OXFORD

Briefings in Bioinformatics, 2023, 1–8
https://doi.org/10.1093/bib/bbad122
Problem Solving Protocol

AGAT-PPIS: a novel protein–protein interaction site predictor based on augmented graph attention network with initial residual and identity mapping

Yuting Zhou, Yongquan Jiang and Yan Yang

Corresponding author. Yongquan Jiang, No. 111, Erhuanlu Beiyiduan, Chengdu City, Sichuan Province, 610031, China. Tel.: 0086-28-66366272; Fax: 0086-28-66367278; E-mail: yqjiang@swjtu.edu.cn

Abstract

Identifying protein–protein interaction (PPI) site is an important step in understanding biological activity, apprehending pathological mechanism and designing novel drugs. Developing reliable computational methods for predicting PPI site as screening tools contributes to reduce lots of time and expensive costs for conventional experiments, but how to improve the accuracy is still challenging. We propose a PPI site predictor, called Augmented Graph Attention Network Protein-Protein Interacting Site (AGAT-PPIS), based on AGAT with initial residual and identity mapping, in which eight AGAT layers are connected to mine node embedding representation deeply. AGAT is our augmented version of graph attention network, with added edge features. Besides, extra node features and edge features are introduced to provide more structural information and increase the translation and rotation invariance of the model. On the benchmark test set, AGAT-PPIS significantly surpasses the state-of-the-art method by 8% in Accuracy, 17.1% in Precision, 11.8% in F1-score, 15.1% in Matthews Correlation Coefficient (MCC), 8.1% in Area Under the Receiver Operating Characteristic curve (AUROC), 14.5% in Area Under the Precision-Recall curve (AUPRC), respectively.

Keywords: protein–protein interaction site prediction, graph neural network, deep learning, initial residual and identity mapping

INTRODUCTION

As the main undertaker of life activities, protein participates in various crucial processes of biological life. However, >80% of the proteins perform functions in the form of complexes instead of acting alone [1]. Therefore, protein–protein interaction (PPI) plays a vital role in many biological processes, such as signal transduction, transport and cellular metabolism [2]. The PPI site refers to the interfacial residues of a protein interacting with another protein. Predicting PPI site contributes to constructing PPI networks [3], understanding the effect of mutation, improving the accuracy of protein–protein docking, predicting protein function [4], comprehending disease mechanism [5], designing novel drugs [6] and developing new therapeutic approaches [7]. Since the laboratory methods for detecting PPI site are expensive and time-consuming [8], such as two-hybrid screening and affinity purification coupled to mass spectrometry [9], it is necessary to develop computational algorithms with high accuracy for predicting PPI site as auxiliary tools of conventional experiments.

The existing methods can be divided into sequence-based methods and structure-based methods. Almost all of these methods are based on machine learning, which includes traditional machine learning algorithms (such as naive Bayesian classifier [10], random forest [11], XGBoost [12], Support Vector Machines (SVM) [13], logistic regression [14]) and deep learning algorithms. Recently, more and more researchers have used deep learning algorithms and achieved good results. For example, Convolutional Neural Network (CNN) was used in ConvsPPIS [15], DELPHI [16], DeepPPISP [17] and the method proposed by Xie *et al.* [18]; Recurrent Neural Network was used in DELPHI [16] and DLPred [19]; MaSIF-site [20] used geometric deep learning to capture surface fingerprints based on protein structures. The protein sequence and structure information play an essential part in the PPI site prediction, especially the protein tertiary structure features. Therefore, one of the crucial challenges is how to embed these important sequence, structure and biochemical physical features into the representation of residues appropriately to improve the accuracy of subsequent prediction. The protein tertiary structure features were rarely considered or underutilized in previous studies, major reasons of which include that the sparse and irregular distribution of residues makes the traditional neural network (such as CNN) unable to capture the structural relationship between residues. At the same time, it is difficult to guarantee the rotation and translation invariance of the protein structure in three-dimensional coordinates using the traditional neural network. To solve the above problems, graph neural network is a suitable choice to construct the model, with representing proteins in the form of graphic data structure. A small number of studies have used graph neural network. For example, EGRET

Yuting Zhou received the BS degree from Southwest Minzu University, Chengdu, China, in 2021. Currently , she is a MS degree candidate in the School of Computing and Artificial Intelligence, Southwest Jiaotong University. Her research focuses on the exploration of artificial intelligence technology in bioinformatics.
Yongquan Jiang is an Associate Researcher of the Artificial Intelligence Research Institute, Southwest Jiaotong University, Chengdu, China. His research interests include artificial intelligence and bioinformatics. He received the Ph.D. degree from Southwest Jiaotong University.
Yan Yang is a Profesor of the School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China. Her research interests include artificial intelligence, big data analysis and mining, ensemble learning, and cloud computing. She received the Ph.D. degree from Southwest Jiaotong University. She is a Member of IEEE and ACM.
Received: December 29, 2022. **Revised:** March 1, 2023. **Accepted:** March 12, 2023

[21] introduced an Edge Aggregated Graph Attention Network and used transfer learning to obtain the feature embedding representation of proteins from protein sequences; GraphPPIS [22] constructed a deep graph convolutional network (GCN) model for PPI site prediction; GraphBind [23] used hierarchical graph neural network to predict the nucleic-acid-binding sites of proteins. At present, the PPI site prediction using graph neural network is still in its early phase, and it has great research potential.

In this study, a new PPI site prediction model based on graph neural network is proposed, called Augmented Graph Attention Network Protein-Protein Interacting Site (AGAT-PPIS). Besides, two kinds of node features and edge features related to protein structure are introduced to provide more abundant learning content. The model uses initial residual and identity mapping structure to connect AGAT layers. In order to integrate more protein structure information, two kinds of protein geometry information are added as edge features in AGAT. Hence the node embedding can be updated by using weighted neighbor node features and neighbor edge features at the same time, which makes the protein sequence and structure information better fused into the residue embeddings. The combination of initial residual and identity mapping structure and AGAT deepens the network, making the model learn the node representation more sufficiently.

MATERIAL AND METHODS

Datasets and evaluation metrics

The datasets used in this experiment were fine-tuned according to the datasets used by GraphPPIS [22], including the training set (Train_335) and test sets (Test_60, Test_315 and UBtest_31). The evaluation metrics are the same with GraphPPIS (See Supplementary for more information). The details of the adjusted datasets used in experiments (Train_335-1, Test_60, Test_315-28, Ubtest_31-6) can be found in Supplementary Table S1 and Table S2.

Protein representation

Each protein was represented as an undirected graph $\mathbf{G} = (V, E, A)$, where $V = \{v_i\}_{i=1,\ldots,N_v}$ denotes the set of feature vectors for N_v nodes, $v_i \in \mathbb{R}^{D_v}$ denotes the feature vector of node i. $E = \{e_{ij} | A_{ij} = 1\}$ denotes the set of feature vectors for N_e edges, and $e_{ij} \in \mathbb{R}^{D_e}$ denotes the edge feature vector between node i and j. A is the adjacency matrix of graph G, with the shape of $N_v \times N_v$. Its specific calculation method will be introduced later. The elements of E are determined according to the adjacency matrix A: if $A_{ij} = 1$ then $e_{ij} \in E$, and if $A_{ij} = 0$ then $e_{ij} \notin E$.

Node features

The residue features were extracted from two aspects, including sequence and structure information of the protein. Specifically, the raw feature representation of each node was concatenated by the following four parts: evolutionary conversation profiles, secondary structure profiles, residue atomic features and residue pseudo-position embedding feature, among which the last two features were newly introduced. Finally, the raw node feature matrix $X^{raw} \in \mathbb{R}^{N_v \times 62}$ for each protein chain that consists of N_v residues was constructed, with 62-dimensional feature vector for each residue.

Evolutionary conversation profiles

The evolutionary conversation features used in this experiment include position-specific scoring matrix (PSSM) and Hidden Markov Models matrix (HMM). PSSM was generated by running the alignment tool PSI-BLAST v2.10.1 [24] with three iterations and the E-value of 0.001. The HMM matrix was generated by running the HHblits v3.0.3 [25] algorithm with default parameters. The shapes of the PSSM and HMM matrix were both $N_v \times 20$. And the values in both were normalized to scores between 0 and 1 by Equation (1), where x represents the original value, and the x_{min} and x_{max} are the minimum and maximum of this feature type in the training set.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Secondary structure profiles

For each protein sequence, the Define Secondary Structure of Proteins (DSSP) matrix that contains secondary structure attributes was calculated by DSSP algorithm tool [26], with the shape of $N_v \times 14$. Each residue in the protein sequence has a corresponding 14-dimensional vector, among which 9-dimensional features are nine secondary structure states expressed in one-hot form; 4-dimensional features are obtained by transforming the torsion angles PHI and PSI of peptide backbone by sine and cosine; one-dimensional feature is converted from solvent accessible surface area to relative solvent accessibility.

Atomic features

According to the protein pdb files from the Protein Data Bank (PDB) website and the inherent properties of the residues, seven features of each atom (excluding hydrogen atom) that makes up the residue were extracted: atomic mass, B-factor, whether it is a residue side-chain atom, electronic charge, the number of hydrogen atoms bonded to it, whether it is in a ring and the van der Waals radius of the atom. Because the number of atoms of different residues is different, the respective average of the seven features of all the atoms that make up the residue were taken as the seven atomic features of the residue, as shown in the Equation (2), where $\{f_{i,j}\}_{i=1,\ldots,7,\ \ j=1,\ldots,N_a}$ denotes the feature i of the atom j in the residue, N_a denotes the number of the residue atoms and $\{x_i\}_{i=1,\ldots,7}$ denotes the atomic feature i of the residue. Ultimately, the atomic feature matrix with the shape of $N_v \times 7$ can be generated for the query protein that contains N_v residues.

$$x_i = \frac{1}{N_a}\left(\sum_{j=1}^{j=N_a} f_{i,j}\right) \tag{2}$$

Pseudo-position embedding feature

The pseudo-position embedding feature of residues contains the relative position information of each residue to the reference residue. In this paper, the coordinate of the residue side-chain centroid (SC) was adopted as the residue pseudo-position. The pseudo-position matrix with the shape of $N_v \times 3$ can be generated for the protein that contains N_v residues, which includes the three-dimensional coordinates of all the residues in the protein. The first residue of the protein sequence was set as the reference residue node o by default, and its pseudo-position is set as the reference position P_o. The calculation method of the pseudo-position embedding PE_i of the node i is shown in Equation (3), where $\left|\overrightarrow{P_oP_i}\right|$ denotes the module of the vector that from P_o to the position P_i of the node i, that is, the Euclidean distance between node o and i. And r is a hyperparameter. Finally, each query protein can obtain a pseudo-position embedding matrix with the shape of $N_v \times 1$.

$$PE_i = \frac{1}{r}\left|\overrightarrow{P_oP_i}\right| \tag{3}$$

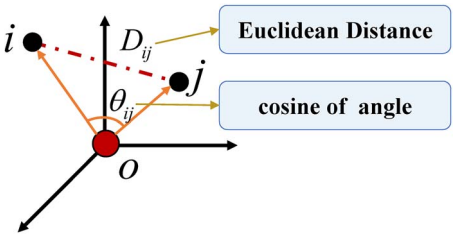

Figure 1. The diagram of the edge features. The node *O* is the reference node. If there is an edge between node *i* and *j*, the edge features will be composed of the Euclidean distance D_{ij} between the two nodes and the cosine of the angle θ_{ij}, where θ_{ij} refers to the angle between the vector from node *O* to *i* and the vector from node *O* to *j*.

Edge features

The adjacency matrix *A* exhibits whether there is an edge between two nodes, and it can be generated by the following two steps. Firstly, the distance matrix *D* of the protein with the shape of $N_v \times N_v$ needs to be calculated. The element D_{ij} in *D* is the Euclidean distance between the pseudo-position P_i of the node *i* and the pseudo-position P_j of the node *j*, that is, the module of the vector $\left|\overrightarrow{P_iP_j}\right|$, as shown in Equation (4).

$$D_{ij} = \left|\overrightarrow{P_iP_j}\right| \tag{4}$$

Then, a distance threshold *t* is selected to transform *D* into *A* with the same shape of $N_v \times N_v$ according to the rules shown in Equation (5). In other words, the elements in *A* will be set to 1 if the distance values of the corresponding positions in *D* are less than or equal to *t*, indicating that there is an edge between the two nodes; and they will be set to 0 if the distance values are over *t*, indicating that there is no edge between the two nodes.

$$A_{ij} = \begin{cases} 1, & if\ D_{ij} \le t \\ 0, & if\ D_{ij} > t \end{cases} \tag{5}$$

The set of the raw edge feature vectors can be expressed as $E^{raw} = \left\{e_{ij}^{raw} \,|A_{ij} = 1\right\}$, where $e_{ij}^{raw} \in \mathbb{R}^2$ denotes the original edge feature vector between node *i* and *j*. e_{ij}^{raw} contains two-dimensional attributes related to geometric knowledge: (i) the Euclidean distance D_{ij} between node *i* and *j*; (ii) the cosine of the angle θ_{ij} between $\left|\overrightarrow{P_oP_i}\right|$ and $\left|\overrightarrow{P_oP_j}\right|$, as shown in Figure 1 and the Equation (6), where · represents the dot product of vectors. The values of the original edge feature vectors e_{ij}^{raw} were also normalized to scores between 0 and 1. As the relative position relationship between nodes can almost be fixed via these two edge features, it increases the translation and rotation invariance of the model.

$$\cos\left(\theta_{ij}\right) = \frac{\overrightarrow{P_oP_i} \cdot \overrightarrow{P_oP_j}}{\left|\overrightarrow{P_oP_i}\right| \left|\overrightarrow{P_oP_j}\right|} \tag{6}$$

The architecture of AGAT-PPIS

Figure 2 shows the overall architecture of AGAT-PPIS. It connects the eight AGAT layers through the initial residual and identity mapping structure, utilizes the node features and edge features that contain protein sequence and structure information to aggregate and update the node embedding representations, and finally outputs the prediction results through a multilayer perceptron (MLP) (see Supplementary for details). The message passing method of graph neural network enables each node to aggregate information from neighboring nodes in the geometric spatial structure that may not participate in effective information fusion because they are far apart in the protein sequence. The initial residual and identity mapping alleviates the over-fitting and over-smoothing as much as possible while deepening the networks. Moreover, because the involved protein geometric structure features are constructed by the relative position between nodes, this model becomes more invariant of translation and rotation.

Augmented graph attention network

The AGAT is a variant network of improving Graph Attention Network (GAT) [27] in this paper. The calculation mechanism of AGAT is shown in Equations (7)–(9). The differences between AGAT and GAT lie in Equations (7) and (9), where edge features are added when calculating attention scores and updating node embeddings, so it is called AGAT. In Equations (7)–(9), $\overrightarrow{h_i}$ and $\overrightarrow{h_j}$ denote the input feature vectors of node *i* and its neighbor node $j \in N_i$ (N_i denotes the set of neighbor nodes of node i); $\overrightarrow{\xi_{ij}}$ denotes the edge feature vector between the node *i* and *j*; $\|$ denotes the concatenating operation; W^{α}, W^{v}, W^{ρ} and W^{ε} represent the learnable parameter matrix of the linear layer at different positions in the network; Ω denotes the Leaky ReLU activation function; γ_{ij} denotes the weight of node *j* that between 0 and 1, obtained by using softmax function to convert its attention score e_{ij}; $\overrightarrow{h_i'}$ denotes the updated embedding of node i; σ denotes the ReLU activation function.

$$e_{ij} = \Omega\left(W^{\alpha}\left[W^{v}\overrightarrow{h_i} \,\middle\|\, W^{v}\overrightarrow{h_j} \,\middle\|\, W^{\rho}\overrightarrow{\xi_{ij}}\right]\right) \tag{7}$$

$$\gamma_{ij} = \mathrm{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N_i}\exp\left(e_{ik}\right)} \tag{8}$$

$$\overrightarrow{h_i'} = \sigma\left(\sum_{j\in N_i}\gamma_{ij}W^{v}\overrightarrow{h_j} + \sum_{j\in N_i}\gamma_{ij}W^{\varepsilon}\left(W^{\rho}\overrightarrow{\xi_{ij}}\right)\right) \tag{9}$$

AGAT with initial residual and identity mapping

In the existing research, Chen *et al.* [28] extended GCN to a deep neural network (GCNII) through the initial residual and identity mapping. Later, GraphPPIS [22] also adopted the GCNII model. However, GCNII cannot distinguish the importance of neighbor nodes and cannot merge more various information about the protein structure flexibly when updating the node embeddings. Hence, the AGAT was combined with the initial residual and identity mapping structure innovatively and successfully in this study, and the calculation mechanisms were modified according to the experimental results and actual requirements. The Equation (11) for our model is obtained by modifying the Equation (10) for GCNII model. In these two equations, σ represents the ReLU activation function, **P** represents the normalized adjacency matrix in the equation of GCN, I_n represents the identity matrix, and $H^{(0)}$ represents the initial node embedding matrix that is not the raw node feature representation matrix *X* but the *X* after linear transformation; $H^{(l)}$ and $H^{(l+1)}$ represent the node feature embeddings before and after the operation of the layer *l* network [refers to GCN in Equation (10) and AGAT in Equation (11)]; $\|$ represents the concatenating operation; $W^{(l)}$ represents the learnable parameter matrix of the linear layer *l*; α and β_l are hyperparameters, where $\beta_l = \log\left(\frac{\lambda}{l} + 1\right)$ is determined by another hyperparameter λ and the layer *l*. Comparing the two equations, it can be found that the adjacency matrix in the Equation (11) have been removed, because the adjacency matrix is not needed after using AGAT instead of GCN. Besides, the latter part of the

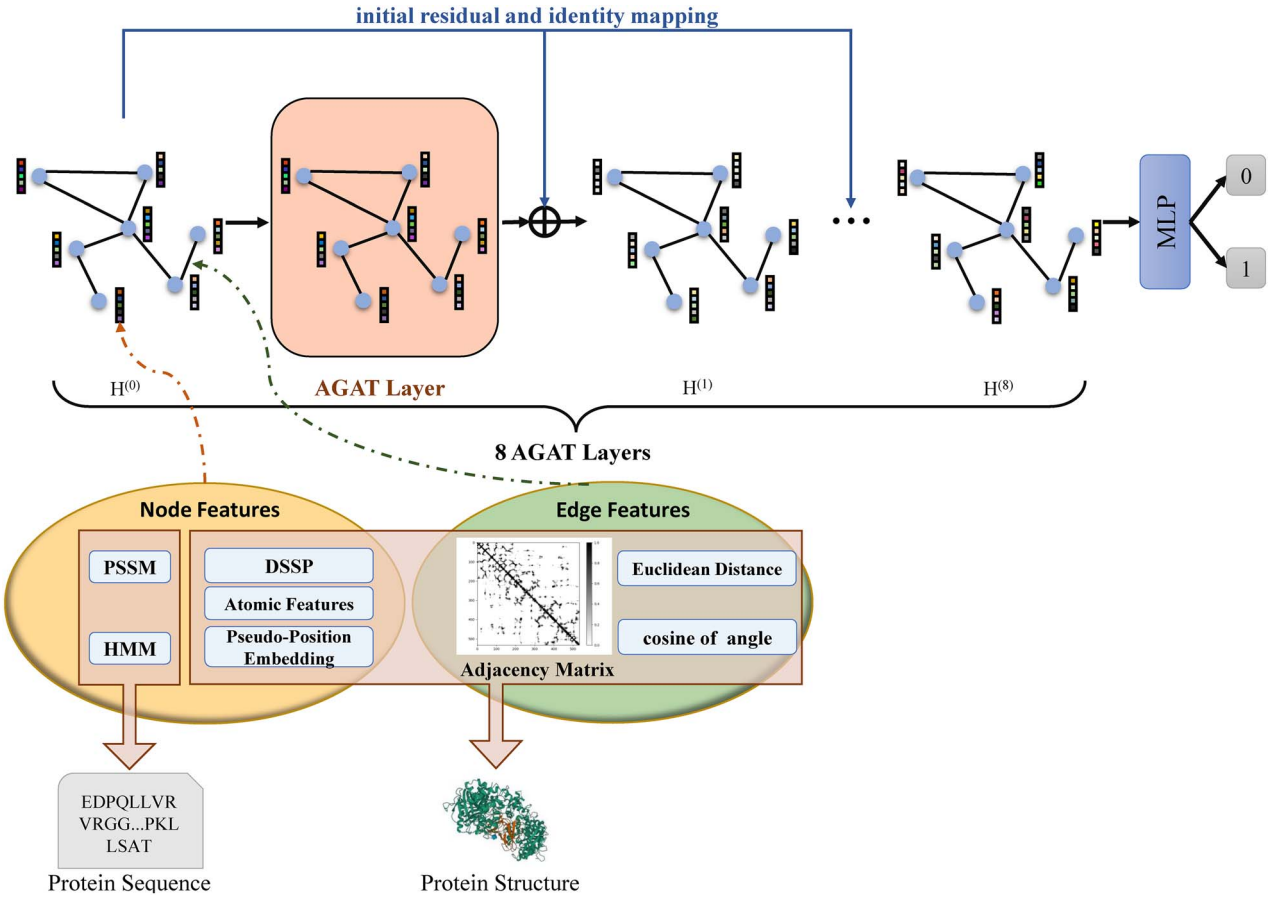

Figure 2. The overall architecture of AGAT-PPIS. The graph is constructed by using node features (including PSSM, HMM, DSSP, Atomic Features, Pseudo-Position Embedding), edge features (including Euclidean Distance, cosine of angle) extracted from protein sequence and structure information' and the adjacency matrix. Then, these features will be inputted into the eight AGAT layers that are connected by the initial residual and identity mapping. The initial node embedding representations $H^{(0)}$ will be updated every time they pass through the AGAT layer. $H^{(1)}$ denotes the node embedding representations after the first AGAT layer. There are eight AGAT layers in all. The eventual node embedding representations $H^{(8)}$ need to go through a MLP to get the prediction result of each residue, in which 0 indicates that the residue is a non-interacting site, and 1 indicates that the residue is an interacting site.

Equation (10) becomes concatenating the input node embedding representations and the initial node embedding representations of the AGAT layer l directly. With the former and latter part of the equation combined, more abundant initial node embedding information can be incorporated in each layer of the model in the two different ways, which further avoids the over-smoothing.

$$H^{(l+1)} = \sigma\left(\left((1-\alpha)PH^{(l)} + \alpha H^{(0)}\right)\left((1-\beta_l)I_n + \beta_l W^{(l)}\right)\right) \quad (10)$$

$$H^{(l+1)} = \sigma\left((1-\beta_l)\left((1-\alpha)H^{(l)} + \alpha H^{(0)}\right) + \left(\beta_l\left(H^{(l)} \| H^{(0)}\right)W^{(l)}\right)\right) \quad (11)$$

RESULTS AND DISCUSSION

Experiment details

In this experiment, 5-fold cross-validation training was conducted on the training set. The average AUROC and AUPRC were taken as the main measurements of the model performance, which were used to select appropriate features and hyperparameters of the model. Finally, the whole training set was also used to train the model. The final hyperparameters of the model were determined according to experience from previous study and some necessary experiments, and they were set as follows. The distance threshold of converting the residues distance matrix into the adjacency matrix was 14 Å; $r = 15$ Å was set to calculate the pseudo-position embedding of residues; the number of AGAT layers was eight; and the length of the node embedding was 256; the parameters involved in the initial residual and identity mapping equations were set to $\alpha = 0.7$, $\lambda = 1.5$; the learning rate of training was 0.001 and the probability of dropout was 0.1. The cross-entropy loss function and Adam optimizer were used to optimize the model and 50 epochs were included in each training. The learning rate decay strategy was used during the training, with which the learning rate would be reduced to 0.6 times and the lowest to 10^{-6} when the AUPRC value does not increase for 10 consecutive epochs. More information about the hyperparameter search of the model can be found in Supplementary Table S4.

Performance comparison with other methods

The performance of AGAT-PPIS was compared with other PPI site predictors on the independent test set Test_60, including PSIVER [10], ProNA2020 [29], SCRIBER [14], DLPred [19], DELPHI [16], DeepPPISP [17], SPPIDER [30], MaSIF-site [20] and GraphPPIS [22]. The top five methods are sequence-based, and the last four methods have used protein structure information. As shown in Table 1, the seven metrics of AGAT-PPIS on the benchmark set Test_60 are significantly better than other methods, with 8% ACC,

Table 1. Performance comparison with other models on Test_60

Method	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
PSIVER	0.561	0.188	0.534	0.278	0.074	0.573	0.190
ProNA2020	0.738	0.275	0.402	0.326	0.176	N/A	N/A
SCRIBER	0.667	0.253	0.568	0.350	0.193	0.665	0.278
DLPred	0.682	0.264	0.565	0.360	0.208	0.677	0.294
DELPHI	0.697	0.276	0.568	0.372	0.225	0.699	0.319
DeepPPISP	0.657	0.243	0.539	0.335	0.167	0.653	0.276
SPPIDER	0.752	0.331	0.557	0.415	0.285	0.755	0.373
MaSIF-site	0.780	0.370	0.561	0.446	0.326	0.775	0.439
GraphPPIS	0.776	0.368	0.584	0.451	0.333	0.786	0.429
AGAT-PPIS	**0.856**	**0.539**	**0.603**	**0.569**	**0.484**	**0.867**	**0.574**

Note: The results of other methods come from the paper GraphPPIS [22] proposed by Yuan Q *et al*. Bold fonts are the best results.

Table 2. Performance comparison of AGAT-PPIS and GraphPPIS on the Test_315-28 and UBtest_31-6

Method	Test_315-28		Btest_31-6		UBtest_31-6	
	MCC	AUPRC	MCC	AUPRC	MCC	AUPRC
GraphPPIS	0.349	0.423	0.352	0.394	0.313	0.339
AGAT-PPIS	**0.481**	**0.572**	**0.485**	**0.583**	**0.327**	**0.365**

Note: The results of GraphPPIS are from running the source code on the independent test sets, and bold fonts are the best results. Test_315-28 comes from the test set specially constructed by Yuan Q *et al*. [22] for further verifying the generalization ability of the model. Proteins in UBtest_31-6 are the monomeric structures corresponding to the 25 protein complexes structures in Test_60, which makes up Btest_31-6.

17.1% Precision, 11.8% F1-score, 15.1% MCC, 8.1% AUROC and 14.5% AUPRC more than the existing best method GraphPPIS.

By continuing to compare the performance on the independent test set Test_315-28 and UBtest_31-6 with GraphPPIS, it turns out that AGAT-PPIS has a significant improvement over GraphPPIS on both datasets, which shows the generalization and robustness of the model further, as shown in Table 2. The PR curves of the two models on four independent test sets are shown in Supplementary Figure S1. Figure 3 shows the results of AGAT-PPIS (**A**) and GraphPPIS (**B**) predicting the interacting sites of the protein 4h3k (PDB ID) chain B from Test_60. This protein has 189 residues, of which 28 are protein–protein interacting sites. AGAT-PPIS predicted 33 interacting sites, of which 27 were true positives and six were false positives, and one interacting site was not predicted. While GraphPPIS predicted 32 interacting sites, of which 19 were true positives, 13 were false positives and nine were unpredicted. By observing the colored parts in Figure 3, it can be found that the prediction results of AGAT-PPIS (**A**) are closer to the actual situation. More protein examples can be found in Supplementary Figure S4–S7 and Table S3.

Ablation experiments

In order to testify the necessity of the components of AGAT-PPIS, the following ablation experiments were implemented by using the average AUROC and AUPRC of 5-fold cross-validation on the training set and the evaluated AUROC and AUPRC on the independent test set Test_60 to compare the model performance.

Feature ablation experiments

As the importance of PSSM, HMM and DSSP matrix has been proved in many previous works, there is no verification here. The importance of two newly added node features was proved by eliminating one of them or both of them from the adopted combined features. As shown in the 'Feature group' section in Table 3, the performance of the model deleted atomic features degrades obviously, which indicates that atomic features play an important role in improving the model performance. While the model performance of removing pseudo-position embedding features degrades subtly. It is probably because the pseudo-position embedding feature has only one dimension, while the dimension of atomic features is seven, so it can provide much less useful information than atomic features. However, removing the pseudo-position embedding feature does cause the performance degradation, so the pseudo-position embedding feature is indeed positive to improving the performance. In addition, although the performance of removing atomic features is very close to that of removing both atomic features and pseudo-position embedding feature on 5-fold cross-validation, the former is better than the latter on Test_60, indicating that pseudo-position embedding feature can increase the robustness of the model. Consequently, the newly added two node features are not redundant.

Model structure ablation experiments

In model structure ablation experiments, three experiments were executed, including using GCN, GAT and AGAT as basic models connected by the initial residual and identity mapping structure. As shown in the 'Base model' section in Table 3, although the GAT's results are better than GCN's on 5-fold cross-validation, the GAT's AUROC on the Test_60 is a little lower than GCN's, while AGAT's results on 5-fold cross-validation and Test_60 are obviously better than the former two, which demonstrates that AGAT can not only improve the model performance but also enhance the robustness and generalization ability of the model. The F1-score, MCC, AUROC, AUPRC metrics on Test_60 and Test_315–28 using three basic models are displayed in Supplementary Figure S2. It is evident that the model with AGAT achieves the best results, which further manifests that using AGAT as the basic model is helpful to improve the overall model performance.

Impact of the residue pseudo-position

For exploring the impact of residue pseudo-position on the model performance, three methods of calculating residue pseudo-position were compared, including the alpha-C atom (CA), the

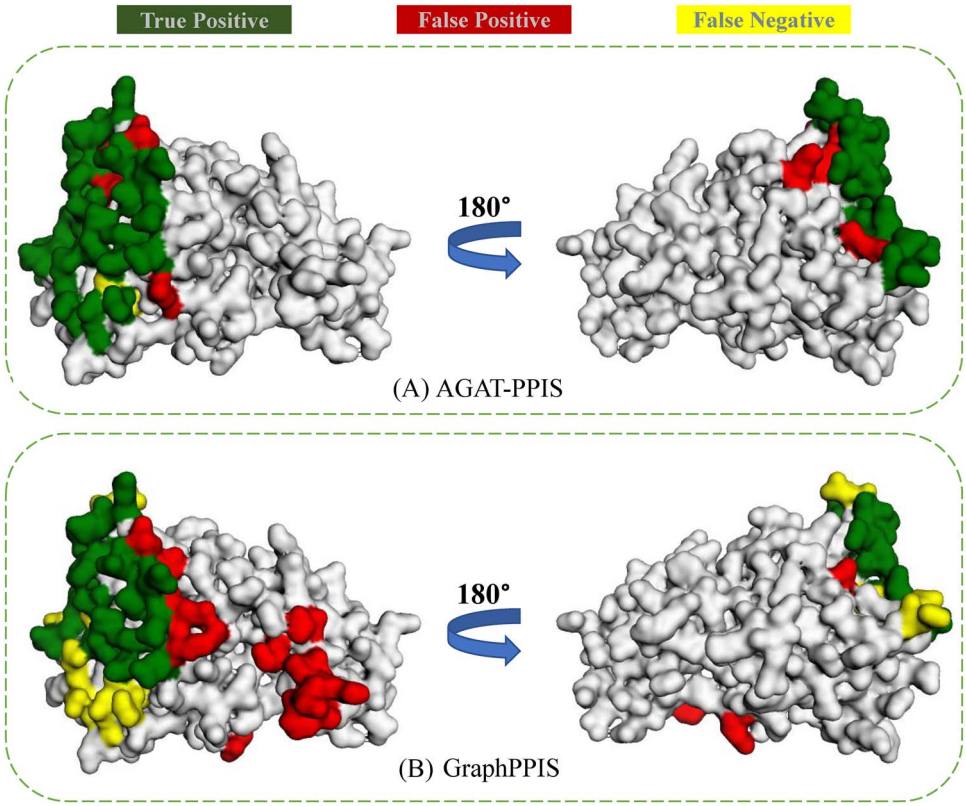

Figure 3. An example of PPI site prediction (PDB ID: 4h3k, chain B) by AGAT-PPIS model (**A**) and GraphPPIS model (**B**). TP, FP and FN are colored in green, red and yellow, respectively. The left and right parts show the views before and after 180° rotation, respectively.

Table 3. The AUROC and AUPRC on 5-fold cross-validation and independent test set Test_60

		CV AUROC	CV AUPRC	Test AUROC	Test AUPRC
Feature group	–Atomic Features	0.796	0.444	0.811	0.467
	–Pseudo-position Embedding	0.853	0.551	0.861	0.567
	–Atomic Features – Pseudo-position Embedding	0.795	0.446	0.796	0.443
	All	**0.854**	**0.558**	**0.867**	**0.574**
Base model	GCN	0.836	0.517	0.848	0.549
	GAT	0.843	0.536	0.843	0.563
	AGAT	**0.854**	**0.558**	**0.867**	**0.574**
The pseudo-position of residue	CA	0.834	0.513	0.843	0.537
	C	0.847	0.538	0.853	0.548
	SC	**0.854**	**0.558**	**0.867**	**0.574**

Note: The Feature group part shows the results of feature ablation experiments, including eliminating atomic features only, eliminating pseudo-position embedding feature only, eliminating both and eliminating neither. The Base model part shows the results of the model structure ablation experiments, including model with GCN, GAT and AGAT as the basic model. The pseudo-position of residue part shows the results of the experiments of evaluating the impact of residue pseudo-position, including using the CA, C and SC as the residue pseudo-position. The bold fonts are the best results.

residue centroid that contains both backbone and side-chain atoms (C) and the residue side-chain centroid (SC) as the residue pseudo-position. As shown in 'The pseudo-position of residue' section in Table 3, it can be found that when taking the residue SC as the residue pseudo-position, the model achieves the best performance, while the other two methods are slightly inferior to it. Supplementary Figure S3 shows the ROC and PR curves of the three pseudo-positions on the Test_60. It may be concluded that the side chains of residues play a more important role in the PPI site prediction.

CONCLUSIONS

In this study, a novel PPI site prediction model called AGAT-PPIS is proposed. It connects eight AGAT layers with the initial residual and identity mapping structure and exhibits better robustness and generalization than the state-of-the-art method on all independent test sets. Definitely, there are still great potentials for further improvement. For example, self-supervised learning strategies can be applied to pre-train a model with large quantities of protein sequences and structures, and then

the model can be fine-tuned to use for the prediction tasks. Because self-supervised learning strategies can learn abundant underly information about protein sequence and structure from mass data, which is very important information for the PPI site prediction. It is also considered to apply this model to the prediction of protein binding sites with other ligands, hoping to summarize a general model.

Key Points

- AGAT-PPIS is a protein–protein interaction site predictor based on deep graph neural networks, which treats protein–protein interaction site prediction as a classification task of graph nodes.
- Two kinds of node features and two kinds of edge features are introduced to improve the performance of the model.
- The base model of AGAT-PPIS is AGAT, which is a variant of GAT, with adding edge features into the process of GAT calculating the attention scores and updating the node embeddings.
- The combination of the initial residual and identity mapping and AGAT makes the model to learn the high-order embedding representations of resides more sufficiently, and so the model performance surpasses others significantly.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

DATA AVAILABILITY

The datasets and the pre-computed features used in this study are available at https://github.com/AILBC/AGAT-PPIS.

ACKNOWLEDGMENTS

The research work was supported by the National Natural Science Foundation of China (No. 61976247) and the Fundamental Research Funds for the Central Universities (No. 2682021ZTPY110).

REFERENCES

1. Berggård T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* 2007;**7**:2833–42.
2. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform* 2017;**19**:821–37.
3. Li X, Li W, Zeng M, *et al.* Network-based methods for predicting essential genes or proteins: a survey. *Brief Bioinform* 2020;**21**:566–83.
4. Orii N, Ganapathiraju MK. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PloS One* 2012;**7**:e49029.
5. Kuzmanov U, Emili A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Med* 2013;**5**:37–12.
6. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 2007;**450**:1001–9.
7. Petta I, Lievens S, Libert C, *et al.* Modulation of protein–protein interactions for the development of novel therapeutics. *Mol Ther* 2016;**24**:707–18.
8. Hamp T, Rost B. More challenges for machine-learning protein interactions. *Bioinformatics* 2015;**31**:1521–5.
9. Wodak SJ, Vlasblom J, Turinsky AL, *et al.* Protein-protein interaction networks: the puzzling riches. *Curr Opin Struct Biol* 2013;**23**:941–53.
10. Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 2010;**26**:1841–8.
11. Northey TC, Barešić A, Martin AC. IntPred: a structure-based predictor of protein–protein interaction sites. *Bioinformatics* 2018;**34**:223–9.
12. Deng A, Zhang H, Wang W, *et al.* Developing computational model to predict protein–protein interaction sites based on the XGBoost algorithm. *Int J Mol Sci* 2020;**21**:2274.
13. Wang B, Mei C, Wang Y, *et al.* Imbalance data processing strategy for protein interaction sites prediction. *Ieee Acm T Comput Bi* 2019;**18**:985–94.
14. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 2019;**35**:i343–53.
15. Zhu H, Du X, Yao Y. ConvsPPIS: identifying protein–protein interaction sites by an ensemble convolutional neural network with feature graph. *Curr Bioinform* 2020;**15**:368–78.
16. Li Y, Golding GB, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* 2021;**37**:896–904.
17. Zeng M, Zhang F, Wu FX, *et al.* Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 2020;**36**:1114–20.
18. Xie Z, Deng X, Shu K. Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. *Int J Mol Sci* 2020;**21**:467.
19. Zhang B, Li J, Quan L, *et al.* Sequence-based prediction of protein–protein interaction sites by simplified long short-term memory network. *Neurocomputing* 2019;**357**:86–100.
20. Gainza P, Sverrisson F, Monti F, *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92.
21. Mahbub S, Bayzid MS. EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Brief Bioinform* 2022;**23**:bbab578.
22. Yuan Q, Chen J, Zhao H, *et al.* Structure-aware protein–protein interaction site prediction using deep graph convolutional network. *Bioinformatics* 2022;**38**:125–32.
23. Xia Y, Xia CQ, Pan X, *et al.* GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res* 2021;**49**:e51–1.
24. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
25. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers Original Res Biomol* 1983;**22**:2577–637.
27. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. *Stat* 2017;**1050**:20.

28. Chen M, Wei Z, Huang Z, *et al*. Simple and deep graph convolutional networks. *Int Conf Mach Learn* 2020;**119**:1725–35.
29. Qiu J, Bernhofer M, Heinzinger M, *et al*. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J Mol Biol* 2020;**432**: 2428–43.
30. Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions. *Proteins Struct Funct Bioinf* 2007;**66**:630–45.