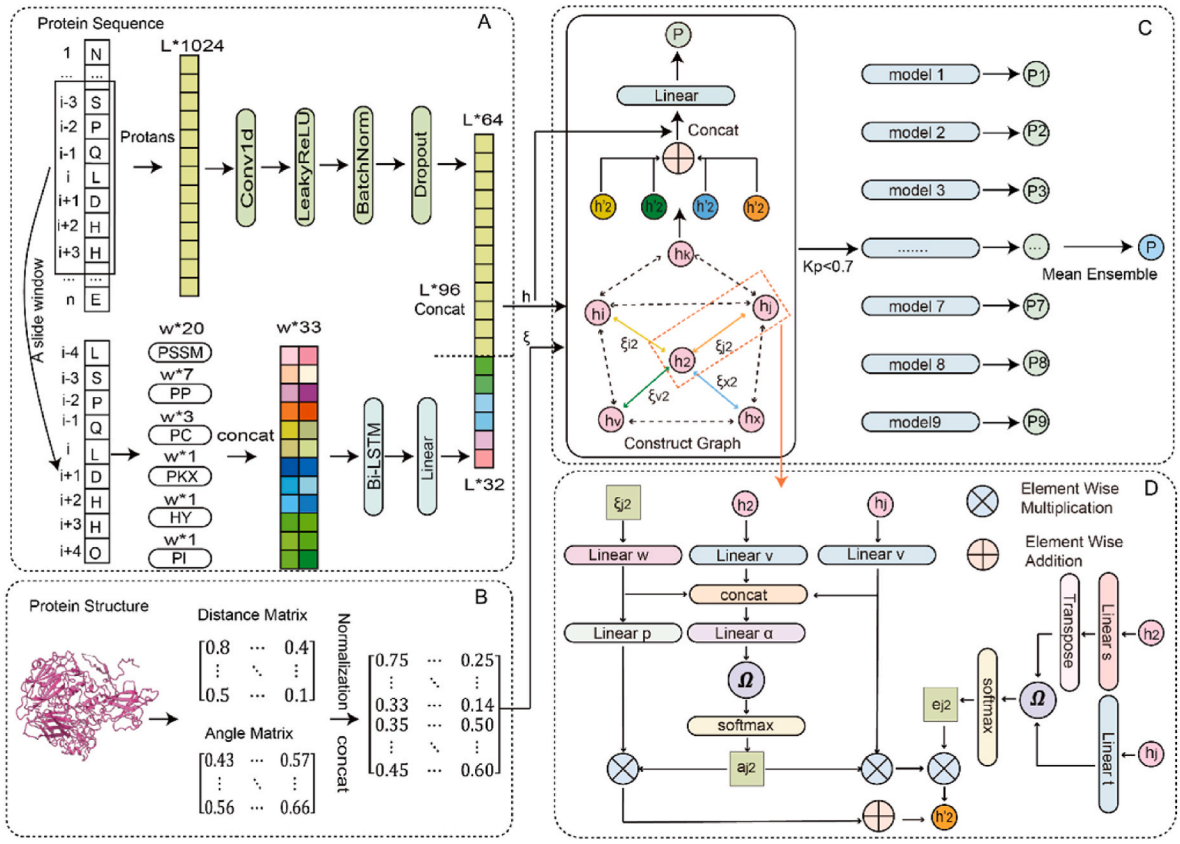

Fig. 1. Architecture of TargetPPI. Panel A illustrates the extraction of global and local features, with detailed descriptions provided in Section 2.3. Panel B presents the computation of protein edge-level feature information, including distance and angle calculations, as described in Section 2.4. Panel C depicts the edge aggregation process through Graph Attention layers incorporating Node Similarity, where a graph is constructed based on the distance information derived in Panel B. The feature data from Panel A are utilized as node information, while the matrix from Panel B serves as edge-level information. Finally, Panel D details the specific computational relationship between node 2 and node *j*.

and 3 826 511 non-binding residues. The maximum sequence identity between any two protein sequences in Dset_9982 is less than or equal to 25 %. Furthermore, the sequence similarity between Dset_9982 and the seven test sets does not exceed 25 %. During training, the training set is randomly divided into 10 subsets, with 9 subsets used for training and 1 subset reserved for validation. This splitting ensures the independence of the training and validation datasets, providing a reliable foundation for model performance evaluation. Detailed information about these datasets is provided in Table 1. Consistent with the approaches outlined in Refs. [4,32] and HN-PPISP [10], we standardize the protein sequence length to 500 by truncating sequences longer than 500 residues and padding shorter sequences with zeros to ensure uniformity.

2.2. Graph representation of proteins

In this study, we represent the three-dimensional structure of the protein as a graph G, where each node corresponds to a protein residue. Each node is connected to k ($k = 21$) neighboring nodes, selected based on the average atomic distance between residues. Specifically, for a given node i, we compute the Euclidean distances between all atom pairs involving this residue and other residues, and take the mean (i.e., the average atomic distance) to quantify their overall spatial proximity. The atomic coordinates are obtained from the three-dimensional protein structure file. Finally, the k residues with the smallest average atomic distances to node i are selected as its neighboring nodes.

2.3. Node-level feature representation

In this study, two types of features, global sequence and local contextual features, are utilized to represent node information. As global sequence features, we employ the ProtT5-XL [18] embeddings generated by the pre-trained protein language model ProtTrans. ProtT5-XL generates protein sequence embeddings in the form of an $L \times 1024$ matrix, where L represents the total number of residues in the protein, and each residue is represented as a 1024-dimensional embedding vector. ProtT5-XL (hereafter referred to as ProtT5) is well-suited for processing long protein sequences or those lacking high-quality multiple sequence alignments, providing high-resolution embedding for individual residues. Previous studies have demonstrated the significant performance of ProtT5 embeddings in PPIs site prediction [14,17,19]. Given its outstanding performance in PPIs site prediction, ProtT5 embeddings are adopted in this study.

For local contextual features, we incorporate six sequence-derived features, PSSM, physicochemical characteristics (PC), physical properties (PP), hydropathy (HY), PKX, and position information (PI). The PSSM is generated using PSI-BLAST [33], with three iterations and an E-value threshold of 0.001 to capture evolutionary information for each amino acid residue. The resulting PSSM is then normalized, representing each residue as a 20-dimensional feature vector.

The other features are derived from established studies [8,27,34] and include: physicochemical properties, each residue is represented by a 3-dimensional vector, incorporating atomic composition, electrostatic charges and potential hydrogen bonding. Physical properties, each residue is represented by a 7-dimensional vector including steric parameters (graph shape index), polarizability, volume (normalized van der Waals volume), hydrophobicity, isoelectric point, helix probability, sheet probability, and polarizability [35]. Hydropathy, a 1-dimensional vector representing the hydrophilicity or hydrophobicity of the amino

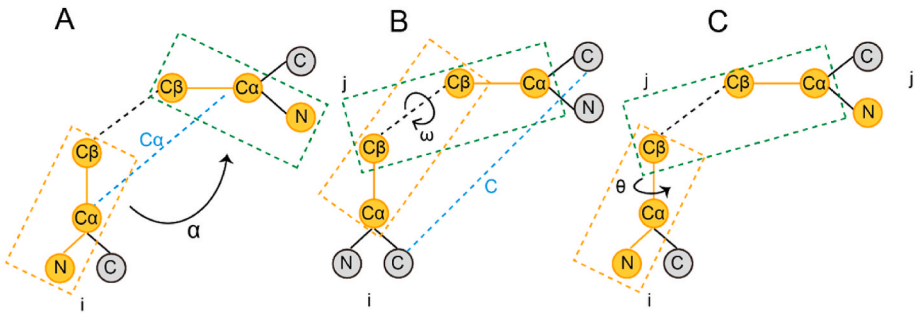

Fig. 2. Graphical representation of edge features. (A) The α angle and $C\alpha$ distance between residues i and j. (B) The ω angle and C distance between residue i and residue j. (C) The θ angle between residues i and j.

acid. PKX, a 1-dimensional vector representing the negative logarithm of the dissociation constant for any other group in the molecule. Position information, inspired by DELPHI [6], each residue is represented by a 1-dimensional vector, where the position of the *i-th* amino acid in the sequence is expressed as i/L, with i being the residue index and L the total number of residues in the protein.

These features are concatenated to construct a feature representation of $L \times 33$, where 33 denotes the dimensionality of the feature vector for each residue. Subsequently, a sliding window approach with a window size of $W = 27$ (Supplementary Text S1 and Table S1) is employed to incorporate contextual information from neighboring residues. This process yields a final representation of $L \times 33 \times W$, where W captures the local sequential context, enabling a more comprehensive characterization of each residue.

2.4. Edge-level feature representation

In the directed graph G representing a protein P, the edge features for an edge from node j to node i are defined as ξ_{ji}. A total of six edge features are utilized, including three distance features and three angle features, as shown in Fig. 2. The distance features include: (1) D_{ij}, the average distance between all atoms of residues i and j, (2) $C_{\alpha ij}$, the distance between the C_α between residues i and j, and (3) C_{ij}, the distance between the carbon atoms of the carboxyl groups of residues i and j. The angle features are defined as follows: (1) α_{ij}, the absolute value of the angle between the surface normal of the planes formed by the C_α atoms, carboxyl carbons, and amino nitrogen of residues i and j, (2) ω_{ij}, the dihedral angle between the planes formed by $C_{\alpha i}$, $C_{\beta i}$, $C_{\beta j}$ of residue i and $C_{\beta i}$, $C_{\beta j}$, $C_{\alpha j}$ of residue j, and (3) θ_{ij}, the dihedral angle between the planes formed by N_i, $C_{\alpha i}$, $C_{\beta i}$ of residue i and $C_{\beta j}$, $C_{\alpha j}$ of residue j. To extend the representation of edge features, we incorporate four additional features inspired by trRosetta [22], building on the EGRET framework. These include two additional distance features (C_α and C) and two additional angle features (ω and θ). This enhancement further enriches the description of edge characteristics and improves the representation of protein structures.

2.5. Architecture of EGR-NS

As shown in Fig. 1, the architecture of EGR-NS can be discussed in two distinct parts: (1) global and local feature extraction, (2) Edge Aggregation through Graph Attention layers with Node Similarity (EAGA-NS).

For global sequence feature extraction, the input tensor has a shape of $L \times 1024$, which is processed through a one-dimensional convolutional neural network (Conv1D) with a kernel size of 21. This convolution operation reduces the dimensionality of the features, yielding an output tensor with a shape of $L \times 64$. For the extraction of local contextual features, the input sequence, with a shape of $L \times W \times 33$ (where w represents the sliding window size), is fed into a bidirectional LSTM layer with 16 units. The Bi-LSTM layer captures the contextual dependencies along the sequence and produces an output tensor of shape $L \times W \times 32$. Subsequently, a linear layer aggregates the features along the w-dimension, resulting in a reduced-dimensional tensor of shape $L \times 32$. Finally, the processed global sequence features with dimensions $R^{L\times 64}$ and the local contextual features with dimensions $R^{L\times 32}$ are concatenated along the feature dimension, forming a final node feature representation $\boldsymbol{H} = \{\boldsymbol{h}_i | i = 1,2,\ldots,L\} \in R^{L\times 96}$. This concatenation effectively integrates complementary information from both global sequence features and local context, enhancing the ability to capture both global sequence information and local contextual. In EAGA-NS, the primary function is to generate updated node features $\widehat{\boldsymbol{h}}_i$ base on the input node feature $\boldsymbol{h}_i$ and edge feature $\boldsymbol{\xi}_{ji}$. (where i represents the *i-th* residue, and j represents residues in the neighborhood of residue i) The updated node features $\widehat{\boldsymbol{h}}_i$ are computed as shown in Equation (1).

$$
\begin{cases}
\widehat{\boldsymbol{h}}_i = \sigma\left(\sum_{j\in N_i} a_{ij}\boldsymbol{w}^v\boldsymbol{h}_j \times \sum_{j\in N_i} t_{ij}\boldsymbol{w}^v\boldsymbol{h}_j + \sum_{j\in N_i} a_{ij}\boldsymbol{w}^e\boldsymbol{\xi}_{ji}\right) \Big\| \boldsymbol{h}_i \\
a_{ij} = softmax_j(e_{ij}) = \dfrac{exp(e_{ij})}{\sum_{k\in N_i} exp(e_{ki})} \\
e_{ij} = \Omega\left(\boldsymbol{w}^a\left[\boldsymbol{w}^v\boldsymbol{h}_i \| \boldsymbol{W}^v\boldsymbol{h}_j \| \boldsymbol{w}^\rho\xi_{ji}\right]\right) \\
t_{ij} = softmax_j(s_{ij}) = \dfrac{exp(s_{ij})}{\sum_{k\in N_i} exp(s_{ki})} \\
s_{ij} = (\boldsymbol{w}^s\boldsymbol{h}_i)^T(\boldsymbol{w}^t\boldsymbol{h}_j)
\end{cases}
\tag{1}
$$

where a_{ij} represents the edge attention distribution, which is derived by normalizing the edge attention coefficients e_{ij} using the *softmax* function. Similarly, t_{ij} represents the node similarity distribution, obtained by normalizing the node similarity coefficients s_{ij} using the *softmax* function. The vectors $\boldsymbol{w}^v$, $\boldsymbol{w}^e$, $\boldsymbol{w}^\rho$, $\boldsymbol{w}^a$, $\boldsymbol{w}^s$, and $\boldsymbol{w}^t$ are learnable parameter during the training process. The updated node features $\widehat{\boldsymbol{h}}_i$, obtained through Equation (1), are passed through a linear layer to output the predicted binding site probability P_i. In contrast to EGRET, we introduce a node similarity function in EAGA-NS for drawing more attention to the discriminative information embedded in the node feature representation.

2.6. Model selection process

To mitigate prediction variance and enhance model performance, this study adopts the model selection strategy from D-PPIsite [5]. This method is selected due to its demonstrated effectiveness in reducing prediction variance and improving generalization in PPIs site prediction. In this study, five independent model trajectories are trained. After eliminating models with an *MCC* below 0.4, a set of 200 candidate models is obtained. The final selection is made by ranking these 23 models according to their *MCC* values on the validation set. The model

with the highest *MCC* is selected as the primary candidate, followed by the others in descending order of their *MCC* scores. This ranking ensures that the selected models not only exhibit strong predictive performance but also possess sufficient diversity, thus facilitating the effectiveness of ensemble learning.

The mean ensemble method is employed to determine the optimal number of models to integrate into TargetPPI. This approach involves testing model quantities ranging from 1 to 23 on the validation set. Ultimately, nine top-performing models are selected for integration based on their performance.

$$out_{min} = \frac{1}{n}\sum_{1}^{n} out_i \tag{2}$$

In this context, out_i denotes the output of the *i-th* selected model, and k_p quantifies the similarity between two models, defined as follows: k_p, represents the similarity score; a, The number of common binding sites predicted by both $model_i$ and $model_j$; b, The number of binding sites predicted by $model_i$ but classified as non-binding by $model_j$; c, The number of binding sites predicted by $model_j$ but classified as non-binding by $model_i$; d, The number of common non-binding sites predicted by both $model_i$ and $model_j$

$$\theta_1 = \frac{a+d}{m} \tag{3}$$

$$\theta_2 = \frac{(a+b)(a+c)(c+d)(b+d)}{m^2} \tag{4}$$

$$k_p = \frac{\theta_1 - \theta_2}{1-\theta_2} \tag{5}$$

2.7. Implementation details

The PyTorch platform (version 1.10.2) is used to implement the training of the TargetPPI model. The hyperparameters are set as follows: the optimizer is Adaptive Moment Estimation (*Adam*), with a learning rate of 0.001, $\beta 1$ and $\beta 2$ are set to 0.9 and 0.99, respectively, the batch size is 100, and the number of epochs is 100. The loss function used is binary cross-entropy (BCE loss).

$$BCEloss = -\frac{1}{N}\sum_{i}^{N}[l_i \log(y_i) + (1-l_i)\log(1-y_i)] \tag{6}$$

where N is the number of samples in the dataset, l_i is the true label of the sample, and y_i is the predicted value from the model. We train five mutually independent training trajectories, and each epoch within each trajectory is evaluated. When the *MCC* exceeds 0.40, the model is saved. As a result, a total of 200 candidate models are saved across the five training trajectories. For these 200 models, we apply the ensemble strategy mentioned in Section 2.6 to obtain the final model. Training is conducted on a GPU (NVIDIA GeForce RTX 3090), with the duration of each independent training session approximately 2 days.

2.8. Evaluation metrics

Five evaluation metrics, accuracy (*ACC*), precision (*PRE*), sensitivity (*SEN*), specificity (*SPE*), *F*1 score (*F*1) and Matthews Correlation Coefficient (*MCC*) are employed for PPIs site prediction. Since PPIs site prediction is an imbalanced classification problem, *F*1 score and *MCC* are key metrics that provide a more comprehensive evaluation.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$PRE = \frac{TP}{TP+FP} \tag{7}$$

Table 2
Global and Local Features ablation results of TargetPPI on the Dset_448 dataset.

Model	feature	ACC	PRE	SPE	SEN	MCC	F1
Global feature	ProtT5	0.871	0.526	0.925	0.526	0.452	0.526
	MSA	0.842	0.412	0.909	0.412	0.321	0.412
Local feature	Without PSSM	0.813	0.309	0.891	0.309	0.201	0.309
	Without PP	0.846	0.433	0.911	0.433	0.344	0.433
	Without HY	0.846	0.433	0.911	0.433	0.344	0.433
	Without PKX	0.844	0.428	0.910	0.428	0.338	0.428
	Without PI	0.839	0.407	0.907	0.407	0.315	0.407
Without local features		0.871	0.526	0.925	0.526	0.452	0.526
Without global features		0.850	0.448	0.913	0.448	0.361	0.448
Global and local features		**0.878**	**0.552**	**0.930**	**0.552**	**0.481**	**0.552**

$$SEN = \frac{TP}{TP+FN} \tag{8}$$

$$SPE = \frac{TN}{TN+FP} \tag{9}$$

$$F1 = \frac{2*SEN*PRE}{SEN+PRE} \tag{10}$$

$$MCC = \frac{TP*TN - FN*FP}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \tag{11}$$

where metrics are calculated based on true positives (*TP*), which refer to residues correctly predicted as interaction sites; true negatives (*TN*), referring to residues correctly predicted as non-interaction sites; false positives (*FP*), which are residues incorrectly predicted as interaction sites; and false negatives (*FN*), referring to residues incorrectly predicted as non-interaction sites. Typically, a threshold is established such that the number of predicted interaction residues aligns with the actual number, thereby ensuring that the *SEN*, *PRE*, and *F*1 score yield identical values. In addition, the Area Under the Receiver Operating Characteristic Curve (*AUC*) and the Area Under the Precision-Recall Curve (*AUPR*) are employed as supplementary evaluation metrics, representing the areas under the *ROC* and *PR* curves, respectively.

3. Results and discussion

3.1. Improving performance by global and local features

To systematically evaluate the contributions of global and local features to the predictive performance of the EGR-NS model, we first analyze the impact of different global features. A comparative study between the ProtT5 feature and the MSA feature reveals that ProtT5 is more effective in encoding global sequence information. Specifically, the use of the ProtT5 leads to improvements of 2.9 % in *ACC*, 11.4 % in *PRE*, 1.6 % in *SPE*, an 11.4 % in *SEN*, a 13.1 % in the *MCC*, and an 11.4 % in the *F*1.

To evaluate the contribution of each local feature, an ablation study is performed by individually removing each feature and assessing the resulting model performance. As shown in Table 2, the exclusion of any single local feature consistently reduces the *MCC* score, suggesting that these features provide complementary information and jointly contribute to improved model performance. Additionally, the effect of varying sliding window sizes on predictive performance is assessed. The experimental results demonstrate that the optimal sliding window size is 27. Detailed data and information are provided in Supplementary Text S1 and Table S1.

To assess the complementary effect of global and local features,

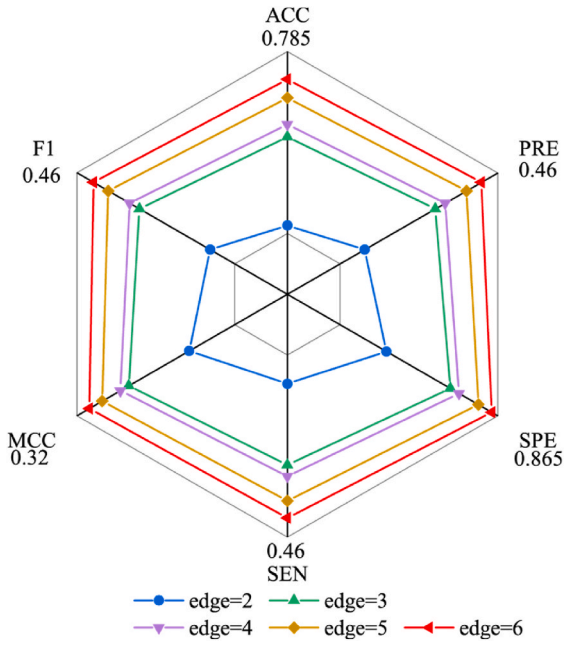

Fig. 3. The impact of different numbers of edge features on the model performance on Dset_70.

Table 3
Performance on Dset_70 with each edge feature used in isolation.

Edge feature	ACC	PRE	SPE	SEN	MCC	F1
D_{ij}	0.760	0.394	0.851	0.394	0.245	0.394
$C_{\alpha ij}$	0.782	0.448	0.864	0.448	0.312	0.448
C_{ij}	0.763	0.402	0.853	0.402	0.254	0.402
α_{ij}	0.769	0.416	0.856	0.416	0.272	0.416
ω_{ij}	0.763	0.403	0.853	0.403	0.255	0.403
θ_{ij}	0.760	0.395	0.851	0.395	0.245	0.395

additional ablation studies are conducted by separately removing either the global or local feature components from the model. The results indicate that the integration of global and local features substantially enhances model predictive performance. When compared to the use of global features alone, the inclusion of local features leads to a 0.7 % increase in *ACC*, a 2.6 % improvement in *PRE*, a 0.6 % increase in *SPE*, a 2.6 % enhancement in *SEN*, a 2.9 % rise in *MCC*, and a 2.6 % increase in the *F*1. Similarly, when compared to the use of local features alone, the same performance metrics demonstrate significant improvement (see Table 2 for detailed information). These findings suggest a complementary relationship between global and local features, where their integration effectively augments the predictive capability of model.

3.2. Performance comparison by different numbers of edge features

To evaluate the impact of the newly introduced edge features on model performance, we conduct a series of controlled experiments based on the two edge features originally used in the EGRET, as shown in Fig. 3. Specifically, we design five experimental settings to assess the contribution of each feature incrementally: (1) Baseline model: Retaining only the two original edge features (D_{ij} and α_{ij}) in EGRET. (2) Adding the $C_{\alpha ij}$ distance feature to enhance backbone connectivity. (3) Introducing the dihedral angle ω_{ij} to refine residue orientation. (4) Incorporating the C_{ij} distance feature to better capture functional group interactions. (5) Adding the θ_{ij} angle feature to improve local geometric representation. As illustrated in Fig. 3, the progressive inclusion of new edge features consistently improves key evaluation metrics including *ACC*, *PRE*, *SPE*, *SEN*, *MCC*, and *F*1. This trend highlights the crucial role of edge information in capturing the spatial organization of proteins and amino acid interactions, ultimately enhancing prediction accuracy. Additionally, we conduct a control experiment where all edge features are removed to assess the overall contribution of edge information. The detailed results are presented in Supplementary Table S5.

In addition to the above progressive experiments, we conducted a comprehensive ablation study in which each of the six edge features was individually evaluated to assess its independent contribution to model performance. The results, presented in Table 3, highlight the varying degrees of effectiveness among these features. (1) When using only the general distance feature D_{ij} the model exhibits the lowest overall performance ($ACC = 0.760$, $MCC = 0.245$, $F1 = 0.394$), as it provides only a coarse measure of residue separation and fails to encode detailed structural interactions necessary for accurate prediction. (2) In contrast, the distance feature $C_{\alpha ij}$ achieves the best performance across all metrics ($ACC = 0.782$, $MCC = 0.312$, $F1 = 0.448$). This highlights the importance of backbone spatial proximity and connectivity, which are crucial for modeling residue-residue interactions. (3) The carboxyl carbon distance feature C_{ij} also contributes positively to prediction performance ($ACC = 0.763$, $MCC = 0.254$, $F1 = 0.402$), as it provides additional geometric information relevant to residue interaction sites. (4) The angle feature α_{ij} representing the orientation between residues, leads to a moderate improvement ($ACC = 0.769$, $MCC = 0.272$, $F1 = 0.416$), suggesting its role in enhancing the structural understanding of local residue arrangements. (5) The dihedral angle feature ω_{ij} which reflects torsional relationships between side-chain atoms, further improves performance ($ACC = 0.763$, $MCC = 0.255$, $F1 = 0.403$), indicating its utility in capturing local conformational properties.(6) Lastly, the angle θ_{ij} while similar in performance to D_{ij} ($ACC = 0.760$, $MCC = 0.245$, $F1 = 0.395$), provides an additional layer of local geometric context, though its impact appears limited when used in isolation.

These findings suggest that while each edge feature offers distinct structural or geometric information, their individual predictive capabilities vary significantly. Notably, $C_{\alpha ij}$ emerges as the most informative standalone feature. Moreover, we further examine the performance of model when using both real and predicted protein structures as inputs; the detailed results of this comparison are provided in Supplementary Text S2, Table S2, and Fig. S1.

3.3. EAGN-NS demonstrates superior performance compared to EGRET

To ensure a fair comparison with EGRET, we reproduce its training process on the same dataset and evaluate its predictive performance on the test set. EAGN refers to the network framework layer of EGRET. This study introduces Node Similarity (NS) into the EAGN structure, constructing EAGN-NS. Under otherwise identical conditions, we conduct experiments by solely incorporating node similarity to verify its effectiveness. Unlike EGRET, EAGN-NS enhances the attention mechanism by incorporating node similarity. Specifically, the similarity between two residues is computed using the cosine similarity of their feature representations, reflecting their functional resemblance. The resulting similarity score is integrated into the attention computation to modulate the aggregation weights. This modification enables the model to emphasize interactions between functionally similar residues, allowing it to more effectively identify key interaction sites and improve prediction performance.

Consequently, compared to EAGN, EAGN-NS demonstrates significant improvements across multiple key evaluation metrics: *PRE* increases by 2.9 %, *MCC* improves by 3.6 %, *AUC* rises by 1.6 %, and *AUPR* increases by 2.3 %. These results indicate that incorporating node similarity enhances the ability of the model to learn from structural and sequential patterns, thereby improving the accuracy of PPIs site prediction. Additionally, we evaluate the performance of three other network frameworks: CNN, Transformer, and GCN. The experimental results show that EAGN-NS outperforms commonly used architectures

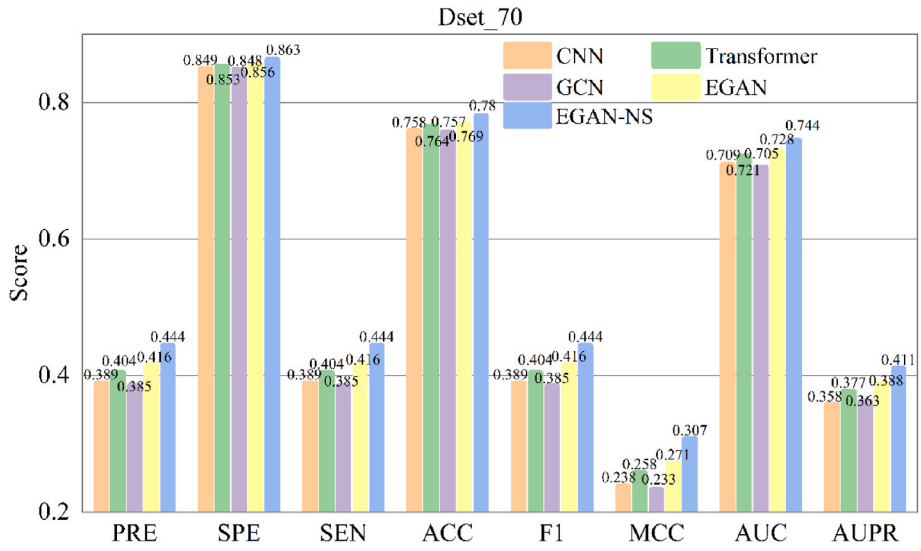

Fig. 4. Performance metrics of different network frameworks on Dset_70.

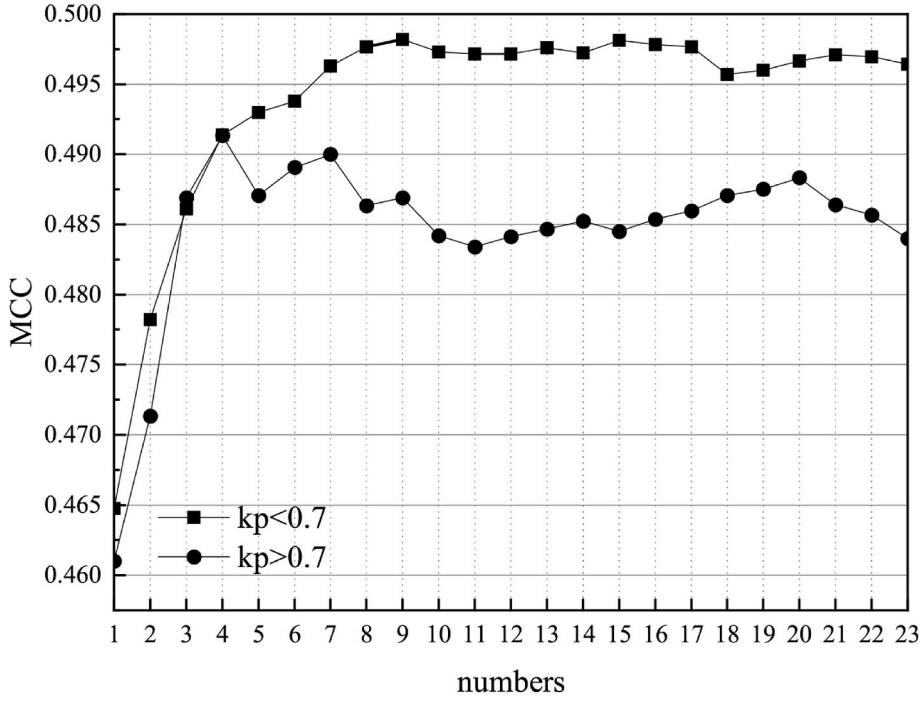

Fig. 5. Performance metrics of $k_p > 0.7$ and $k_p < 0.7$ on Dset_448.

such as Transformer, CNN, and GCN in predicting PPI sites (as shown in Fig. 4). Specific numerical data are provided in Supplementary Table S6. The performance comparison of different frameworks across multiple datasets can be found in the Supplementary Fig. S2–S7. Additionally, to evaluate the robustness of our preprocessing strategy, we perform experiments across three datasets (Dset_186, Dset_60, and Dset_448) with varying sequence lengths. The results demonstrate stable model performance and are presented in Supplementary Text S5 and Table S9. Furthermore, we compare the performance of Bi-LSTM and standard LSTM [36] within the EGR-NS framework to assess the benefits of bidirectional sequence modeling. Detailed analysis and corresponding metrics are provided in Supplementary Text S6 and Table S8.

3.4. *Improving performance by varying the number of integrated models*

In this section, to evaluate the effect of integrating different numbers of models on prediction performance, we adopt the mean ensemble strategy introduced in D-PPIsite [5]. Initially, a total of 500 individual models are trained. Among these, 300 models with *MCC* lower than 0.40 are discarded, yielding 200 candidate models (as described in Section 2.4). From the pool of 200 models, 46 models are randomly selected and split into two groups, each containing 23 models. The grouping is based on the pairwise similarity (k_p) between models: in the first group, the similarity between any two models is less than 0.7, whereas in the second group, the pairwise similarity exceeds 0.7. Within each group, the models are ranked according to their *MCC* values on the validation set, with the highest-*MCC* model ranked first and others sorted in descending order accordingly.

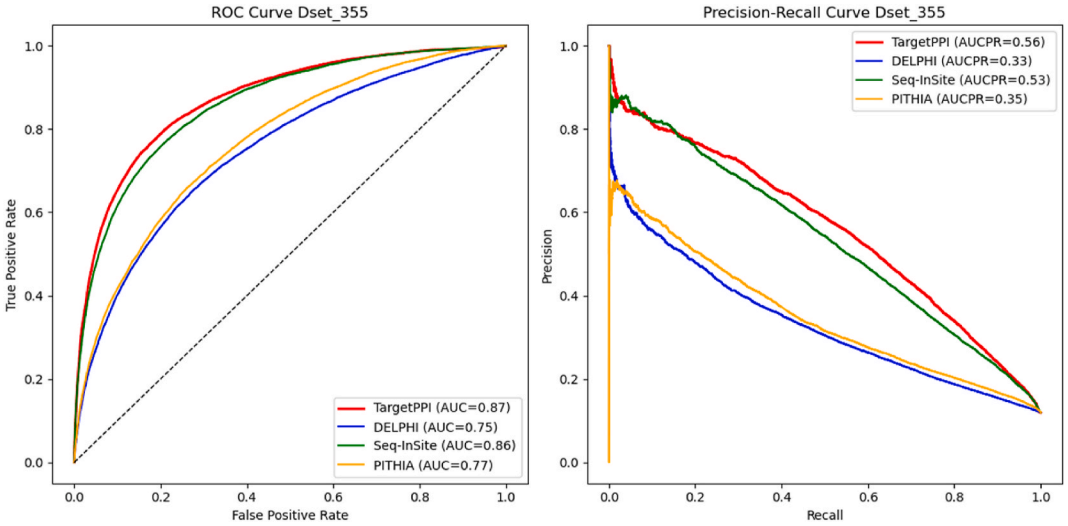

Fig. 6. Comparison of ROC and precision-recall curves between TargetPPI and other sequence-based methods on Dset_355.

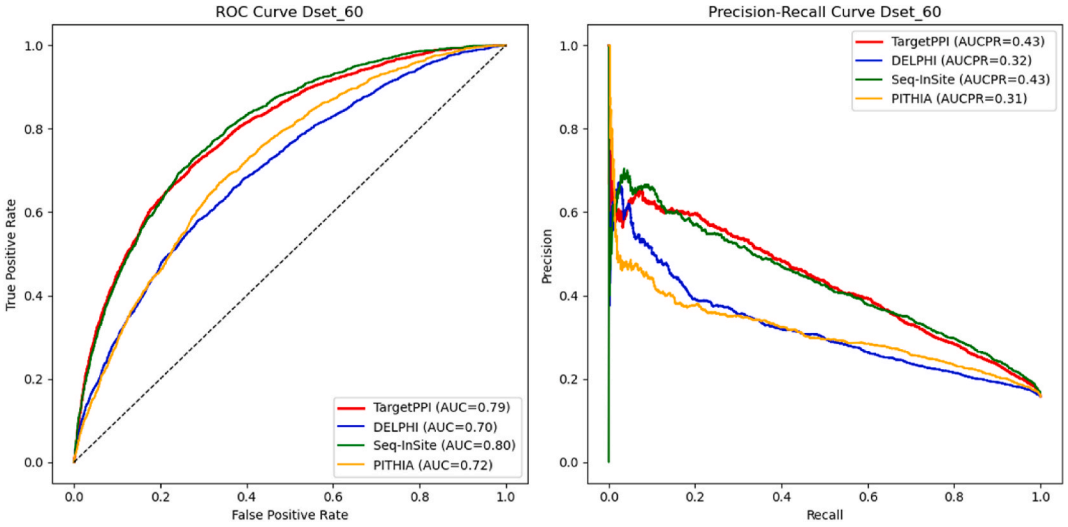

Fig. 7. Comparison of ROC and precision-recall curves between TargetPPI and other sequence-based methods on Dset_60.

For each group, we evaluate the *MCC* on the test set (Dset_448) by varying the number of integrated models from 1 to 23. As illustrated in Fig. 5, for the group with $k_p < 0.7$, the *MCC* reaches its maximum when 9 models are integrated, after which it begins to fluctuate. In contrast, for the group with $k_p > 0.7$, the *MCC* peaks at the integration of 7 models, followed by similar fluctuations. Moreover, when comparing the two groups under the same number of integrated models at their respective optimal points, the $k_p < 0.7$ group consistently yields a higher *MCC* than the $k_p > 0.7$ group. Detailed results are provided in Supplementary Table S7. These findings indicate that integrating models with low similarity contributes to enhanced prediction performance in TargetPPI. Based on this analysis, the optimal number of models for integration in this study is determined to be nine.

3.5. Comparison with sequence-based models

In this section, we evaluate the performance of TargetPPI against nine sequence-based models, including DELPHI [6], ProNA2020 [37], EnsemPPIS [17], PIPENN [38], ProB-Site [9], D-PPIsite [5], SENSDeep [11], PITHIA [14], and Seq-Insite [15], across seven test datasets (Dset_448, Dset_335, Dset_164, Dset_186, Dset_72, Dset_70, and Dset_60). Additional comparisons with other methods are provided in Supplementary Text S3 and Table S3. For a more comprehensive evaluation, Dset_448 and Dset_70 are selected as representative datasets for in-depth analysis. The prediction results for the competing models are primarily obtained from the dataset published by Seq-Insite or reported in their respective original studies. On most test datasets, TargetPPI demonstrates superior performance compared to other sequence-based approaches, including the currently leading Seq-Insite method. However, as shown in Fig. 6, on Dset_335, the performance of Seq-InSite (as reported in its original paper) is significantly better than the results from our tested models and its supplementary materials. We contact the authors of Seq-Insite regarding these discrepancies but do not receive a clear response. On Dset_60, as shown in Fig. 7, although Seq-InSite outperforms TargetPPI in certain regions of the ROC curve, TargetPPI achieves a substantially higher area under the *AUPR*. Furthermore, we conduct a comparative analysis of residue-type-level prediction performance between TargetPPI and Seq-Insite on the Dset_355, Dset_60, and Dset_70 datasets, systematically evaluating the prediction performance across different residue types. Detailed information can be found in Supplementary Text S4 and Table S4.

As shown in Table 4, TargetPPI outperforms all competing methods on all evaluation metrics for Dset_448. In particular, given the imbalanced distribution of binding and non-binding sites, metrics such as

Table 4
Performance comparison of TargetPPI with sequence-based and structure-based prediction methods.

Dataset	Model	ACC	PRE	SPE	SEN	MCC	F1	AUC	PR
Dset_448									
	PITHIA[a]	0.840	0.408	0.907	0.408	0.315	0.408	0.778	0.387
	PIPENN*	0.860	0.470	0.870	0.470	0.254	0.385	0.729	0.357
	D-PPIsite*	0.859	0.480	0.919	0.480	0.399	0.480	0.824	0.479
	Seq-Insite*	0.874	0.535	0.927	0.535	0.462	0.535	0.859	0.552
	TargetPPI	**0.882**	**0.566**	**0.932**	**0.566**	**0.498**	**0.566**	**0.872**	**0.579**
Dset_335									
	PITHIA[a]	0.793	0.444	0.873	0.444	0.317	0.444	0.768	0.442
	D-PPIsite*	0.871	0.460	0.927	0.460	0.387	0.460	0.822	0.448
	Seq-Insite*	0.841	**0.573**	0.902	**0.573**	0.476	**0.573**	0.853	**0.617**
	Seq-Insite[f]	0.886	0.525	0.935	0.525	0.460	0.525	0.859	0.532
	TargetPPI	**0.893**	0.553	**0.939**	0.553	**0.492**	0.552	**0.871**	0.559
Dset_164									
	DELPHI[a]	0.765	0.274	0.914	0.274	0.189	0.274	0.711	0.237
	D-PPIsite*	0.851	0.299	0.917	0.299	0.216	0.299	0.740	0.254
	PITHIA*	0.815	0.360	0.892	0.360	0.252	0.360	0.731	0.340
	TargetPPI	**0.808**	**0.470**	**0.883**	**0.470**	**0.352**	**0.466**	**0.779**	**0.453**
Dset_186									
	SENSDeep*	0.776	0.355	0.866	0.355	0.223	0.358	0.685	0.338
	PITHIA*	0.815	0.360	0.892	0.360	0.252	0.360	0.731	0.340
	D-PPIsite*	0.778	0.386	0.864	0.386	0.250	0.386	0.710	0.364
	TargetPPI	**0.827**	**0.433**	**0.898**	**0.433**	**0.331**	**0.433**	**0.791**	**0.413**
Dset_72									
	DELPHI*	0.847	0.274	0.914	0.274	0.189	0.274	0.711	0.237
	SENSDeep*	0.788	0.258	0.832	**0.448**	0.224	0.327	0.714	0.264
	D-PPIsite*	0.851	0.299	0.917	0.299	0.216	0.299	0.740	0.254
	TargetPPI	**0.866**	**0.370**	**0.925**	0.370	**0.295**	**0.370**	**0.792**	**0.324**
Dset_70									
	ProNA2020*	0.741	0.297	N/A	0.229	0.106	0.258	N/A	N/A
	EnsemPPIS*	0.732	0.375	N/A	0.532	0.277	0.440	0.719	0.405
	EGRET*†	0.715	0.358	N/A	0.591	0.270	0.438	0.719	0.405
	Seq-Insite*	0.781	0.447	0.864	0.447	0.311	0.447	0.766	0.440
	TargetPPI	**0.795**	**0.482**	**0.872**	**0.482**	**0.354**	**0.482**	**0.771**	**0.461**
Dset_60									
	DELPHI[a]	0.792	0.343	0.877	0.343	0.219	0.343	0.699	0.319
	SENSDeep*	0.768	0.344	N/A	0.370	0.199	0.339	0.686	0.371
	ProB-Site*	0.799	0.407	N/A	**0.612**	**0.368**	**0.517**	**0.844**	**0.467**
	Seq-Insite*	0.826	0.448	0.897	0.448	0.345	0.448	0.798	0.430
	MaSIF-site*†	0.780	0.370	N/A	0.561	0.326	0.446	0.775	0.439
	GraphPPIS[a†]	0.816	0.417	0.891	0.417	0.308	0.417	0.776	0.406
	RGN[a†]	0.824	0.443	0.896	0.443	0.338	0.443	0.783	0.427
	TargetPPI	**0.829**	**0.459**	**0.899**	0.459	0.357	0.459	0.793	0.432

The results marked with [a] are obtained from Seq-Insite [15], the methods marked by † are structure-based; the rest are sequence-based. Those marked with * are taken from their respective papers, and the results marked with [f] are derived from the supplementary materials provided by Seq-Insite.

MCC, *F*1, *AUC*, and *AUPR* are more critical. For Dset_448, the *MCC*, F1, *AUC*, and *AUPR* values of TargetPPI are 0.498, 0.566, 0.872, and 0.579, which are 3.6 %, 3.1 %, 1.3 %, and 2.7 % higher than those of the second-best method, Seq-Insite [15]. For Dset_70, the *MCC*, *F*1, *AUC*, and *AUPR* values of TargetPPI are 0.354, 0.482, 0.771, and 0.461, which are 4.4 %, 3.5 %, 0.5 %, and 2.1 % higher than those of Seq-Insite [15]. These results demonstrate that TargetPPI significantly outperforms sequence-based methods in terms of performance.

3.6. Comparison with structure-based models

In this section, we evaluate the performance of TargetPPI in comparison with four structure-based models, namely EGRET [19], GraphPPIS [20], RGN [21], and MaSIF-site [39]. Since structure-based models are highly dependent on protein structural information, their evaluation is typically performed on datasets such as Dset_70 and Dset_60. Therefore, we focus on these two datasets for a detailed comparison, with specific data can be found in Table 4. On Dset_70, TargetPPI demonstrates significant improvements over the top-performing EGRET across multiple key evaluation metrics. Specifically, the *MCC*, F1, *AUC*, and *AUPR* values for TargetPPI are 0.354, 0.482, 0.771, and 0.461, respectively, showing increases of 8.4 %, 4.4 %, 5.2 %, and 5.6 % compared to EGRET. On Dset_60, see Fig. 8, TargetPPI shows significant improvements over the top-performing RGN across multiple key evaluation metrics. Specifically, the *MCC*, *F*1, *AUC*, and *AUPR* values for TargetPPI are 0.357, 0.459, 0.793, and 0.432, respectively, representing increases of 1.9 %, 1.6 %, 1 %, and 0.5 % compared to RGN. The improvements of TargetPPI over EGRET and RGN can be mainly attributed to its incorporation of additional structural features—such as dihedral and angular descriptors, and multiple residue-level distance metrics—which enrich the spatial encoding of residue pairs. Furthermore, the introduction of node similarity into the attention mechanism allows the model to better capture functional correlations between residues, ultimately improving the identification of interaction sites.

3.7. Case study

In this section, we present case studies involving two proteins (Q5SK88 and 1ewyC) from Dset_355 and Dset_60. We evaluate the performance of our method, TargetPPI, in comparison with two

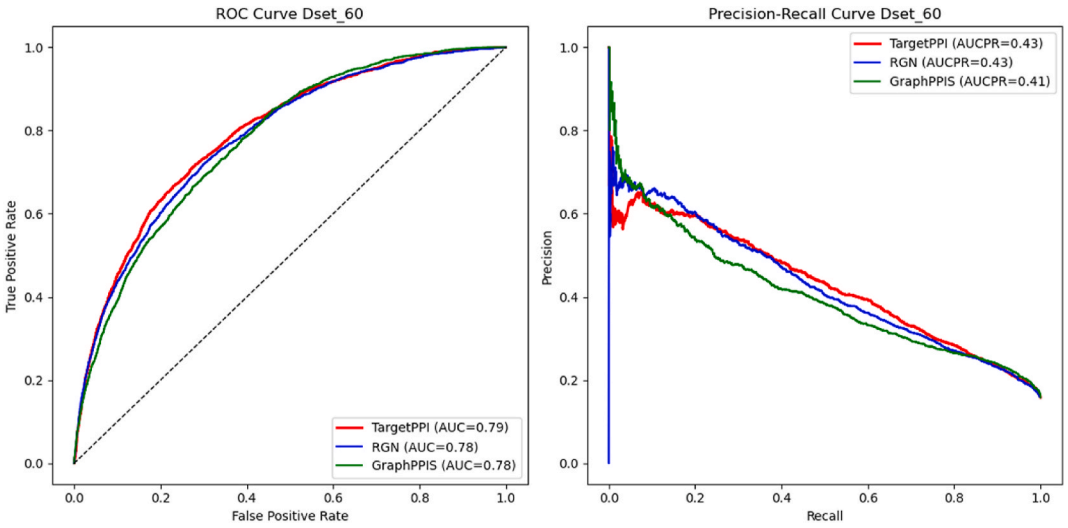

Fig. 8. Comparison of ROC and precision-recall curves between TargetPPI and other structure-based methods on Dset_355.

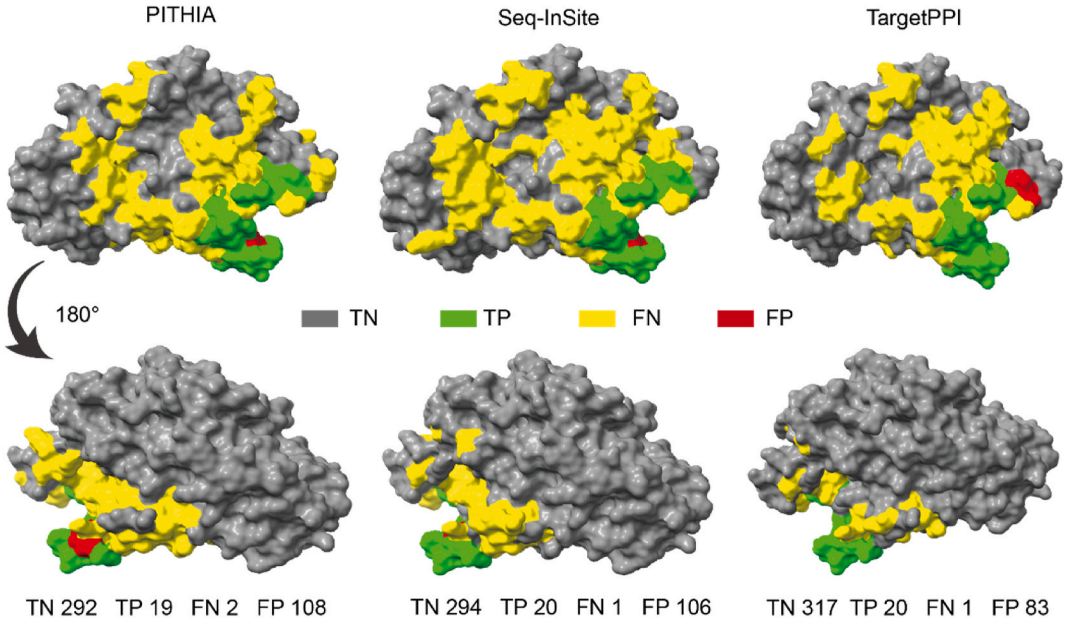

Fig. 9. Visualization of the prediction results for protein Q5SK88 from Dset_355 under PITHIA, Seq-InSite, and TargetPPI.

sequence-based methods, PITHIA [14] and Seq-InSite [15], on Q5SK88, which contains 55 binding sites and 266 non-binding sites, as shown in Fig. 9. Compared to PITHIA and Seq-InSite, the TargetPPI method demonstrates superior performance in predicting false positives, exhibiting fewer false positives. Additionally, TargetPPI shows higher accuracy in predicting true negatives, resulting in improved prediction stability and overall accuracy. We compare the performance of two structural prediction methods, GraphPPIS [20] and RGN [21], with our method of TargetPPI on 1ewyC, which contains 36 binding sites and 193 non-binding sites, as shown in Fig. 10. In these comparisons, TargetPPI outperforms GraphPPIS in terms of prediction accuracy, with higher accuracy in predicting true positives and a significant reduction in false positives. Compared to RGN, TargetPPI achieves better performance in predicting true positives, with fewer false positives. Although TargetPPI slightly underperforms RGN in terms of true negatives and false negatives, its *MCC* score of 0.597 is substantially higher than RGN of 0.433, demonstrating the superior overall performance of TargetPPI in PPIs site prediction.

4. Conclusion

We introduce a new protein-protein interaction site prediction framework, TargetPPI. Experimental results demonstrate that, when compared to widely used structure-based or sequence-based approaches, TargetPPI achieves superior prediction accuracy. Our findings highlight that structural information provides spatial features and constraints not captured by sequence data alone, thereby significantly enhancing the ability of model to predict PPIs sites. Furthermore, while protein sequences may contain rich information, structural data serve as a valuable complement, further refining the prediction of protein properties.

Notably, while the structures predicted by AlphaFold are not flawless, the structural information it provides closely approximates actual protein structures, enabling TargetPPI to effectively utilize this data and improve prediction performance. As protein structure prediction technologies continue to evolve, our study introduces a novel approach for leveraging predicted structural information to enhance PPIs site

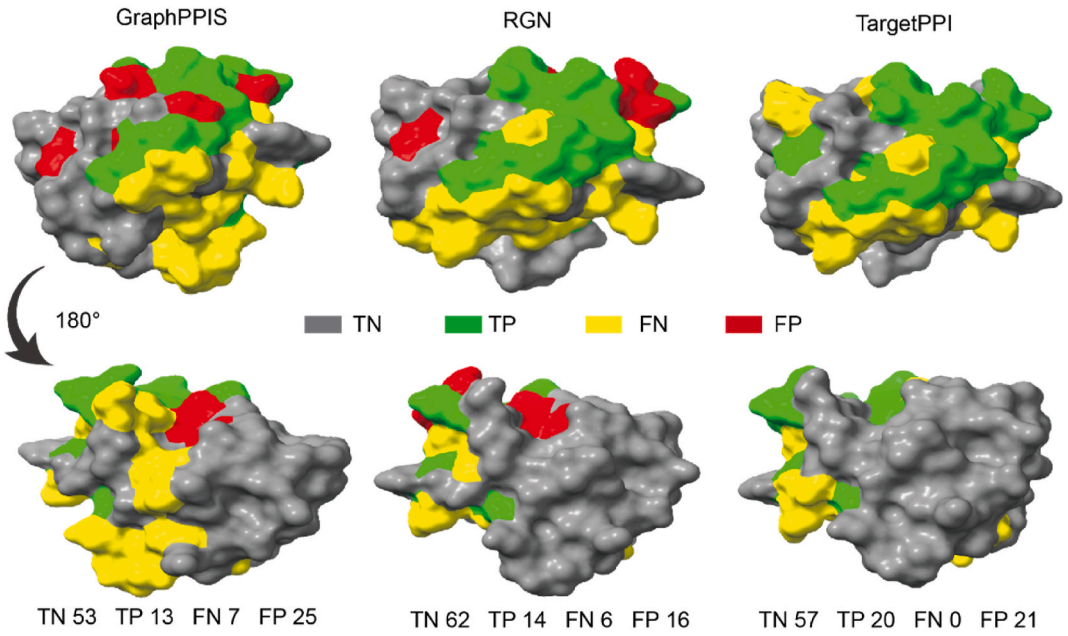

Fig. 10. Visualization of the prediction results for protein 1ewyC from Dset_60 under GraphPPIS, RGN, and TargetPPI.

prediction outcomes. Moving forward, future research could focus on two key areas to further refine PPIs site prediction: (1) incorporating additional structural features into the model to enable deeper exploration of the latent protein space, and (2) enhancing TargetPPI to predict protein-protein binding residue pairs for a given protein pair.

CRediT authorship contribution statement

Qing Zhang: Writing – review & editing, Writing – original draft, Methodology, Data curation. **You-Hang Hu:** Writing – review & editing, Data curation. **Yu Zhou:** Writing – review & editing. **Jun Hu:** Writing – review & editing, Methodology, Funding acquisition. **Xiao-Gen Zhou:** Writing – review & editing, Funding acquisition. **Biao Zhang:** Writing – review & editing, Funding acquisition.

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

This work was supported by the National Natural Science Foundation of China (Nos. 61902352, 62203389, and 62201506), the Natural Science Foundation of Zhejiang Province (No. LY21F020025), the Fundamental Research Funds for the Provincial Universities of Zhejiang (Nos. RF-A20200012 and RF-C2024006), the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (No. 2023R01006), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2025C01121), and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (No. GZC20240159). Zhejiang Key Laboratory of Intelligent Perception and Control for Complex Systems.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ab.2025.115929.

Data availability

Data will be made available on request.

References

[1] S. Jones, J.M. Thornton, Principles of protein-protein interactions, Proc. Natl. Acad. Sci. 93 (1) (1996) 13–20.
[2] J.D. Han, et al., Evidence for dynamically organized modularity in the yeast protein-protein interaction network, Nature 430 (6995) (2004) 88–93.
[3] J. De Las Rivas, C. Fontanillo, Protein-protein interactions essentials: key concepts to building and analyzing interactome networks, PLoS Comput. Biol. 6 (6) (2010) e1000807.
[4] M. Zeng, et al., Protein–protein interaction site prediction through combining local and global features with deep neural networks, Bioinformatics 36 (4) (2020) 1114–1120.
[5] J. Hu, et al., Improving protein-protein interaction site prediction using deep residual neural network, Anal. Biochem. 670 (2023) 115132.
[6] Y. Li, G.B. Golding, L. Ilie, DELPHI: accurate deep ensemble model for protein interaction sites prediction, Bioinformatics 37 (7) (2021) 896–904.
[7] Z.S. Wei, et al., A Cascade random forests algorithm for predicting protein-protein interaction sites, IEEE Trans. NanoBioscience 14 (7) (2015) 746–760.
[8] B. Zhang, et al., Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network, Neurocomputing 357 (2019) 86–100.
[9] S.H. Khan, H. Tayara, K.T. Chong, ProB-Site: Protein binding site prediction using local features, Cells 11 (13) (2022).
[10] Y. Kang, et al., HN-PPISP: a hybrid network based on MLP-mixer for protein-protein interaction site prediction, Briefings Bioinf. 24 (1) (2023).
[11] E. Aybey, Ö. Gümüş, SENSDeep: an ensemble deep learning method for protein-protein interaction sites prediction, Interdiscip Sci 15 (1) (2023) 55–87.
[12] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.
[13] E. Faraggi, Y. Zhou, A. Kloczkowski, Accurate single-sequence prediction of solvent accessible surface area using local and global features, Proteins 82 (11) (2014) 3170–3176.
[14] S. Hosseini, L. Ilie, PITHIA: protein interaction site prediction using multiple sequence alignments and attention, Int. J. Mol. Sci. 23 (21) (2022).
[15] S. Hosseini, G.B. Golding, L. Ilie, Seq-InSite: sequence supersedes structure for protein interaction site prediction, Bioinformatics 40 (1) (2024) btad738.
[16] R M Rao, J Liu, R Verkuil, et al., MSA transformer[C]//International Conference on Machine Learning, PMLR (2021) 8844–8856.
[17] M. Mou, et al., A Transformer-based Ensemble Framework for the Prediction of Protein-Protein Interaction Sites, vol. 6, Research (Wash D C), 2023, p. 240.
[18] A. Elnaggar, et al., ProtTrans: toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2022) 7112–7127.
[19] S. Mahbub, M.S. Bayzid, EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction, Briefings Bioinf. 23 (2) (2022) bbab578.

[20] Q. Yuan, et al., Structure-aware protein–protein interaction site prediction using deep graph convolutional network, Bioinformatics 38 (1) (2021) 125–132.
[21] S. Wang, et al., RGN: residue-based graph attention and convolutional network for protein-protein interaction site prediction, J. Chem. Inf. Model. 62 (23) (2022) 5961–5974.
[22] Z. Du, et al., The trRosetta server for fast and accurate protein structure prediction, Nat. Protoc. 16 (12) (2021) 5634–5651.
[23] X. Zhou, et al., I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction, Nat. Protoc. 17 (10) (2022) 2326–2353.
[24] J. Abramson, et al., Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature 630 (8016) (2024) 493–500.
[25] D. Guo, et al., Graph attention tracking, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[26] J. Jumper, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
[27] Y. Murakami, K. Mizuguchi, Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites, Bioinformatics 26 (15) (2010) 1841–1848.
[28] G Singh, K Dhole, P P Pai, et al., SPRINGS, prediction of protein-protein interaction sites using artificial neural networks[R], Peer J PrePrints (2014), https://doi.org/10.7287/peerj.preprints.266v2.
[29] J. Zhang, L. Kurgan, SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences, Bioinformatics 35 (14) (2019) i343–i353.
[30] S.F. Altschul, et al., Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[31] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions, Nucleic Acids Res. 41 (Database issue) (2013) D1096–D1103.
[32] S. Lu, et al., Attention-based convolutional neural networks for protein-protein interaction site prediction, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021.
[33] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[34] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, Nat. Struct. Biol. 3 (10) (1996) 842–848.
[35] Y. Murakami, K. Mizuguchi, Applying the naïve bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites, Bioinformatics 26 (15) (2010) 1841–1848.
[36] Gong Y Fau - Li, R., et al., A CNN-LSTM Ensemble Model for Predicting Protein-Protein Interaction Binding Sites. (1557-9964 (Electronic)).
[37] J. Qiu, et al., ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence, J. Mol. Biol. 432 (7) (2020) 2428–2443.
[38] B. Stringer, et al., PIPENN: protein interface prediction from sequence with an ensemble of neural nets, Bioinformatics 38 (8) (2022) 2111–2118.
[39] P. Gainza, et al., Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning, Nat. Methods 17 (2) (2020) 184–192.