Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

DSSGNN-PPI: A Protein–Protein Interactions prediction model based on Double Structure and Sequence graph neural networks

Fan Zhang [a,b], Sheng Chang [b], Binjie Wang [a], Xinhong Zhang [c,*]

[a] Huaihe Hospital of Henan University, Kaifeng 475004, China
[b] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China
[c] School of Software, Henan University, Kaifeng, 475004, China

ARTICLE INFO

Dataset link: https://github.com/cstudy1/DSSGNN-PPI

Keywords:
Protein–Protein Interactions
Prediction
Graph neural networks
Dual-level architecture
DSSGNN-PPI

ABSTRACT

The process of experimentally confirming complex interaction networks among proteins is time-consuming and laborious. This study aims to address Protein–Protein Interactions (PPIs) prediction based on graph neural networks (GNN). A novel multilevel prediction model for PPIs named DSSGNN-PPI (Double Structure and Sequence GNN for PPIs) is designed. Initially, a distance graph between amino acid residues is constructed. Subsequently, the distance graph is fed into an underlying graph attention network module. This enables us to efficiently learn vector representations that encode the three-dimensional structure of proteins and simultaneously aggregate key local patterns and overall topological information to obtain graph embedding that adequately represent local and global structural features. In addition, the embedding representations that reflect sequence properties are obtained. Two features are fused to construct high-level protein complex networks, which are fed into the designed gated graph attention network to extract complex topological patterns. By combining heterogeneous multi-source information from downstream structure graph and upstream sequence models, the understanding of PPIs is comprehensively enhanced. A series of evaluation results validate the remarkable effectiveness of DSSGNN-PPI framework in enhancing the prediction of multi-type interactions among proteins. The multilevel representation learning and information fusion strategies provide a new effective solution paradigm for structural biology problems. The source code for DSSGNN-PPI has been hosted on GitHub and is available at https://github.com/cstudy1/DSSGNN-PPI.

1. Introduction

Proteins play an indispensable and crucial role in living organisms, with over 80% of proteins functioning synergistically with others [1]. Signal transduction [2], immune response, cell proliferation, and DNA transcription and replication [3,4] in the life activities of an organism involve proteins extensively. The successful execution of these processes relies on the synergistic collaboration of multiple proteins. This collaborative relationship is termed protein-protein interactions (PPIs). Consequently, delving into PPIs holds profound significance for comprehensively understanding the topological characteristics of biological networks [5], uncovering the evolutionary patterns of PPIs [6], and providing insights for disease diagnosis and drug design [7]. Experimentally, high-throughput technologies such as Yeast Two-Hybrid screening (Y2H) [8], Tandem Affinity Purification (TAP) [9], and Mass Spectrometric Protein Complex Identification (MS-PCI) [10] are frequently employed to identify PPIs. Nevertheless, these experimental methods are costly and time-intensive [11]. Additionally, the detected PPIs may be more susceptible to yielding false positive results [12]. Even if a single experiment detects PPIs, it still may not entirely determine their types [13]. The preceding experimental methods have amassed a substantial volume of PPIs data. This accumulation has facilitated the rapid development of machine learning (ML) based prediction methods for PPIs, resulting in the emergence of numerous models dedicated to predicting PPIs.

Due to the confirmed fact that all protein information is encoded within amino acid sequences, and these sequences are easily obtainable [14], early researchers predominantly employed traditional ML methods for predicting PPIs based on sequence information. Shen et al. pioneered an ML method for predicting PPIs based on protein sequence information [15]. Their approach combined a Support Vector Machine (SVM) with a kernel function and a joint triad feature abstract. In a separate study, Guo et al. proposed an alternative SVM-based method

[*] Corresponding author.
E-mail addresses: zhangfan@henu.edu.cn (F. Zhang), 104754221336@henu.edu.cn (S. Chang), engai01@aliyun.com (B. Wang), zxh@henu.edu.cn (X. Zhang).

https://doi.org/10.1016/j.compbiomed.2024.108669
Received 31 December 2023; Received in revised form 4 April 2024; Accepted 26 May 2024
Available online 29 May 2024

for predicting PPIs, which incorporated Auto-Covariance (AC) [16]. This method was employed to predict PPIs in the yeast dataset. Wong et al. identified room for improvement in the efficiency and accuracy of previous PPIs prediction methods [17]. Therefore, they combined the physicochemical properties of protein sequences with the Random Forest (RF) classifier to predict PPIs. On the other hand, Li et al. constructed a Position-Specific Scoring Matrix (PSSM) based on protein sequence information and integrated it with the RF approach to mine information related to PPIs, specifically for predicting self-interacting proteins [18]. The aforementioned models are all based on protein sequence information and provide solutions for predicting PPIs. However, constrained by the inherent limitations of ML models, they face challenges in precisely extracting PPIs features and performing complex nonlinear transformations.

Compared to traditional ML models, deep learning (DL) has more powerful expressive capabilities and has found extensive applications in the field of bioinformatics, particularly in predicting protein-related structures and functional sites [19]. Du et al. were the first to employ two uncorrelated deep neural networks for processing the sequence information of individual proteins within a protein pair [20]. This model significantly advanced the application of DL in the domain of PPIs prediction. The DNN-PPI proposed by Li et al. utilizes Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) to capture feature information in protein sequences, semantic associations between amino acids, and long-term dependencies, achieving remarkable results across datasets from four different species [21]. Hashemifar et al. employed a deep Siamese-like CNN to predict PPIs [22]. They introduced sliding windows for data augmentation and a random projection module to better capture the symmetry of relationships between protein pairs. The innovative design of this network structure improved the accuracy and generalization ability of PPIs predictions. Subsequently, Chen et al. proposed a Siamese model framework consisting of two residual recurrent convolutional neural network (RCNN) modules, enabling automatic multiscale feature selection for protein sequences. This model demonstrated outstanding performance in both binary and multi-type predictions of PPIs [23].

With the deepening of research in bioinformatics, methods for expressing protein features have continuously advanced. Dutta et al. pioneered the construction of a deep neural network model that integrates heterogeneous information from protein sequences, structures, and genomes [24]. This model extracts features for predicting PPIs, propelling the development in this field. Nambiar et al. introduced PRoBERTa, a Transformer-based pre-trained model with protein sequences as input, utilizing a Transformer to extract the high-level structure of amino acid sequences for protein prediction tasks [25]. Zhang et al. constructed graph embedding vectors for protein biological features, global statistical information of residues, and the relationship graph between proteins and Gene Ontology (GO) functional annotations [26]. They extracted more global features of proteins from this comprehensive approach. The PAthreader model developed by Zhao et al. introduces an innovative three-track alignment algorithm that combines predicted distance profiles, structural profiles, and sequence alignment to efficiently and accurately identify structural and sequence pattern information in proteins [27]. The model extends the model's coverage of protein families and improves the accuracy of template recognition by utilizing the PAcluster80 master structure database, which clusters structures from the PDB and AlphaFold databases.

Yang et al. employed the GNN model S-VGAE [28,29], where they comprehensively utilized information from protein sequences and the topological structure of PPI networks, such as node degrees and neighboring nodes, to predict PPIs through link prediction. However, their approach is limited to binary classification tasks for PPIs. Lv et al. recognized significant shortcomings in existing PPIs prediction evaluation methods, particularly in predicting novel PPIs [30]. In addressing this issue, they not only proposed a new evaluation method but also designed the GNN-PPI model for multi-type PPIs prediction, representing a State-of-the-Art (SOTA) approach at that time. Jha et al. introduced the PPI-GNN model, utilizing GNN to integrate protein structural information and sequence features for predicting PPIs [31]. Specifically, the distance structure graph of two proteins is input into a Siamese-like GNN to infer the internal topology of each protein in an isolated manner, and the graph embedding information of each protein is obtained thus performing PPI binary classification. More recently, Kang et al. introduced the AFTGAN model [32], incorporating an attention-free transformer module to extract protein sequence features and utilizing Graph Attention Networks (GAT) [33] to extract structural features of PPI networks. This model achieved multi-class prediction of PPIs, demonstrating broader application prospects than the PPI-GNN model. While Jha et al. focused on protein structural information, Kang et al. leveraged structural information from PPI networks for predicting PPIs.

Wu et al. skillfully leveraged the multi-scale feature extraction capability in the Inception module to efficiently capture relevant patterns and representations in protein sequences to extract more informative features. They also introduced a feature-relational inference network to enhance the computation of fine similarity between the global features of two proteins, leading to excellent results in the task of predicting PPIs [34]. Kang et al. introduced the Bilateral Branch Learning Network (BBLN) model [35], incorporating cross-modal contrastive learning to capture co-dominant features from GO terms sets and amino acid sequences. Additionally, the model utilizes a multimodal graph learning branch to extract graph-related protein features. The proposed method showcases outstanding generalization capabilities. However, when constructing predictive models, they considered only partial structural information, potentially limiting their ability to comprehensively integrate multidimensional structural knowledge and fully capture the intrinsic mechanisms of PPIs, thereby compromising predictive performance.

Therefore, a predictive model framework Double Structure and Sequence GNN for Protein–Protein Interactions (DSSGNN-PPI) for multi-type prediction of PPIs is proposed. This framework constructs a more comprehensive feature expression space by jointly learning the representations of protein structure graphs and PPI networks. It aims to delve deeper into the intrinsic patterns of PPIs and enhance predictive performance. The main contributions of this paper are summarized as follows:

(i) A new multimodal PPIs prediction model is proposed by using dual-level GNN to model protein residue map and PPI networks respectively;

(ii) The PPI networks are processed by GAT combined with Gate Augmentation mechanism, the global features embedded in each protein residue graph are extracted by GAT, and the sequence embedding features are extracted by ProteinBERT pre-trained model [36] to enrich the node feature information in the upper module of the model.

2. Methods

2.1. Dataset

In this study, multi-type PPIs data from the STRING database were used to train and evaluate the proposed model. STRING is a specialized database system that integrates PPI networks, systematically collecting and integrating both physical and functional associations among proteins [37]. The data come from multiple sources, including co-expression-based interaction predictions, conservative genomic backgrounds, experimental interaction databases, automatic text mining from scientific literature, and known complexes/pathways from curated sources. All these interactions undergo rigorous scoring and evaluation. The STRING database meticulously annotates different types of PPIs, encompassing seven interaction types: catalysis, expression regulation,

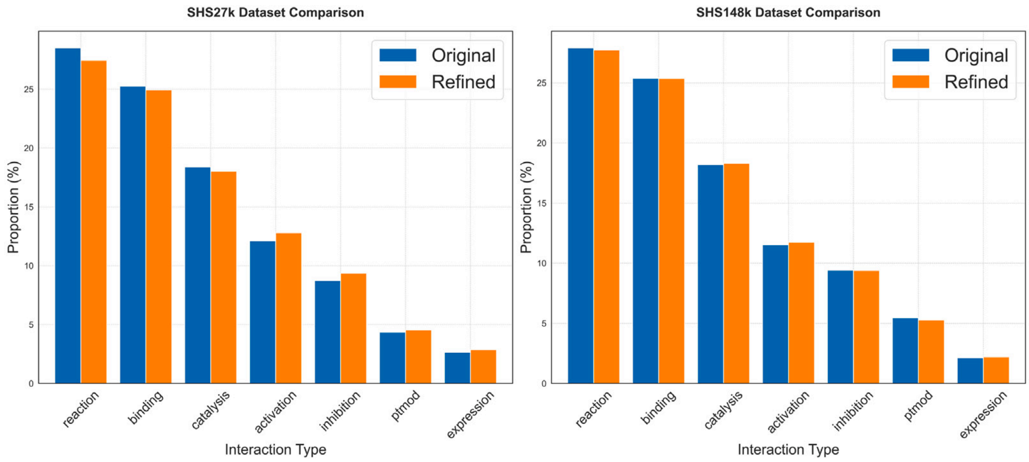

Fig. 1. Comparison of the original and refined datasets.

activation, binding, inhibition, reaction, and post-translational modification. In the investigation, the subsets of SHS27k and SHS148k was used, which were screened by Chen et al. [23] from the STRING Homo sapiens subset with less than 40% protein sequence identity. The methodology for acquiring structural information entailed an initial search in the Protein Data Bank (PDB) [38], subsequently extending to the AlphaFold Database [39] for predictive models when PDB did not yield direct matches. Although the aim is to obtain structural data for all proteins, structural data for some proteins remain inaccessible owing to database limitations or the lack of specific predictive models. These proteins were rigorously documented and excluded from analyses that required structural data, ensuring that study was grounded in proteins with valid structural information and enhancing the credibility of the results.

The dataset refinement had a substantial impact on its composition. The SHS27k dataset, initially containing 1690 proteins, was narrowed down to 1461 proteins after refinement. The SHS148k dataset saw a decrease from 5189 to 4091 proteins. Fig. 1 illustrates a comparative analysis of the protein interaction type proportions between the original and refined datasets. In the original dataset, the sample proportions of the seven interaction types were unbalanced. After extracting the structure, the sample proportion of some interaction types increased compared to the original dataset, especially those with fewer samples. Notably, the refinement process maintained the relative proportions of each interaction type, affirming the preservation of interaction type distribution integrity post-refinement.

In this study, three different dataset partitioning schemes were used: Random sampling, Breadth-First Search (BFS) and Depth-First Search (DFS). 20% of the PPIs data were allocated to the test set, while retaining 80% for the training set. Considering the complex topology of real PPIs, the degree threshold of the root node was set to t=5. Specifically, only nodes with degrees ≤ 5 were chosen as starting points for the partitioning. This setting ensures that the selected root nodes are more marginalized, avoiding the diffusion starting from core hub proteins, which aligns more with the cascading nature of information propagation in biological networks. The degree threshold of the root node is set to 5, mainly based on the following two reasons. Firstly, this threshold helps ensure that the generated subgraphs remain moderate in size and complexity, neither too simple nor too complex. This helps the model learn and generalize patterns of protein interactions efficiently. Secondly, since the subgraph will be used as a test set, limiting the number of connections at the root node helps to simulate the situation of unknown protein interactions in the real world. In the real world, a protein may only interact directly with a limited number of other proteins. In addition, this threshold setting also references the threshold treatment of root nodes in GNN-PPI, AFTGAN, DL-PPI and BBLN models.

This study have conducted experiments with thresholds set to 2, 5, and 7 respectively. The experimental results show that the smaller the degree threshold of the test set root node, the lower the performance of the model. This is because the test set only accounts for 20% of the entire data set, and when the degree threshold of the root node is small, the test set is located at the edge of the entire data set. In contrast, when t is larger, the test set is in the core part of the data set and is in the minority. The model fits the test set more easily on the training set and therefore performs better on the test set. Therefore, t=5 is a reasonable compromise between balancing model performance and data set partitioning.

Lv et al. argued that, compared to simple random sampling, test sets constructed using BFS and DFS algorithms better reflect the predictive performance of models for new PPIs [30]. This is because the latter two algorithms ensure that the test set includes more low-degree proteins located at the network periphery, and predicting interactions involving such proteins is more challenging due to the sparse information. The experimental results confirmed that on test sets generated by BFS/DFS, the performance of various models was significantly lower than their performance on randomly sampled test set.

2.2. Protein sequence feature processing

The protein sequence features utilized in this study include protein sequence information, one-hot encoding of amino acid residues, seven parameter information for amino acids, and seven one-hot encodings based on the similarity derived from the dipolarity and side-chain properties of the 20 amino acids. The efficient sequence processing capabilities of the ProteinBERT pre-trained model were utilized to generate sequence embeddings of proteins. To ensure consistent input dimensions, protein sequences were normalized to 2000 residues. Sequences under 2000 residues were zero-padded, and those over 2000 were truncated. This achieves uniform length without affecting the basic biological details. This method ensures input uniformity and retains crucial protein information. ProteinBERT enhances the typical Transformer/BERT architecture of encoder–decoder frameworks. It was pre-trained on a dataset comprising 106 million protein sequences, including amino acid sequences and associated Gene Ontology (GO) annotations, using self-supervised learning. ProteinBERT has demonstrated predictive performance on multiple evaluation datasets, including protein structure, post-translational modifications, and biophysical properties, showing performance close to or sometimes surpassing the current SOTA methods. Simultaneously, compared to other DL-based protein analysis models, ProteinBERT has a smaller model size and faster computational speed.

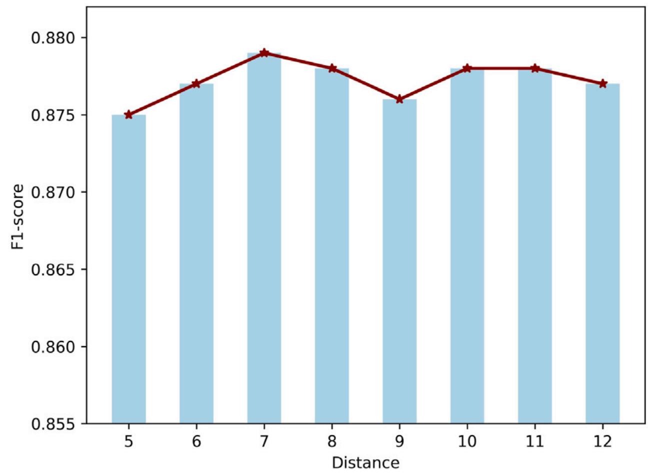

Fig. 2. Sensitivity analysis of residue distance thresholds.

Table 1
Amino acid parameters table.

Name	$\alpha1$	$\alpha2$	υ	π	ρ	φ	ψ
Ala	1.28	0.05	1	0.31	6.11	0.42	0.23
Gly	0	0	0	0	6.07	0.13	0.15
Val	3.67	0.14	3	1.22	6.02	0.27	0.49
Leu	2.59	0.19	4	1.7	6.04	0.39	0.31
Ile	4.19	0.19	4	1.8	6.04	0.3	0.45
Phe	2.94	0.29	5.89	1.79	5.67	0.3	0.38
Tyr	2.94	0.3	6.47	0.96	5.66	0.25	0.41
Trp	3.21	0.41	8.08	2.25	5.94	0.32	0.42
Thr	3.03	0.11	2.6	0.26	5.6	0.21	0.36
Ser	1.31	0.06	1.6	−0.04	5.7	0.2	0.28
Arg	2.34	0.29	6.13	−1.01	10.74	0.36	0.25
Lys	1.89	0.22	4.77	−0.99	9.99	0.32	0.27
His	2.99	0.23	4.66	0.13	7.69	0.27	0.3
Asp	1.6	0.11	2.78	−0.77	2.95	0.25	0.2
Glu	1.56	0.15	3.78	−0.64	3.09	0.42	0.21
Asn	1.6	0.13	2.95	−0.6	6.52	0.21	0.22
Gln	.56	0.18	3.95	−0.22	5.65	0.36	0.25
Met	2.35	0.22	4.43	1.23	5.71	0.38	0.32
Pro	2.67	0	2.72	0.72	6.8	0.13	0.34
Cys	1.77	0.13	2.43	1.54	6.35	0.17	0.41

$\alpha1$: Steric parameter (graph shape index).
$\alpha2$: Polarizability.
υ: Volume (normalized van der Waals volume).
π: Hydrophobicity.
ρ: Isoelectric point.
φ: Helix probability.
ψ: Sheet probability.

Table 2
Amino acid classification information table.

Class	Amino Acids
C1	Ala, Gly, Val
C2	Ile, Leu, Phe, Pro
C3	Tyr, Met, Thr, Ser
C4	His, Asn, Gln, Tpr
C5	Arg, Lys
C6	Asp, Glu
C7	Cys

2.3. Constructing graphs

In DSSGNN-PPI, the protein graph is defined as $G = (V, D, E,$ and $A)$, where V is the set of residue nodes, $M = |V|, (M \le 1000)$. D is the set of distances for the edges. E is the set of undirected edges. If the Euclidean distance between the α-carbon atoms of two residues is less than a threshold distance, there is an edge connecting them. The threshold distance is 7 Angstroms ($\mathring{A}$). After sensitivity analysis, when the distance threshold between residues was set to $7\mathring{A}$, the model performance reached the local optimal value. As shown in Fig. 2. When the threshold is less than or greater than $7\mathring{A}$, the performance of the model will be degraded. This suggests that the $7\mathring{A}$ threshold plays a key role in performance optimization. A is the adjacency matrix, representing connectivity using numerical values. For a graph G with M nodes, the dimensions of the adjacency matrix A are $M \times M$, if the node i is connected to node j, then $A_{ij} > 0$, otherwise, $A_{ij} = 0$. It can be defined as follows:

$$A_{ij} = \begin{cases} 1 & if\ d_{ij} \le 7\mathring{A}\ or\ if\ i = j \\ 0 & other \end{cases} \quad (1)$$

where, d_{ij} is the distance between node i and node j.

The proximity of distances between residues in proteins has a certain impact on the expression of protein functions and interaction types. Therefore, the distance between residues feature is taken into account in the protein graph. The distances between residues are computed as a scalar set, denoted as $D = \{d_{ij}\}$, $d \in \mathbb{N}$ and $i, j < N$. Subsequently, the scalar distances were converted to 8-dimensional vector by Gaussian kernel function. The Gaussian kernel function is defined as follows:

$$e_{ij} = \exp\left(-\frac{\left\|d_{ij} - \sigma_k\right\|^2}{2\sigma^2}\right). \quad (2)$$

By assigning various parameter values to σ_k, the scalar distance of the edges can be expanded into a high-dimensional vector, where $\sigma = 0.15$.

The graph of PPI networks is defined as $H = (P, T,$ and $X)$, where P is the set of protein nodes and T is the set of undirected edges. If there is an interrelationship between protein P_n and P_m, then $t_{mn} \in T$. X is the adjacency matrix using numbers to indicate the connectivity of the PPI networks, the elements in the matrix X can be defined as follows:

$$X_{mn} = \begin{cases} 1 & if\ protein\ n\ interacts\ with\ protein\ m \\ 0 & other \end{cases} \quad (3)$$

According to research by Meiler et al. [40], certain physicochemical properties of amino acids, such as steric parameters, polarizability, volume, hydrophobicity, etc., can influence PPIs. Specifically, these properties affect the formation of hydrogen bonds and hydrophobic interactions between protein molecules. According to the research results of Meiler et al. seven physicochemical property descriptors of the constituent amino acids were chosen and represented as part of the nodes within the protein graph (see details in Table 1). Additionally, one-hot coding information was constructed by categorizing 20 amino acids into seven groups based on their polarity and side-chain polarity (see details in Table 2). These categories reflect amino acid properties from different perspectives. Incorporating these diverse amino acid sequence features provides richer node information for GNN models. This aids in more accurately learning to represent and predict PPIs.

Most protein sequences are longer than 1000. In order to construct a protein graph to obtain its feature representation, 500 residue nodes are extracted from the protein head and tail, thus limiting the graph size to $M \le 1000$. The purpose of this is twofold: (1) It can balance model size and model performance. (2) The head and tail residues of proteins contain more information about protein function and are easier to interact with other proteins, so the head and tail protein residues are extracted to form the diagram. When the sequence information of a protein is extracted and processed with ProteinBert, the sequence length of the protein is limited to ≤ 2000 residues.

2.4. GAT module

In the GAT module of DSSGNN-PPI model, the initial input is a set of features for a group of residue nodes, denoted as $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}$, $\vec{h}_i \in \mathbb{R}^F$. A set of distance features between residues, denoted as

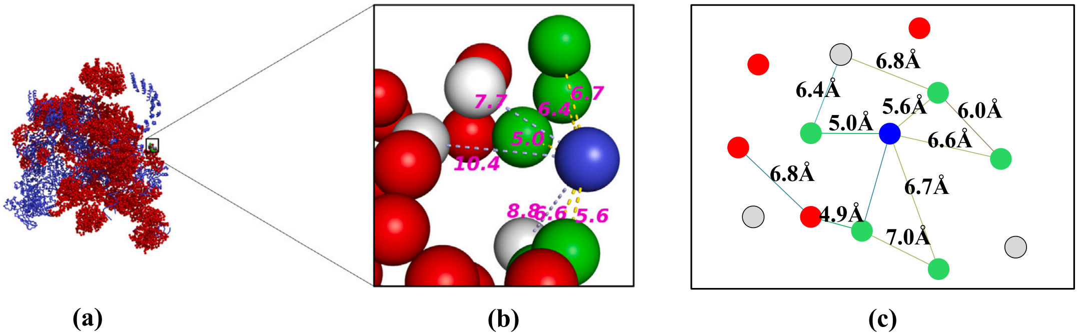

Fig. 3. Flowchart for constructing residue distance graph. (a) Residue representation of protein (PDB: 7DVQ) with spheres for the first and last 500 residues, and simplified sticks for intermediate residues. (b) Demonstration of graph building: the blue node is centered, green nodes within $7\mathring{A}$ are directly connected by edges to the center, while white nodes over $7\mathring{A}$ are excluded. Distances labeled in Angstroms ($\mathring{A}$). (c) Final generated graph with distance attributes on edges.

$\mathbf{e} = \{\vec{e}_{ij}\}$, $\vec{e}_{ij} \in \mathbb{R}^{F_e}$, and $i, j < M$, where M is the number of nodes. $F = 34$ is the number of features in each residue node, and $F_e = 8$ is the number of features in each edge. After passing through the GAT layer, it generates a set of higher-level node features denoted as $\mathbf{h}' = \{\vec{h'}_1, \vec{h'}_2, \dots, \vec{h'}_N\}$, $\vec{h'}_i \in \mathbb{R}^{F'}$, where F' is the latent feature dimension of the GAT layer. The GAT layer transforms the input features into higher-level features through the following operations.

Firstly, the initial features undergo a shared linear transformation parameterized by the weight matrix $\mathbf{W}$, Next, attention coefficients are obtained through element-wise operations on the linearly transformed node features. Specifically, the node features are first scaled and then utilized to calculate attention coefficients via learnable self-attention parameters α, as depicted in Eq. (4):

$$e_{ij} = \alpha\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j, \mathbf{W}_e\vec{e}_{ij}\right), \tag{4}$$

where, $\mathbf{W}_e$ is the linear transformations weight matrix of the edge features. e_{ij} can be interpreted as the contribution of node i to j. Subsequently, the attention coefficients are normalized using the Softmax function:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{\alpha}^{\mathrm{T}}\left[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j \parallel \mathbf{W}_e\vec{e}_{ij}\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\boldsymbol{\alpha}^{\mathrm{T}}\left[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_k \parallel \mathbf{W}_e\vec{e}_{ik}\right]\right)\right)}, \tag{5}$$

where, $\cdot^T$ represents transposition. $\parallel$ is the concatenation operation. LeakyReLU is a variant of ReLU. $\mathcal{N}_i$ is some neighborhood of node i in the graph, and α is the normalized attention coefficient.

Subsequently, the normalized attention coefficients are utilized to compute the feature output for each node. There are two approaches to calculating the node output. One approach is using K separate attention heads to conduct transformations, followed by concatenating their output features (Eq. (6)). This multi-head attention mechanism is appropriate for the hidden layer, as it generates a concatenated hidden representation from the K distinct attentional perspectives.

$$\vec{h}'_i = \parallel_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{k}\mathbf{W}^k\vec{h}_j\right), \tag{6}$$

where, $\parallel$ is the concatenation operation. An alternative approach is to average the feature outputs derived from the K attention heads to acquire the final output. This averaging mechanism is more appropriate for the last layer (i.e., output layer), as it aggregates the attentional information before making predictions, As shown in Eq. (7):

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}i}\alpha_{ij}^{k}\mathbf{W}^k\vec{h}_j\right). \tag{7}$$

2.5. GAT-gate module

The input to this module is the concatenation of the output from the lower module and the output from ProteinBert. Compared to standard GAT, a Gate Augmentation mechanism in the GAT layers is introduced, denoted as GAT-Gate, to improve overall model performance. Building upon Eq. (8) and Eq. (9), the following gating components are incorporated:

$$h_i^{out} = z_i h_i + \left(1 - z_i\right) h'_i, \tag{8}$$

$$z_i = \sigma\left(\mathbf{U}\left(h_i \parallel h'_i\right) + b\right), \tag{9}$$

where, σ denotes a sigmoid activation function. $\mathbf{U} \in \mathbb{R}^{2F\times 1}$ is a learnable vector and b is a learnable scalar value. z_i can be interpreted as how much information about the characteristics of the input nodes is allowed to be passed directly to the next layer.

After propagation through the Dual-Level GNN module, each protein acquires a fixed-length vector representation. Formally, the set of protein vector representations can be denoted as: $\mathbf{P} = \{\vec{P}_1, \vec{P}_2, \dots, \vec{P}_N\}$. When predicting the PPI X_{ij}, the representations of the proteins $\vec{P}_i$ and $\vec{P}_j$ are first combined by dot product. This combined representation is then fed into a Fully Connected layer (FC) that functions as a multi-label prediction classifier, outputting the predicted interaction labels $\hat{y}_{ij}$. During training on the labeled dataset X_{train}, the loss between the true multi-label interactions Y_{train} and predicted $\hat{y}_{ij}$ is computed using a multi-task binary cross-entropy:

$$\mathcal{L} = \sum_{k=0}^{n}\left(\sum_{x_{ij}\in X_{train}} -y_{ij}^{k}\log\hat{y}_{ij}^{k} - \left(1 - y_{ij}^{k}\right)\log\left(1 - \hat{y}_{ij}^{k}\right)\right). \tag{10}$$

2.6. DSSGNN-PPI

The proposed DSSGNN-PPI model utilizes a dual-level GNN architecture with integrated protein sequence information. The lower module employs GAT layers to capture the structural information of proteins. Specifically, subgraphs are constructed for each protein using the first and last 500 residues, with edges based on the spatial distance between residues. The process of constructing these protein distance graphs is illustrated in Fig. 3. Built upon these graphs, GAT layers can learn topology-aware node embedding by aggregating features from neighboring nodes. Through iterative updates across stacked GAT layers, rich graph vector representations of proteins are obtained. The upper module uses GAT-Gate to process the PPI networks, updating protein node representations for final multi-type prediction. The overall framework is shown in Fig. 4.

2.7. Model settings

The lower module of DSSGNN-PPI model incorporates three stacked GAT layers to aggregate features between central residues and their surrounding neighbors, facilitating the learning of local and global structural representations of proteins. The upper module employs three

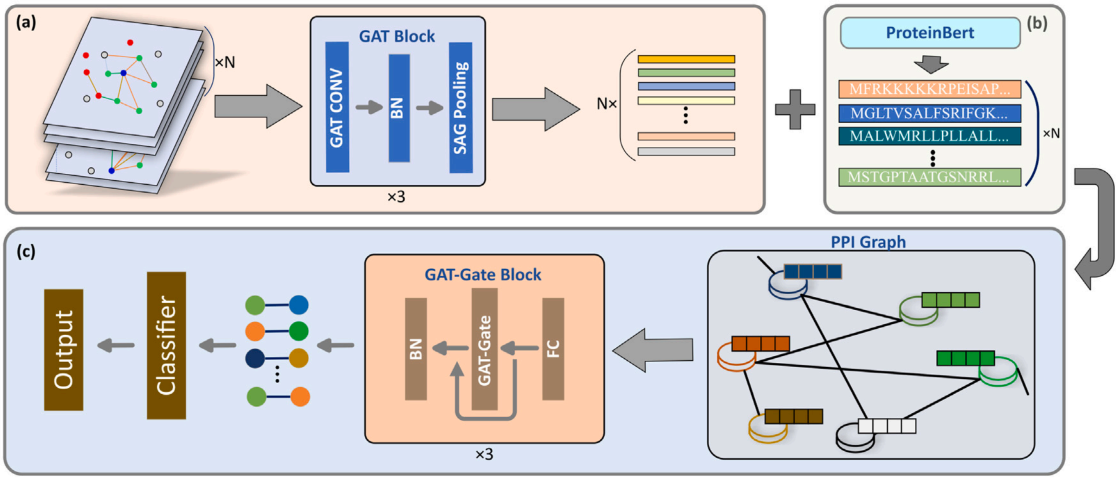

Fig. 4. Framework of DSSGNN-PPI. (a) Protein subgraphs are input into GATs to obtain 256-dim graph embeddings for each protein. (b) Protein sequence information extracted by ProteinBert, concatenated with graph embeddings from (a). (c) Utilizing GAT-Gate module to process the PPI networks and perform multi-class prediction on the resulting feature representation of protein pairs.

Table 3
DSSGNN-PPI experimental results.

Dataset	Partition Scheme	Recall	Precision	Micro-F1
SHS27k	Random	0.863	0.896	0.879
	BFS	0.714	0.739	0.726
	DFS	0.734	0.739	0.753
SHS148k	Random	0.909	0.936	0.921
	BFS	0.719	0.721	0.729
	DFS	0.804	0.839	0.821

successive GAT-Gate layers to capture the higher-order topology of the PPI networks. Between these two GNN modules, the sequence-based protein features derived from ProteinBERT are integrated. The input dimension for the GAT layers is set to 34, with an output dimension of 128 for the GAT-Gate module. Model convergence is expedited through batch normalization, and the ReLU activation function is employed [41]. Optimization is performed using the Adam optimizer [42] with an initial learning rate of 0.001 and weight decay of 5e-4. A dropout of 0.4 is applied to regularize the lower module, with 0.3 for the upper module. The total training process spans 1000 epochs.

3. Results and discussions

3.1. Evaluation criteria

In this study, the predictive performance of the proposed model is evaluated by utilizing evaluation metrics of precision, recall, and micro-F1 score. These statistics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

The micro-F1 score is utilized for multi-label classification scenarios. It is suitable for imbalanced label distributions. Regarding the datasets employed in this study, the different types of PPIs exhibit considerably skewed relative proportions. Therefore, the micro-F1 metric is selected as an appropriate evaluation measure for such data.

$$\text{Precision}_{\text{Micro}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i}, \tag{13}$$

$$\text{Recall}_{\text{Micro}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i}, \tag{14}$$

$$\text{Micro} - \text{F1} = 2\frac{\text{Precision}_{\text{Micro}} \cdot \text{Recall}_{\text{Micro}}}{\text{Precision}_{\text{Micro}} + \text{Recall}_{\text{Micro}}}, \tag{15}$$

where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. i, n denote the ith label type, and there are n labels in total.

In the three partitioning strategy, random sampling, breadth-first search and depth-first search, each has different training sets and test sets. For each data partitioning strategy, experimental results will be repeated under three random seeds to eliminate the influence of randomness. Moreover, the experimental results are all the average values of the three experiments.

3.2. Experimental results

The DSSGNN-PPI model is trained and tested on the SHS27k and SHS148k datasets, with each dataset having three different data partitioning schemes utilized for constructing the test sets. The final experimental results are presented in Table 3.

Observations reveal that the model consistently attains optimal performance metrics on test set constructed through a random sampling scheme, regardless of the dataset utilized (SHS27k or SHS148k). A notable performance gap exists between test sets random sampling and those generated from the BFS and DFS partitions. Despite an increase in data volume, the performance differences among the three data partitions are not sufficiently mitigated. This suggests that test set constructed through random sampling exhibit greater dispersion within the PPI networks, rendering their patterns more predictable. In contrast, test sets constructed via BFS and DFS are positioned at the network's periphery, displaying a tendency to cluster or distribute along a single path. This presents a greater challenge for models to learn both local and global representations simultaneously, imposing more stringent demands on their generalization capabilities.

The three data partitioning schemes have different levels of performance enhancement when the dataset is converted from the smaller SHS27k to the larger SHS148k dataset. Moreover, performance on the DFS test set consistently supersedes that on the BFS test set across dataset variations. This indicates that imputing the latent functions of unknown proteins manifesting in aggregated forms within PPI networks poses greater difficulty for the model. Conversely, proteins dispersed throughout the topological structure stand to benefit from the model's capacity to integrate broader relational contexts, ultimately making more accurate inferences about their properties. This stark difference

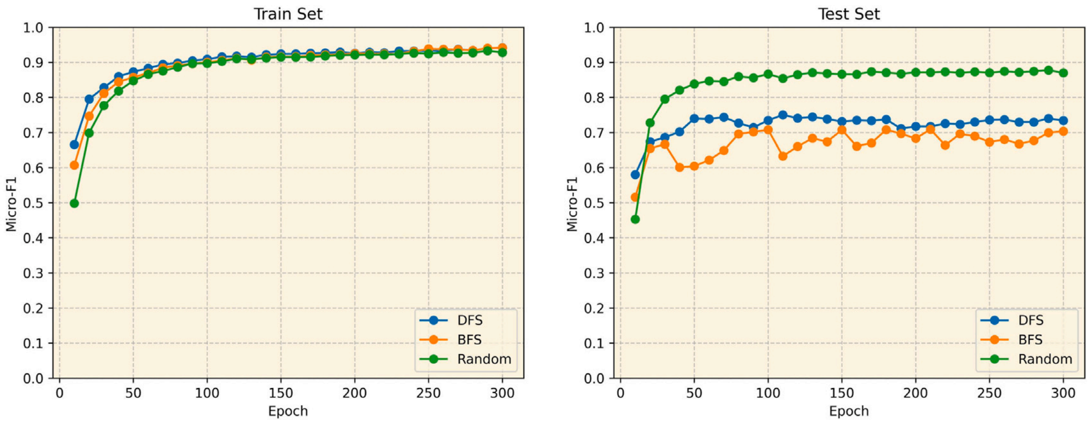

Fig. 5. Variations micro-F1 score over the course of training and testing.

Table 4
Micro-F1 comparison with other methods on SHS27k and SHS148k.

Dataset	Partition Scheme	RF	LR	DPPI	DNN-PPI	PIPR	GNN-PPI	AFTGAN	BBLN	DSSGNN-PPI
SHS27k	Random	0.784	0.717	0.752	0.815	0.838	0.871	0.844	0.865	0.879
	BFS	0.385	0.443	0.475	0.537	0.481	0.656	0.680	0.660	0.726
	DFS	0.352	0.485	0.469	0.486	0.542	0.732	0.739	0.702	0.753
SHS148k	Random	0.822	0.675	0.782	0.875	0.905	0.921	0.920	0.915	0.921
	BFS	0.391	0.481	0.552	0.634	0.657	0.734	0.745	0.657	0.729
	DFS	0.443	0.511	0.537	0.582	0.643	0.793	0.816	0.787	0.821

Table 5
Division results of the test set SHS27k by different partitioning schemes.

Partition scheme	No.	Vision node	Invisible node	X_BS	X_ES	X_NS
Random	1	1375	86	1133	87	2
	2	1378	83	1128	91	3
	3	1373	88	1126	92	4
BFS	1	1411	50	0	1138	112
	2	1350	111	0	1015	212
	3	1377	84	0	1087	171
DFS	1	1350	111	0	1075	150
	2	1382	79	0	1139	156
	3	1382	79	0	1150	141

underscores the notion that the network environs within which unknown proteins are situated exert a substantially greater impact on predictive judgments, as compared to isolated protein node targets.

Fig. 5 presents the evolution of the Micro-F1 scores for both the training and testing datasets throughout the training process. It is observed that, under the BFS partitioning strategy, the model exhibits superior Micro-F1 scores on the training set, indicating a faster convergence rate compared to other partitioning methods. Additionally, a comparison between the Micro-F1 scores variations of models under the Random partitioning strategy for both training and testing sets reveals a high degree of consistency, thereby suggesting a similar data distribution across the two datasets. Furthermore, an examination of the Micro-F1 scores for models under BFS and DFS partitioning strategies on the testing dataset demonstrates that DFS consistently outperforms BFS. This finding further corroborates the inherent challenges associated with predicting interactions of unknown proteins that appear in clusters within the PPI networks.

3.3. Comparison with other methods

In order to demonstrate the performance of the DSSGNN-PPI model, comparisons were made with various representative methods on two datasets, SHS27k and SHS148k. The comparative methods encompass two ML models (namely, RF [17] and LR [43]) along with five popular DL architectures (DNNPPI [21], DPPI [22], PIPR [23], GNN-PPI [30], AFTGAN [32]), and BBLN [33]. Table 4 presents a comparison of Micro-F1 metrics on both datasets under the three data partitioning schemes. DL methods usually have better performance in most cases compared to ML methods. Particularly, on random sampling scheme test sets, most methods exhibit excellent predictive capabilities. However, on test sets partitioned by BFS and DFS, the performance of non-GNN-based methods significantly decreases, except for the four methods leveraging GNN architectures. This suggests that GNN-based models can effectively learn the global topological structure of PPI networks and enhance the model's generalization capacity to localized regions by aggregating representations from neighboring nodes. Therefore, when the test set contains a large proportion of low-degree boundary proteins, GNN models still demonstrate strong robustness, consistently achieving high-level performance. DSSGNN-PPI model consistently achieves superior performance over the three other GNN models (GNN-PPI, AFTGAN, and BBLN), as evidenced by higher Micro-F1 scores on the SHS27k dataset for all partitioning schemes. Notably, DSSGNN-PPI excels under both BFS and DFS data partitioning, underlining its robust predictive ability. This model demonstrates an exceptional capacity to capture predictive patterns effectively, even within smaller datasets. By integrating local and global features end-to-end, DSSGNN-PPI effectively models the complex dependencies in PPI networks, leading to superior predictive performance.

Table 4 demonstrates that DL-based model performance generally scales up with the size of the dataset, transitioning from the smaller SHS27k to the larger SHS148k. This trend supports the idea that a larger volume of data facilitates the learning process for these DL models, particularly in capturing complex protein interaction patterns. Notably, the performance gap between the four GNN models narrows on the large-scale dataset, with all models showing a strong predictive performance in this context.

To visually compare the prediction performance of different models on the SHS27k dataset, the Precision–Recall (PR) curves obtained from the prediction of the proposed DSSGNN-PPI model with BBLN, AFTGAN and GNN-PPI models on the test set are detailed and analyzed in this dataset using both random and DFS data partitioning schemes. In

Table 6
Comparison of the generalization capacity of different interaction types.

Multi Labels	Type ratio (%)	Random partition		BFS partition		DFS Partition	
		GNN-PPI	DSSGNN-PPI	GNN-PPI	DSSGNN-PPI	GNN-PPI	DSSGNN-PPI
Reaction	27.68	0.643	0.716	0.618	0.671	0.638	0.708
Binding	25.28	0.703	0.755	0.695	0.717	0.712	0.741
Ptmod	5.15	0.553	0.635	0.18	0.281	0.288	0.538
Activation	11.91	0.645	0.698	0.49	0.662	0.541	0.692
Inhibition	9.39	0.336	0.563	0.219	0.444	0.144	0.521
Catalysis	18.26	0.67	0.722	0.612	0.699	0.677	0.709
Expression	2.32	0.324	0.337	0.181	0.236	0.235	0.241

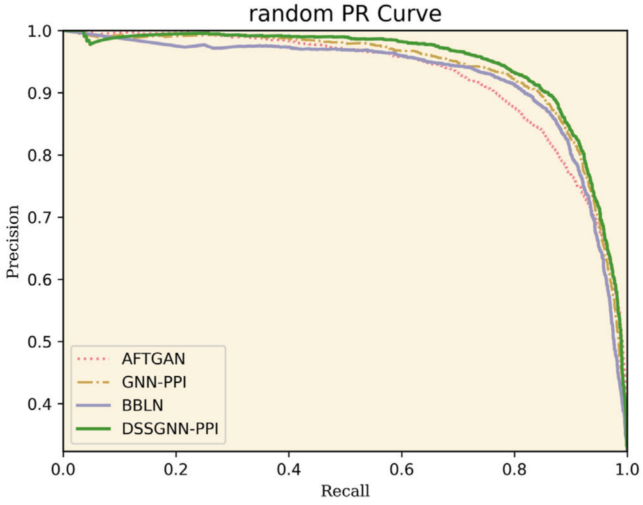

Fig. 6. PR curves for random data partitioning for four models.

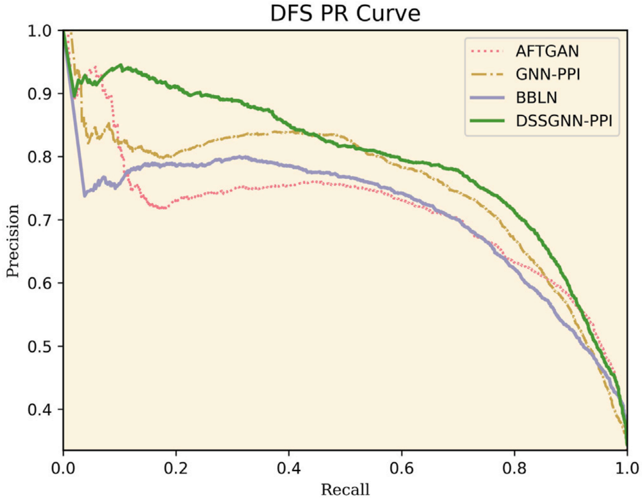

Fig. 7. PR curves under DFS data partitioning for four models.

Figs. 6 and 7, the PR curves of the four models on the test set using random and DFS data partitioning schemes are given, respectively. It can be observed that the proposed DSSGNN-PPI model achieves optimal performance under both random and DFS partitioning schemes, especially maintaining a considerable accuracy advantage over the entire recall range. The GNN-PPI model follows in second place, while AFTGAN lags significantly behind the three approaches. This trend manifests across partitioning strategies and is further evidenced by comparisons of the Area Under the Precision–Recall curve (AUPR) — a consolidated global metric. The DSSGNN-PPI model maintains a stable and statistically significant performance advantage over other methods across different data-splitting strategies.

In addition, PR curves categorized by interaction type are generated so that the model's fitness across subclasses can be visually assessed. As shown in Figs. 8 and 9, all models exhibit optimal performance in predicting the catalysis interaction type, with corresponding precision–recall curves proximal to the ideal maximized corner-simultaneously conferring high precision and recall. This suggests that GNN-based learning methods can efficiently identify binding patterns between enzymatic and substrate complexes. More importantly, the proposed DSSGNN-PPI method not only excels at catalysis prediction but also consistently displays advantageous or comparable performance across all interaction types. Such comprehensive superiority, both visually and quantitatively, verifies the framework's robustness in predicting diverse interaction categories. In contrast, the PR curves for the other two modeled partial interaction types show greater volatility, indicating less stable performance.

The model's capacity to predict interactions between unknown proteins serves as an important benchmark of practical utility, while also enabling more intuitive assessments of generalization. The test set was further categorized into three key categories: protein pairs where Both constituents were Seen during training (BS), Either of the pair proteins was Seen (ES), and pairs where Neither protein was previously Seen (NS). The objective is to predict interaction relationships amongst proteins across all three test subsets. Table 5 describes the results of the SHS27k partitioning test set. Fig. 10 illustrates the Micro-F1 score averages for different models across the three test set categories. It reveals that DSSGNN-PPI and GNN-PPI consistently surpass AFTGAN and BBLN in performance on all test sets. Encouragingly, DSSGNN-PPI model consistently exhibits superior performance in all scenarios, surpassing other GNN-based methods. Particularly noteworthy is its outstanding performance on the NS set, where DSSGNN-PPI model achieves a significantly improved Micro-F1 score of 0.601, notably surpassing the second-best score of 0.512. This strongly validates that the designed framework effectively generalizes to completely unknown PPIs and accurately predicts the potential binding relationships among them. The observed performance differences underscore the robust application prospects of DSSGNN-PPI method in real-world scenarios, especially when facing emerging biological research questions and unknown PPI networks.

3.4. Analysis of model generalization capability

In the prediction of multiple types of PPIs, the model's generalization ability is crucial. If the model can still perform well on unknown PPIs prediction tasks, it holds more value in practical applications. Typically, models are trained and tested on the same dataset, which only allows for evaluating the model's performance on a limited number of unknown proteins. To evaluate the generalization ability of the model under different dataset partitions, the model is trained on the SHS27k dataset and tested on the larger SHS148k dataset. Then, the generalization performance of the proposed DSSGNN-PPI model is compared with the SOTA method GNN-PPI under three data partitioning modes. As Fig. 11 illustrates, the performance of both models

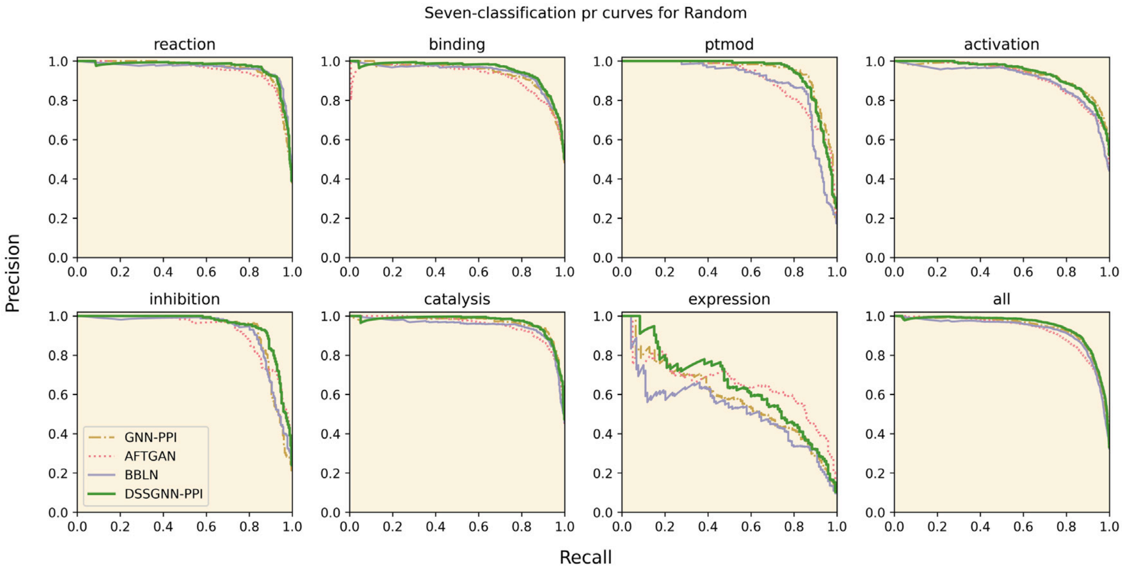

Fig. 8. PR curves for interaction types across four models in random data partitioning.

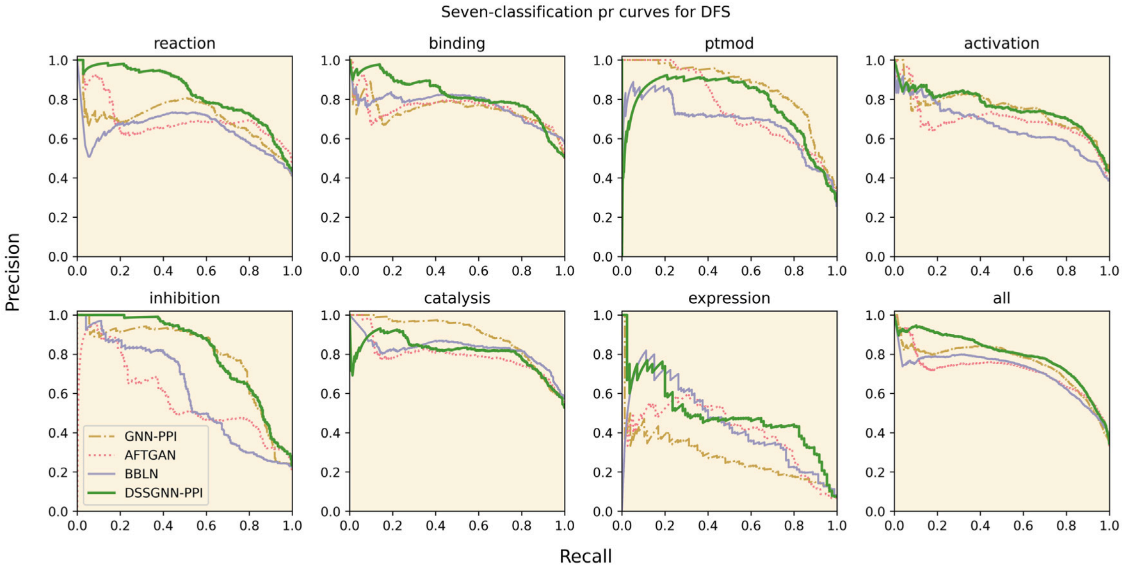

Fig. 9. PR curves for interaction types across four models in DFS data partitioning.

significantly degrades on the unknown test set across the three data partitioning modes, especially under random data partitioning. This outcome demonstrates that the generalization ability of models trained in the random mode is poor, failing to effectively capture the internal structure and patterns of the PPI networks. Upon comparing the two models, DSSGNN-PPI model exhibits better generalization ability, with performance degradation of less than 0.9 under both BFS and DFS data partitioning, while GNN-PPI model suffers more significant performance degradation. The DSSGNN-PPI model's effectiveness in predicting multi-type protein interactions, even when dealing with unknown protein data, renders it a robust and reliable choice for practical applications.

In addition, Table 6 details the results of model performance comparison for different types of PPI prediction tasks on the SHS148k dataset. By analyzing these data in the table, it can be observed that there is a close relationship between the generalization performance of the model and the disease correlation of the interaction types in the dataset. For interaction types with a higher prevalence, such as reaction, binding, and catalysis, both the DSSGNN-PPI and GNN-PPI models exhibit higher generalization abilities across all three data partitioning methods. On the other hand, for the least prevalent ptmod type, the robustness of the DSSGNN-PPI model significantly surpasses that of the GNN-PPI model, maintaining a high-performance level regardless of the data partitioning method. This emphasizes the advantage of DSSGNN-PPI in handling sparsely represented interaction types. In short, the proposed DSSGNN-PPI model shows a substantial performance advantage over the GNN-PPI in predictive metrics across various PPI types, demonstrating superior generalization capability of the DSSGNN-PPI model.

3.5. Visual analysis

To intuitively compare the predictive performance of DSSGNN-PPI and GNN-PPI models, the technique based on unified visualization encoding and labeling is employed. Considering the overall weaker performance of AFTGAN and BBLN in the earlier quantitative analysis, only the models with better performance were selected for comparison. Specifically, proteins are labeled according to seven interaction types

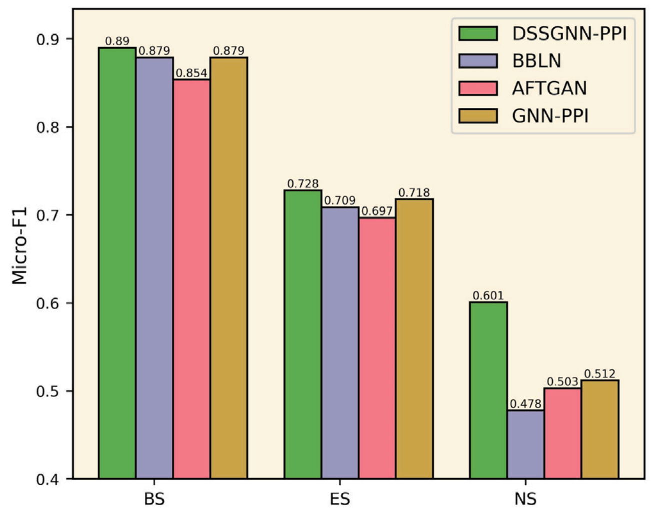

Fig. 10. Micro-F1 comparison of DSSGNN-PPI, BBLN, AFTGAN, and GNN-PPI in BS, ES, and NS subsets.

Table 7
Comparison results of ablation experiments.

Partition scheme	Model	Recall	Precision	Micro-F1
Random	DGAT-PPI	0.861	0.857	0.859
	Top-GAT	0.865	0.893	0.874
	NDGNN-PPI	0.858	0.893	0.875
	NSGNN-PPI	0.850	0.888	0.869
	DSSGNN-PPI	0.863	0.896	0.879
BFS	DGAT-PPI	0.748	0.700	0.723
	Top-GAT	0.570	0.660	0.612
	NDGNN-PPI	0.718	0.716	0.717
	NSGNN-PPI	0.662	0.665	0.664
	DSSGNN-PPI	0.731	0.697	0.726
DFS	DGAT-PPI	0.687	0.760	0.722
	Top-GAT	0.587	0.726	0.650
	NDGNN-PPI	0.730	0.750	0.740
	NSGNN-PPI	0.724	0.625	0.671
	DSSGNN-PPI	0.768	0.738	0.753

and then implemented 2D visualization of high-dimensional protein representations based on UMAP. As shown in Figs. 12 and 13, DSSGNN-PPI model effectively clustered proteins into four major classes, representing four main interaction types (reaction, binding, activation, and expression). It is important to note that, because each protein typically participates in multiple types of interactions, the boundaries between nodes are not strictly segregated. There is a significant overlap between the reaction and binding types, validating the coexistence of these two types in many crucial biological processes, such as enzyme-catalyzed reactions [44], kinase phosphorylation [45], and ligand–receptor activation [46], all of which involve both chemical reactions and molecular binding. Thus, the functional space learned by DSSGNN-PPI model exhibits continuity consistent with the real topology of PPI networks. Comparing the classification performance of the two models, DSSGNN-PPI model roughly divides proteins into four classes, and the boundaries also retain information on the coexistence of various PPI types. In contrast, GNN-PPI struggles to cluster proteins effectively, and the intersections between different types of proteins are more pronounced. This difference indicates that through the deep fusion of sequence embedding and structural information, along with hierarchical and progressive network topology learning, the DSSGNN-PPI model can learn protein functional embedding that are more biologically meaningful. The above comparison, both quantitative and visual, provides a valuable reference case for protein research based on GNN.

3.6. Ablation analysis

To validate the practical effectiveness of the constituent modules in DSSGNN-PPI, module ablation experiments were performed. Distance feature vectors were introduced in constructing the protein graph. When processing the protein graph, GAT was used to aggregate local neighborhood contexts while merging protein sequence information. Processing of PPI networks integrates gate mechanisms.

In order to evaluate the necessity and utility of distance and sequence features as well as the two network modules, the following variant systems were compared:

(i) Removal of the Gate Augmentation mechanism, resulting in a simple model based on the stacking of multiple GAT layers (DGAT-PPI).

(ii) Removal of the lower-layer GAT module, retaining only the upper-layer GAT-Gate (Top-GAT).

(iii) Removal of the residue distance features added during protein graph construction (NDGNN-PPI).

(iv) Removal of protein sequence embeddings added to protein nodes (NSGNN-PPI).

The performance comparison of DSSGNN-PPI with each variant is presented in Table 7. It is evident that utilizing only the top GAT-Gate module (Top-GAT) on the BFS and DFS test sets exhibits a significant performance gap compared to the other two models that incorporate the underlying GNN. This underscores the critical importance of integrating protein structural information to enhance the model's ability to generalize to network boundary regions. Furthermore, when comparing DSSGNN-PPI and DGAT-PPI, the introduction of Gate Augmentation mechanism contributes explicitly to performance improvement. This underscores the capability to enhance the model's proficiency in modeling PPI networks and its adeptness in handling long-range dependencies. Comparing the NDGNN-PPI and other variants of the model, the predictive performance of the model decreased slightly after removing the inter-residue distance feature between amino acids used to construct the protein map. This underscores that the spatial distance information encoded in the residue map better reflects the three-dimensional structure of proteins, contributing to learning more explicit structured representations through GATs. Sequences or conformations alone cannot provide this additional geometrically constrained information. When the sequence embedding information is removed, the performance of the model decreases to varying degrees under the three data partitions. This also proves that using sequence embedding information can effectively enhance the model's understanding and learning of protein characteristics. The sequence embeddings capture local and global pattern information underlying the protein sequences, helping the model to construct more accurate representations of the protein nodes. Thus, the ablation experimental results affirm the effectiveness of the design for cross-level representation learning and information fusion across three levels.

4. Conclusion

A novel dual-level GNN architecture named DSSGNN-PPI is proposed, which is designed for the multiclassification prediction of PPIs. In the bottom layer, residue distance-based amino acid maps and ProteinBERT sequence coding were utilized to efficiently capture protein structure and sequence information, respectively. In the top layer of the PPI networks, a sophisticated GAT-Gate module integrates the underlying protein embedding representations, thereby enhancing the modeling of network topology. The fusion of dual features (sequence and structure) combined with contextual information across different layers empowers the model to gain profound insights into complex PPIs. Finally, the obtained protein node embedding are fed into an FC classifier for multi-label PPIs prediction. After completing the evaluation under multiple datasets and divisions, the results demonstrate the predictive effectiveness, generalization ability, and robustness of the framework. The analysis of ablation experiments also verified the necessity of each component module for the model effect.

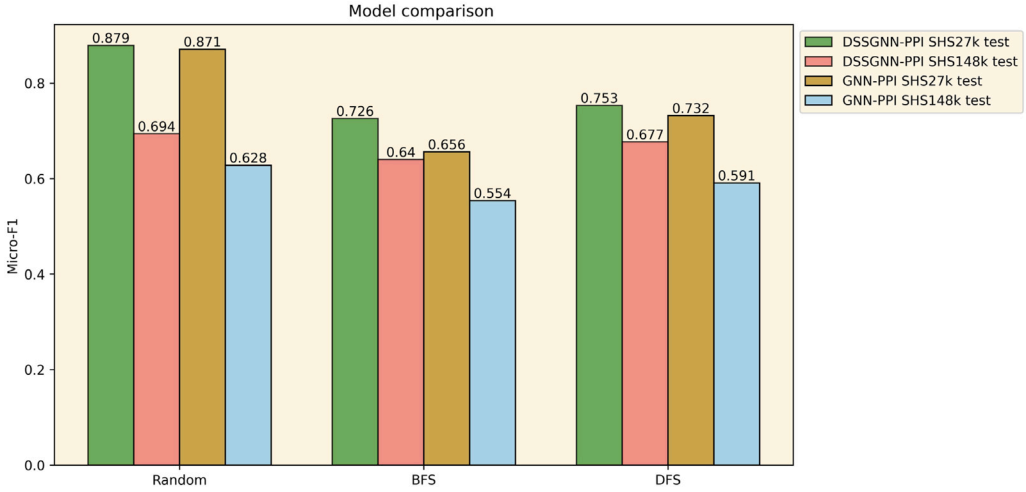

Fig. 11. Generalization evaluation of DSSGNN-PPI and GNN-PPI on training set homologous test set and unknown test set.

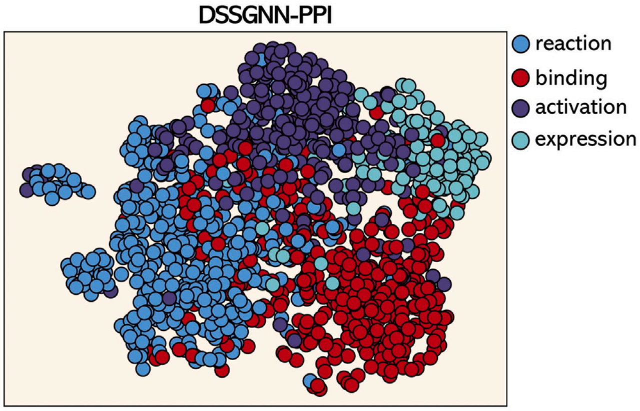

Fig. 12. DSSGNN-PPI Protein embedding visualization with UMAP.

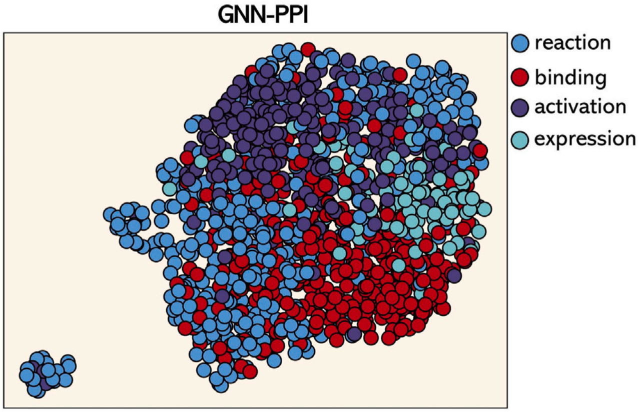

Fig. 13. GNN-PPI Protein embedding visualization with UMAP.

CRediT authorship contribution statement

Fan Zhang: Writing – review & editing, Writing – original draft, Resources, Methodology, Formal analysis, Conceptualization. **Sheng Chang:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Binjie Wang:** Validation, Resources, Formal analysis. **Xinhong Zhang:** Writing – review & editing, Supervision, Methodology, Conceptualization.

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.

Data availability

The source code for DSSGNN-PPI has been hosted on GitHub and is available at https://github.com/cstudy1/DSSGNN-PPI.

Acknowledgments

This research was supported by Key scientific and technological project of Henan Province (No. 232102210003).

References

[1] Tord Berggard, Sara Linse, Peter James, Methods for the detection and analysis of protein–protein interactions, Proteomics 7 (16) (2007) 2833–2842.

[2] Antti Virkamaki, Kohjiro Ueki, C. Ronald Kahn, et al., Protein–protein interaction in insulin signaling and the molecular mechanisms of insulin resistance, J. Clin. Invest. 103 (7) (1999) 931–943.

[3] Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, et al., Structure-based prediction of protein–protein interactions on a genome-wide scale, Nature 490 (7421) (2012) 556–560.

[4] Igor A. Sedov, Yuriy F. Zuev, Recent advances in protein–protein interactions, Int. J. Mol. Sci. 24 (2) (2023) 1282.

[5] Eric Alm, Adam P. Arkin, Biological networks, Curr. Opin. Struct. Biol. 13 (2) (2003) 193–202.

[6] Julian Mintseris, Zhiping Weng, Structure, function, and evolution of transient and obligate protein–protein interactions, Proc. Natl. Acad. Sci. 102 (31) (2005) 10930–10935.

[7] Chandra Sekhar Pedamallu, Janos Posfai, Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information, Source Code Biol. Med. 5 (2010) 1–6.

[8] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, Yoshiyuki Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. 98 (8) (2001) 4569–4574.

[9] Anne-Claude Gavin, Markus Bosche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jorg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (6868) (2002) 141–147.

[10] Stanley Fields, Ok-kyu Song, A novel genetic system to detect protein–protein interactions, Nature 340 (6230) (1989) 245–246.

[11] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, et al., Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry, Nature 415 (6868) (2002) 180–183.

[12] Sean R. Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank C.P. Holstege, Jonathan S. Weissman, Nevan J. Krogan, Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae, Mol. Cell. Proteomics 6 (3) (2007) 439–450.

[13] Javier De Las Rivas, Celia Fontanillo, Protein–protein interactions essentials: key concepts to building and analyzing interactome networks, PLoS Comput. Biol. 6 (6) (2010) e1000807.

[14] Christian B. Anfinsen, The formation and stabilization of protein structure, Biochem. J. 128 (4) (1972) 737.

[15] Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, Hualiang Jiang, Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. 104 (11) (2007) 4337–4341.

[16] Yanzhi Guo, Lezheng Yu, Zhining Wen, Menglong Li, Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic Acids Res. 36 (9) (2008) 3025–3030.

[17] Leon Wong, Zhu-Hong You, Shuai Li, Yu-An Huang, Gang Liu, Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor, in: Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III 11, Springer, Cham, 2015, pp. 713–720.

[18] Jian-Qiang Li, Zhu-Hong You, Xiao Li, Zhong Ming, Xing Chen, PSPEL: in silico prediction of self-interacting proteins from amino acids sequences using ensemble learning, IEEE/ACM Trans. Comput. Biol. Bioinform. 14 (5) (2017) 1165–1172.

[19] Wei Xiong, Hui Liu, Jihong Guan, Shuigeng Zhou, Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks, BMC Bioinform. 14 (12) (2013) 1–13.

[20] Xiuquan Du, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, Yanping Zhang, DeepPPI: boosting prediction of protein–protein interactions with deep neural networks, J. Chem. Inf. Model. 57 (6) (2017) 1499–1510.

[21] Hang Li, Xiu-Jun Gong, Hua Yu, Chang Zhou, Deep neural network based predictions of protein interactions using primary sequences, Molecules 23 (8) (2018) 1923.

[22] Somaye Hashemifar, Behnam Neyshabur, Aly A. Khan, Jinbo Xu, Predicting protein–protein interactions through sequence-based deep learning, Bioinformatics 34 (17) (2018) i802–i810.

[23] Muhao Chen, Chelsea J.-T. Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, Wei Wang, Multifaceted protein–protein interaction prediction based on siamese residual RCNN, Bioinformatics 35 (14) (2019) i305–i314.

[24] Pratik Dutta, Sriparna Saha, Amalgamation of protein sequence, structure and textual information for improving protein-protein interaction identification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 6396–6407.

[25] Ananthan Nambiar, Maeve Heflin, Simon Liu, Sergei Maslov, Mark Hopkins, Anna Ritz, Transforming the language of life: transformer neural networks for protein prediction tasks, in: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Association for Computing Machinery, 2020, pp. 1–8.

[26] Fan Zhang, Yawei Zhang, Xiaoke Zhu, Xiaopan Chen, Fuhao Lu, Xinhong Zhang, DeepSG2PPI: A protein-protein interaction prediction method based on deep learning, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023).

[27] Kailong Zhao, Yuhao Xia, Fujin Zhang, Xiaogen Zhou, Stan Z Li, Guijun Zhang, Protein structure and folding pathway prediction based on remote homologs recognition using PAthreader, Commun. Biol. 6 (1) (2023) 243.

[28] Fang Yang, Kunjie Fan, Dandan Song, Huakang Lin, Graph-based prediction of protein-protein interactions with attributed signed graph embedding, BMC Bioinform. 21 (1) (2020) 1–16.

[29] Thomas N. Kipf, Max Welling, Variational graph auto-encoders, 2016, arXiv preprint arXiv:1611.07308.

[30] Guofeng Lv, Zhiqiang Hu, Yanguang Bi, Shaoting Zhang, Learning unknown from correlations: graph neural network for inter-novel-protein interaction prediction, 2021, arXiv preprint arXiv:2105.06709.

[31] Kanchan Jha, Sriparna Saha, Hiteshi Singh, Prediction of protein–protein interaction using graph neural networks, Sci. Rep. 12 (1) (2022) 8360.

[32] Yanlei Kang, Arne Elofsson, Yunliang Jiang, Weihong Huang, Minzhe Yu, Zhong Li, AFTGAN: prediction of multi-type PPI based on attention free transformer and graph attention network, Bioinformatics 39 (2) (2023) btad052.

[33] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[34] Jiahui Wu, Bo Liu, Jidong Zhang, Zhihan Wang, Jianqiang Li, DL-PPI: a method on prediction of sequenced protein–protein interaction based on deep learning, BMC Bioinform. 24 (1) (2023) 473.

[35] Yan Kang, Xinchao Wang, Cheng Xie, Huadong Zhang, Wentao Xie, BBLN: A bilateral-branch learning network for unknown protein–protein interaction prediction, Comput. Biol. Med. 167 (2023) 107588.

[36] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022) 2102–2110.

[37] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al., The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (D1) (2023) D638–D646.

[38] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.

[39] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.

[40] Jens Meiler, Michael Muller, Anita Zeidler, Felix Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Mol. Model. Annu. 7 (9) (2001) 360–369.

[41] Xavier Glorot, Antoine Bordes, Yoshua Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, PMLR, 2011, pp. 315–323.

[42] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[43] Yael Silberberg, Martin Kupiec, Roded Sharan, A method for predicting protein-protein interaction types, PLoS One 9 (3) (2014) e90904.

[44] Jiali Gao, Shuhua Ma, Dan T. Major, Kwangho Nam, Jingzhi Pu, Donald G. Truhlar, Mechanisms and free energies of enzymatic reactions, Chem. Rev. 106 (8) (2006) 3188–3209.

[45] Tony Pawson, John D. Scott, Protein phosphorylation in signaling–50 years and counting, Trends Biochem. Sci. 30 (6) (2005) 286–290.

[46] Jerome N. Feige, Laurent Gelman, Liliane Michalik, Beatrice Desvergne, Walter Wahli, From molecular action to physiological outputs: peroxisome proliferator-activated receptors are nuclear receptors at the crossroads of key cellular functions, Prog. Lipid Res. 45 (2) (2006) 120–159.

Fan Zhang received the B.S. degree from North China University, China, the M.S. degree from Jiangsu University, China, and the Ph.D. degree in computer application technology from Beijing University of Technology, China. He is currently a professor with the School of Computer and Information Engineering, Henan University, China. He is also work for the Henan Key Laboratory of Big Data Analysis and Processing, the Henan Engineering Laboratory of Spatial Information Processing, and the Huaihe Hospital of Henan University, Kaifeng, China. In addition, he is the vice president of the Graphic and Image Society of Henan Province, the director of the Image Processing and Pattern Recognition Institute of Henan University, the distinguished professor of Huaihe Hospital of Henan University, and the visiting professor of Harvard Medical School, Harvard University, USA. His research focuses on medical image processing, pattern recognition, artificial intelligence, and bioinformatics.

Sheng Chang is currently a graduate student at the School of Computer and Information Engineering, Henan University. His research interests are artificial intelligence and biological information processing.

Binjie Wang is currently a radiologist at Huaihe Hospital of Henan University. His research interests are radiology and medical image processing.

Xinhong Zhang is currently a professor with the School of Software at Henan University, China. She is also work for the Image Processing and Pattern Recognition Institute of Henan University. In addition, she is the director of the Graphic and Image Society of Henan. Her research focuses on digital image processing, pattern recognition, bioinformatics, and artificial intelligence.