GRIFFITH UNIVERSITY

Structure-aware protein-protein interaction site prediction using deep graph convolutional network

Author

Yuan, Qianmu, Chen, Jianwen, Zhao, Huiying, Zhou, Yaoqi, Yang, Yuedong

Published

2021

Journal Title

Bioinformatics

Version

Accepted Manuscript (AM)

DOI

10.1093/bioinformatics/btab643

Downloaded from

http://hdl.handle.net/10072/408619

Griffith Research Online

https://research-repository.griffith.edu.au

Subject Section

Structure-aware protein-protein interaction site prediction using deep graph convolutional network

Qianmu Yuan[1], Jianwen Chen[1], Huiying Zhao[2], Yaoqi Zhou[3,4,5,*], Yuedong Yang[1,6,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China
[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China
[3]Peking University Shenzhen Graduate School, Shenzhen 518055, China
[4]Shenzhen Bay Laboratory, Shenzhen 518055, China
[5]Institute for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia
[6]Key Laboratory of Machine Intelligence and Advanced Computing of MOE, Sun Yat-sen University, Guangzhou 510000, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein-protein interactions (PPI) play crucial roles in many biological processes, and identifying PPI sites is an important step for mechanistic understanding of diseases and design of novel drugs. Since experimental approaches for PPI site identification are expensive and time-consuming, many computational methods have been developed as screening tools. However, these methods are mostly based on neighbored features in sequence, and thus limited to capture spatial information.

Results: We propose a deep graph-based framework GraphPPIS (deep Graph convolutional network for Protein-Protein Interacting Site prediction) for PPI site prediction, where the PPI site prediction problem was converted into a graph node classification task and solved by deep learning using the initial residual and identity mapping techniques. We showed that a deeper architecture (up to 8 layers) allows significant performance improvement over other sequence-based and structure-based methods by more than 12.5% and 10.5% on AUPRC and MCC, respectively. Further analyses indicated that the predicted interacting sites by GraphPPIS are more spatially clustered and closer to the native ones even when false-positive predictions are made. The results highlight the importance of capturing spatially neighboring residues for interacting site prediction.

Availability: The datasets, the pre-computed features, and the source codes along with the pre-trained models of GraphPPIS are available at https://github.com/biomed-AI/GraphPPIS. The GraphPPIS web server is freely available at https://biomed.nscc-gz.cn/apps/GraphPPIS.

Contact: yangyd25@mail.sysu.edu.cn or zhouyq@szbl.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein interactions (PPI) play crucial roles in many biological processes such as signal transduction, transport and metabolism (Zhang and Kurgan, 2018). Identification of the residues involved in physical contacts in protein-protein complexes (PPI sites) helps construct protein-protein interaction networks (De Las Rivas and Fontanillo, 2012), predict protein functions (Orii and Ganapathiraju, 2012), reveal molecular mechanisms of diseases (Kuzmanov and Emili, 2013), and design novel drugs (Wells and McClendon, 2007). However, conventional experimental methods for PPI site detection such as two-hybrid assay and affinity purification are costly and time-consuming (Shoemaker and Panchenko, 2007). Therefore, it is necessary to develop complementary computational methods capable of making accurate PPI site prediction.

The methods for computational PPI site prediction can be classified into two classes according to the information used for prediction. The first class are sequence-based approaches that require readily available protein

© The Author(s) (2021). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

sequences only. Nonetheless, their predictive accuracies are usually quite limited (Zhang and Kurgan, 2018). By comparison, the second class are structure-based approaches that infer PPI sites from known structures. These methods are often more accurate and practically useful because there are a growing number of structurally resolved proteins with unknown functions (Porollo and Meller, 2007). Moreover, protein-protein complex structures are much more challenging to solve experimentally, and only about 50% structures in Protein Data Bank (PDB) (Berman, et al., 2000) are protein-protein complexes.

The majority of existing PPI site prediction methods are machine-learning-based techniques (Esmaielbeiki, et al., 2015; Zhang and Kurgan, 2018) since sequence conservation (Capra and Singh, 2007) and interaction energy scoring (Liang, et al., 2006) can only capture part of the determinants for protein-protein interactions. In these machine-learning-based methods, protein sequences are represented by feature groups, which generally include evolutionary information (Murakami and Mizuguchi, 2010), secondary structure (Guharoy and Chakrabarti, 2007), and relative solvent accessibility (RSA) (Porollo and Meller, 2007). Additional features were also explored for this problem, such as high-scoring segment pair (HSP) (Li and Ilie, 2017), physical and physicochemical characteristics (Zhang, et al., 2019), one-hot vectors (Zeng, et al., 2020) or embeddings (Li, et al., 2020) of amino acids. With these well-designed features, various machine learning strategies have been employed, including Naïve Bayes classifier in PSIVER (Murakami and Mizuguchi, 2010), random forest in IntPred (Northey, et al., 2017), SVM in (Wang, et al., 2019), XGBoost in IHT-XGB (Deng, et al., 2020) and neural network in SCRIBER (Zhang and Kurgan, 2019) and ProNA2020 (Qiu, et al., 2020). More recent methods utilized deep contextual learning. For example, convolutional neural network (CNN) was used to explore the context of consecutive residues (Xie, et al., 2020; Zhu, et al., 2020). DLPred (Zhang, et al., 2019) developed a simplified long short-term memory (LSTM) network to perform multi-task learning for concurrent prediction of PPI sites and RSA. DELPHI (Li, et al., 2020) introduced three novel features (HSP, position information, and 3-mer embedding) and implemented an ensemble framework with a CNN and a recurrent neural network (RNN) component. DeepPPISP (Zeng, et al., 2020) developed a CNN-based architecture input with the whole protein chain to capture global information. MaSIF-site (Gainza, et al., 2020) used geometric deep learning to capture surface fingerprints that are important for PPI solely based on protein structures. Considering that protein is a folded chain, SPPIDER (Porollo and Meller, 2007) measured impacts from spatially neighboring residues by adopting weighted averages over features of spatially nearest neighbors. Such selection of neighbors, however, is based on a somewhat arbitrary distance cutoff, and the linear weighting can't reflect the complex relations between spatially neighboring residues.

In the last few years, graph convolutional network (GCN) (Kipf and Welling, 2017) and its variants have been successfully applied to a wide range of tasks with graph-structured data, including genomic analysis (Rao, et al., 2021), protein solubility prediction (Chen, et al., 2021), and drug discovery (Song, et al., 2020). Despite their successes, most GCN-based models adopt shallow architectures that are unable to extract information from high-order neighbors. On the other hand, stacking more layers and adding non-linearity may bring about over-smoothing (Li, et al., 2018), that is, the representations of the nodes in GCN tend to converge to a certain value, leading to poor performance in node classification tasks. Recently, it has been proven that the over-smoothing problem can be effectively solved by two simple techniques, initial residual connection and identity mapping (Chen, et al., 2020).

In this study, we applied a deep Graph convolutional network for structure-based prediction of Protein-Protein Interacting Site (GraphPPIS). Specifically, we considered a protein as an undirected graph and PPI site prediction as a graph node classification problem, where we integrated evolutionary and structural information to construct node features and calculated pairwise amino acid distances to construct the adjacency matrix. Afterwards, initial residual and identity mapping were employed for implementing a deep graph convolutional framework to capture information from high-order amino acid neighbors. GraphPPIS was found to outperform other sequence-based and structure-based methods through various evaluations because the spatial information and deep graph convolution technique have prevented spatially isolated false positives. Predicted binding patches, even when falsely predicted, are often near the native PPI sites. To the best of our knowledge, this is the first work that utilized deep graph convolutional network for PPI site prediction, which can be easily extended to structure-based prediction of other functional sites.

2 Methods

Here, the PPI site prediction task is treated as a graph node classification problem, where our goal is to train a deep graph convolutional network that takes inputs of node features and amino acid distance maps and outputs the probability of being a PPI site for each node.

2.1 Datasets

We adopted three publicly available and widely used benchmark datasets from previous studies: Dset_186, Dset_72 (Murakami and Mizuguchi, 2010) and Dset_164 (Dhole, et al., 2014), which were named by the numbers of proteins in the datasets. Concretely, Dset_186 was constructed based on a collection of known protein-protein complexes in PDB, which was then refined by a six-step filtering process including exclusion of structures with more than 30% missing residues, removal of chains with identical UniprotKB/Swiss-Prot accessions, removal of transmembrane proteins, removal of oligomeric structures (higher than dimeric), removal of proteins with buried surface accessibility and interface polarity under a certain threshold, and removal of redundant proteins. Dset_72 was built based on the protein-protein docking benchmark set version 3.0 (Hwang, et al., 2008) and Dset_164 was constructed on newly annotated protein-protein complexes in PDB (Jun. 2010 to Nov. 2013). These two datasets have all been filtered with the same criteria as for Dset_186. In these datasets, a protein-interacting residue was defined as a surface residue (RSA > 5%) that lost more than 1 $Å^2$ absolute solvent accessibility after protein-protein complex formation. To ensure that the training and test sets obey similar distributions in terms of interacting percentages, we integrated the three datasets into a fused dataset. Subsequently, we used BLASTClust (Altschul, et al., 1990) to remove redundant proteins with more than 25% sequence similarities over 90% overlap on either sequence as in Dset_186 and obtained 395 protein chains, from which we randomly selected 335 protein chains for training (Train_335) and used the remaining 60 chains as independent test (Test_60).

To further demonstrate the generalization of our model, we built another independent test set (Test_315) based on newly solved protein complexes in PDB (Jan. 2014 to May 2021). This dataset has been filtered

Table 1. Statistics of the three benchmark datasets along with the training and test sets used in this study. The columns give, in order, the dataset name, the number of interacting and non-interacting residues in each dataset, and the percentage of the interacting residues out of total.

Dataset	Interacting residues	Non-interacting residues	% of interacting residues
Dset_186	5517	30702	15.23
Dset_72	1923	16217	10.60
Dset_164	6096	27585	18.10
Train_335	10374	55992	15.63
Test_60	2075	11069	15.79
Test_315	9355	55976	14.32
UBtest_31	841	5813	12.64

with the same criteria as for Dset_186 and we have further removed redundant protein chains sharing sequence identity > 25% over 30% overlap or E-value < 1e-6 with any sequence in the above three datasets using BLASTClust and DIAMOND (Buchfink, et al., 2015), respectively.

Protein-protein binding often involves conformational changes by induced fit or conformational selection (Hammes, et al., 2009). To evaluate the robustness of GraphPPIS and the impact of conformational changes on method performance, we collected the corresponding unbound structures for the proteins in the independent Test_60. Specifically, 31 of the 60 proteins have known monomeric structures in PDB (**Supplementary Table S2**), which form an additional unbound test set (UBtest_31). Details of the statistics of these datasets are given in **Table 1**.

2.2 Protein representation

In our framework, a protein consisting of n amino acid residues is represented by an undirected graph $G = (V, E)$, with V denoting the amino acids (nodes) and E denoting the contacts of amino acids according to pairwise distances (edges). Thus, the protein graph G can be represented by a node feature matrix $\boldsymbol{X}$ and an adjacency matrix $\boldsymbol{A}$.

2.2.1 Node features

We employed two groups of amino acid features to train our model: evolutionary information (PSSM and HMM) and structural properties (DSSP), which were concatenated and formed the final node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times 54}$ with n denoting the length of a protein sequence.

Evolutionary information. Evolutionarily conserved residues may contain motifs related to important protein properties such as protein binding propensity. Here, we employed the position-specific scoring matrix (PSSM) and hidden Markov models (HMM) profile. Concretely, PSSM was generated by running PSI-BLAST v2.10.1 (Altschul, et al., 1997) to search the query sequence against the UniRef90 database (Suzek, et al., 2007) with three iterations and an E-value of 0.001. The HMM profile was produced by running HHblits v3.0.3 (Remmert, et al., 2012) to align the query sequence against the UniClust30 database (Mirdita, et al., 2017) with default parameters. Each amino acid was encoded into a 20-dimensional vector in PSSM or HMM, and the values were normalized to scores between 0 to 1 using Equation (1), where v is the original feature value, and Min and Max are the smallest and biggest values of this feature type observed in the training set.

$$v_{norm} = \frac{v - Min}{Max - Min} \tag{1}$$

Structural properties. Three types of structural properties were computed by the program DSSP (Kabsch and Sander, 1983): 1) 9-dimensional one-hot secondary structure profile where the first 8 dimensions represent 8 categories of secondary structure states, and the last dimension represents unknown secondary structure. 2) Peptide backbone torsion angles PHI and PSI, which were converted to a 4-dimensional feature vector using sine and cosine transformations. 3) Solvent accessible surface area (ASA), which was then normalized to relative solvent accessibility (RSA) by the maximal possible ASA of the corresponding amino acid type. This 14-dimensional structural feature group is named DSSP hereinafter.

2.2.2 Adjacency matrix

Consistent with the convention of graph node classification problems, we used an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ to represent the edges in a protein graph, which was obtained by two steps: 1) According to the PDB file of a protein, we acquired the coordinate of the Cα atom of each amino acid residue, and then calculated the Euclidean distances between all residue pairs, which formed a distance map. 2) We transformed this protein distance map into an adjacency matrix by converting any value less than or equal to the chosen cutoff to 1 and any value greater than the cutoff to 0. The cutoff was set to 14 Å based on the model's performance on the training data (**Supplementary Table S3**). It is worth noting that under this condition, each node in the protein graph is self-looped.

We also explored a different strategy, which was first introduced in (Chen, et al., 2019), to process protein distance maps into continuous matrices in which values are in the range of 0 to 1. In this scenario, the distance between residue i and j, namely d_{ij}, is normalized to s_{ij} if d_{ij} is less than or equal to the preset cutoff using the following equation:

$$s_{ij} = \frac{2}{1 + \frac{\max(d_0, d_{ij})}{d_0}} \tag{2}$$

where d_0 is set to 4 Å as (Chen, et al., 2019) suggests.

2.3 The architecture of GraphPPIS

Figure 1 shows the overall network architecture of the proposed model GraphPPIS**,** where the *L*-layer GCN with initial residual and identity mapping aggregates node features over spatial neighbors according to the adjacency matrix, and finally a fully connected layer is employed to convert the output of the last graph convolutional layer to PPI site prediction.

2.3.1 Graph Convolutional Network (GCN)

Given a protein with n amino acids, the protein graph is represented by the node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ and the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, with m denoting the feature dimension of each node ($m = 54$ in this work). The adjacency matrix can be normalized to $\boldsymbol{P} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is the diagonal degree matrix of $\boldsymbol{A}$. Consequently, the graph convolutional operation is computed as follows:

$$\boldsymbol{H}^{(l+1)} = \sigma\left(\boldsymbol{P}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{3}$$

where σ denotes the ReLU function, $\boldsymbol{H}^{(l)}$ and $\boldsymbol{H}^{(l+1)}$ denote the hidden states before and after the convolutional operation of the l+1$^{\text{th}}$ layer, and $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d\times d}$ is a trainable weight matrix, with d as the dimension of node representation in the hidden states.

2.3.2 GCN with initial residual and identity mapping

L-layer GCN has been proven to essentially simulate an *L*-order polynomial filter with fixed coefficients, which limits the expressive power of the

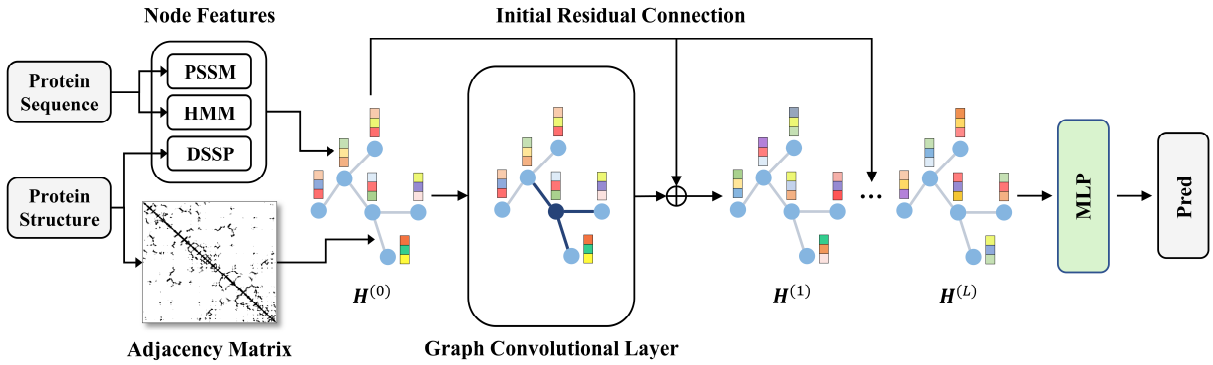

Figure 1. The network architecture of the proposed GraphPPIS model. Extracted from the sequence and structure of a protein, the node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ and the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are used to represent the protein graph, which is input into an L-layer graph convolutional network with initial residual and identity mapping. Here, $\boldsymbol{H}$ denotes the hidden state of the network and L is set to 8 in this work. The output of the L^{th} layer ($\boldsymbol{H}^{(L)}$) is converted to residue-level protein-interacting probabilities by the final MLP module.

model and leads to over-smoothing (Wu, et al., 2019). As a result, GCN achieves its best performance in many graph-based problems via shallow architectures, which are incapable of extracting information from high-order neighbors. Recently, (Chen, et al., 2020) has extended GCN to a truly deep model via initial residual connection and identity mapping:

$$\boldsymbol{H}^{(l+1)} = \sigma\left(\left((1-\alpha)\boldsymbol{P}\boldsymbol{H}^{(l)} + \alpha\boldsymbol{H}^{(0)}\right)\left((1-\beta_l)\boldsymbol{I}_n + \beta_l\boldsymbol{W}^{(l)}\right)\right) \quad (4)$$

where α and β_l are hyperparameters and $\boldsymbol{P}$ is the normalized adjacency matrix. Note that compared with GCN (Equation (3)), there are two modifications: 1) The smoothed representation $\boldsymbol{P}\boldsymbol{H}^{(l)}$ is combined with the first layer $\boldsymbol{H}^{(0)}$ by an initial residual connection. 2) An identity matrix $\boldsymbol{I}_n$ is added to the l^{th} weight matrix $\boldsymbol{W}^{(l)}$. Intuitively, initial residual ensures that the final representation of each node reserves at least a fraction of the input features even if we stack many layers, which partially relieves over-smoothing. Note that $\boldsymbol{H}^{(0)}$ does not have to be the feature matrix $\boldsymbol{X}$, which means that a fully connected layer can be applied on $\boldsymbol{X}$ to obtain the initial hidden representation $\boldsymbol{H}^{(0)}$. On the other hand, identity mapping ensures that a deep GCN model achieves at least the same performance as its shallow version dose by making the decay of the weight matrix adaptively increases as the network goes deeper. In practice, we set $\beta_l = \log(\frac{\lambda}{l}+1)$, where λ is a hyperparameter.

2.3.3 Multilayer Perceptron

The output of the last graph convolutional layer is input to the multilayer perceptron (MLP) to predict the protein-interacting probabilities of all n amino acid residues:

$$\boldsymbol{Y}' = Softmax\left(\boldsymbol{H}^{(L)}\boldsymbol{W} + b\right) \quad (5)$$

where $\boldsymbol{H}^{(L)} \in \mathbb{R}^{n\times d}$ is the output of the L^{th} graph convolutional layer; $\boldsymbol{W} \in \mathbb{R}^{d\times 2}$ is the weight matrix; $b \in \mathbb{R}$ is the bias term, and $\boldsymbol{Y}' \in \mathbb{R}^{n\times 2}$ is the predictions of n amino acid residues. The softmax function normalizes the output of the network into a probability distribution over the two predicted classes (non-interacting and interacting).

2.4 Implementation details

We performed 5-fold cross-validation on the training data, where the data were split into five folds randomly. Each time, a model was trained on four folds and evaluated on the remaining one fold. This process was repeated five times and the performances on the five folds were averaged as the overall validation performance, which was used to choose the best feature combination and optimize all hyperparameters through grid search (**Supplementary Table S3**). To reduce the fluctuation from the random split of the folds, we repeated this cross-validation process five times with five different random seeds and calculated the average performance in the feature ablation study. Subsequently, the final model was trained using all training data and tested on the independent test set.

Specifically, we utilized an 8-layer GraphPPIS framework with 256 hidden units and the following set of hyperparameters: $\alpha = 0.7$, $\lambda = 1.5$, learning rate of 0.001, weight decay of 0 and batch size of 1. The dropout rate was set to 0.1 to avoid overfitting and the cutoff used to transform protein distance maps into adjacency matrices was set to 14 Å. We implemented the proposed model with Pytorch 1.6.0 (Paszke, et al., 2019) and employed cross-entropy loss and Adam optimizer (Kingma and Ba, 2015) for optimization. The training process lasted at most 50 epochs and took approximately 15 minutes on an Nvidia GeForce GTX 1080 Ti GPU.

2.5 Evaluation metrics

Similar to previous studies, we used accuracy (ACC), precision, recall, F1-score (F1), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) to measure the predictive performance:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F1 = 2\times\frac{Precision\times Recall}{Precision+Recall} \quad (9)$$

$$MCC = \frac{TP\times TN - FN\times FP}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \quad (10)$$

where true positives (TP) and true negatives (TN) denote the number of interacting and non-interacting sites identified correctly, and false positives (FP) and false negatives (FN) denote the number of incorrectly predicted interacting and non-interacting sites, respectively. AUROC and AUPRC are independent of thresholds, thus revealing the overall performance of a model. The other metrics were calculated using a threshold to

convert predicted interacting probabilities to binary predictions, which was determined by maximizing F1-score for each model. We used AUPRC for the above hyperparameter selection as it is more sensitive and informative and it emphasizes more on the minority class in imbalanced two-class classification tasks (Saito and Rehmsmeier, 2015).

2.6 False positive distribution analysis

Since native protein binding sites generally cluster together as a patch on the protein surface, we analyzed the spatial distribution of the false-positive sites predicted by different models. For a given protein in the test set, different models may predict different numbers of false positives. We took the top k false positives with the highest predicted scores for each model, where k is the lowest number of false positives predicted by all models. We measured the spatial dispersion of the false positives in a protein by an indicator similar to the univariate variance:

$$Dev(\boldsymbol{X}) = \frac{\sum_{i=1}^{k} d(\boldsymbol{x}_i, \overline{\boldsymbol{x}})^2}{k} \quad (11)$$

where $\boldsymbol{X}$ is the set of all top k false positives of the protein; $\boldsymbol{x}_i \in \boldsymbol{X}$ is the i^{th} false-positive site; $\overline{\boldsymbol{x}}$ is the average of the Cα atom coordinates of the false positives, representing their centroid; $d(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$ denotes the Euclidean distance between the i^{th} false-positive site and the centroid; and $Dev(\boldsymbol{X})$ denotes the deviation of all false positives to their centroid, indicating the dispersion of the false positives. On the other hand, we also analyzed the average distance between the false positives and the native PPI sites, by calculating the minimal distance between a false-positive site and all true PPI sites and averaging over all false positives of the protein:

$$Dis(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{k} \min_{\boldsymbol{y} \in \boldsymbol{Y}} d(\boldsymbol{x}_i, \boldsymbol{y})}{k} \quad (12)$$

where $\boldsymbol{Y}$ denotes the set of all true PPI sites. A small $Dev(\boldsymbol{X})$ and $Dis(\boldsymbol{X}, \boldsymbol{Y})$ means that the false-positive interacting sites are clustered and adjacent to the native PPI sites.

3 Results and Discussions

3.1 Feature importance and model ablation

We evaluated the performance of GraphPPIS by AUROC and AUPRC using 5-fold cross-validation (CV) and the independent Test_60. As shown in **Table 2**, we obtained AUROC values of 0.783±0.002 and 0.786; as well as AUPRC values of 0.423±0.003 and 0.429 on the 5-fold CV and independent test, respectively. The consistent performances on the cross-validation and independent test indicate the robustness of our model.

To demonstrate the relative importance of each feature we adopted, we conducted feature ablation experiments by using one feature individually or excluding one feature from the final feature combination. As shown in **Table 2**, when using single feature group as node features, the structural feature group (DSSP) gives better performance than sequence-based evolutionary feature PSSM or HMM, indicating that structural features such as secondary structure and relative solvent accessibility are more directly relevant to the identification of PPI sites. Interestingly, although PSSM performs worse than HMM when using single feature group, the removal of PSSM causes a larger performance drop than HMM (AUPRC of 0.429 to 0.394 and 0.404, respectively). In addition, the removal of DSSP leads to the greatest drop of AUPRC from 0.429 to 0.283. The performance reduction when removing any feature group suggests that the combined feature groups are nonredundant.

Since the computations of PSSM and HMM are time-consuming, we have also developed a simplified version of GraphPPIS to replace these two evolutionary features with amino acid encoding using BLOSUM62, a widely used substitution matrix for protein sequence alignment built using sequences with less than 62% similarities (Mount, 2008). This computationally efficient version of GraphPPIS achieves slightly worse yet satisfactory performance (only 2% reduction in AUROC for the test set) and is available on our web server as well.

Table 2. The AUROC and AUPRC of the 5-fold cross-validation (CV) and independent test (Test_60) using a single feature individually or excluding each feature in turn from the final feature combination. GraphPPIS* is a computationally efficient version which replaces evolutionary information with BLOSUM62 encoding.

Feature group	CV AUROC	CV AUPRC	Test AUROC	Test AUPRC
PSSM	0.659±0.004	0.263±0.003	0.654	0.244
HMM	0.684±0.002	0.280±0.006	0.684	0.275
DSSP	0.726±0.003	0.320±0.006	0.709	0.301
-PSSM	0.769±0.001	0.386±0.003	0.766	0.394
-HMM	0.779±0.001	0.397±0.002	0.776	0.404
-DSSP	0.687±0.003	0.281±0.003	0.691	0.283
GraphPPIS	**0.783±0.002**	**0.423±0.003**	**0.786**	**0.429**
GraphPPIS*	0.761±0.001	0.378±0.003	0.768	0.388

We further conducted a model ablation study to demonstrate the indispensability of initial residual and identity mapping utilized by GraphPPIS for solving over-smoothing, and the benefit of using a deeper graph convolutional network. As shown in **Supplementary Figure S1**, GraphPPIS achieves better performance when the network goes deeper, and reaches the best performance with 8 layers, while maintaining similar results as we further increase the network's depth.

3.2 Evaluating the effect of protein distance map

As mentioned in **Section 2.2.2**, we transformed protein distance maps into discrete adjacency matrices by converting any value in the distance maps less than or equal to the preset cutoff to 1 and any value greater than the cutoff to 0. **Figure 2** shows the performance of GraphPPIS on Test_60 at different distance map cutoffs. Note that the cutoff of 3.8 Å is the limit, when each amino acid residue in a protein only connects with itself and its one-hop sequence neighbors, because the distance between two sequentially neighboring Cα atoms is 3.8 Å. The other end of the cutoff is +∞, when each residue connects with all residues in the protein, corresponding to an adjacency matrix with 1 for all elements. As shown by the red line in **Figure 2**, as the cutoff increases, the AUPRC of GraphPPIS increases rapidly due to the introduction of more informative edges in the protein graphs, and the model reaches its best performance at the cutoff of 14 Å. Subsequently, the performance slowly descends as the cutoff continues to increase, indicating that an overlarge cutoff will bring redundant and harmful connections.

Furthermore, we have tested the normalized protein distance maps in which values are in the continuous range of 0 to 1 (Equation (2)). As shown by the blue line in **Figure 2**, the performance of GraphPPIS using continuous distance maps increases gradually along with the cutoff and reaches the maximal AUPRC at around 30 Å with only a slight decrease

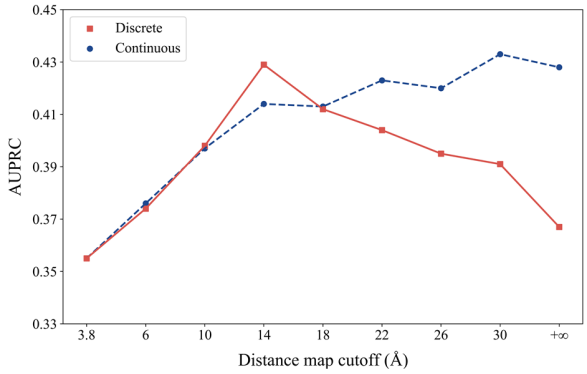

Figure 2. Performance of GraphPPIS on Test_60 using discrete (red) and continuous (blue) distance maps with different cutoffs.

using +∞ cutoff. This is reasonable because Equation (2) will convert a large distance to a small value when an overlarge cutoff is chosen. Practically, satisfactory performance can be obtained simply not setting any cutoff when using continuous maps. In general, these two types of maps may end up with similar performance with a proper choice of the cutoff for discrete maps. Throughout this paper, the results of GraphPPIS are based on discrete maps at the 14 Å cutoff because they allow faster convergence in the training step.

3.3 Performance comparison with other methods

We compared GraphPPIS with five sequence-based (PSIVER, ProNA2020, SCRIBER, DLPred and DELPHI) and three structure-based (DeepPPISP, SPPIDER and MaSIF-site) predictors. Note that our test sets may be part of the training sets in other methods. If true, the results reported here would be upper limits for other methods.

As shown in **Table 3**, there is a huge performance gap between GraphPPIS and the five sequence-based methods (AUPRC ranging from 0.190 to 0.319), whose performance rank is consistent with what reported previously (Li, et al., 2020). The poor performance of DeepPPISP may be ascribed to its lack of the essential RSA feature in amino acid representation, as we achieved similar performance with a 2-layer bidirectional LSTM model using all features except RSA (**Supplementary Table S4** shows detailed hyperparameters). For a more direct comparison, we also used the same training and test data as DeepPPISP to train and test our model. Still, our method outperforms DeepPPISP significantly (**Supplementary Table S5)**. On the other hand, SPPIDER has confirmed the importance of spatially neighboring residues as it improves over LSTM with all node features and GraphPPIS without contact information (by setting the distance map cutoff to 3.8 Å). The performance of GraphPPIS without contact information is marginally poorer than that of LSTM. This is understandable because LSTM is more powerful for sequential data in the absence of spatial information. Nevertheless, incorporating the spatial information coupled with the deep nonlinear architecture, GraphPPIS improves over LSTM by 18.2% and 21.5% and over SPPIDER by 15.0% and 16.8% on AUPRC and MCC, respectively. Note that GraphPPIS outperforms MaSIF-site on F1, MCC and AUROC, but is slightly worse than MaSIF-site on AUPRC. A closer inspection indicates that there are 35 proteins in Test_60 sharing more than 25% similarities with the training set of MaSIF-site, which may lead to performance overestimation of MaSIF-site. The precision-recall curves and the ROC curves of GraphPPIS and other methods on Test_60 can be found in **Supplementary Figure S2** and **Figure S3**. After further inspection, we noted that there are still 4 proteins in Test_60 sharing more than 25% similarities over 30% coverage with our training set. Removing these proteins from Test_60 yielded essentially the same result (**Supplementary Table S6**), which highlights the robustness of our trained model.

To further demonstrate the generalization and stability of our method, we compared GraphPPIS with other structure-based methods on our newly built independent dataset Test_315. As shown in **Table 4**, GraphPPIS achieves similar performance as in Test_60 for newly solved proteins (AUPRC of 0.423 and MCC of 0.336). More importantly, GraphPPIS consistently outperforms DeepPPISP and SPPIDER significantly and outperforms MaSIF-site by 13.7% and 10.5% on AUPRC and MCC, respectively. The improved performance of GraphPPIS over other structure-based methods on Test_315 can be further illustrated by the

Table 3. Performance comparison with other methods on Test_60. Predictions by the programs marked with * were obtained from (Li, et al., 2020). The results of DeepPPISP and MaSIF-site were obtained from their source codes. Predictions by PSIVER, ProNA2020 and SPPIDER were directly generated from their web servers. ProNA2020 only makes binary predictions and thus, AUROC and AUPRC are not calculated.

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**
LSTM[a] (no RSA)	0.662	0.245	0.551	0.340	0.178	0.661	0.267
LSTM[a] (all features)	0.746	0.323	0.551	0.407	0.274	0.750	0.363
GraphPPIS (no contact)	0.741	0.317	0.554	0.403	0.269	0.744	0.355
GraphPPIS	0.776	0.368	**0.584**	**0.451**	**0.333**	**0.786**	0.429

[a] Replace the deep graph convolution module in GraphPPIS with a 2-layer bidirectional LSTM module.

GraphPPIS

Table 4. Performance comparison with other structure-based methods on the independent Test_315 and UBtest_31. Here, Btest_31 represents the 31 bound structures in Test_60 that have known monomeric structures.

Method	Test_315		Btest_31		UBtest_31	
	MCC	AUPRC	MCC	AUPRC	MCC	AUPRC
DeepPPISP	0.169	0.256	0.163	0.223	0.162	0.217
SPPIDER	0.294	0.376	0.240	0.315	0.222	0.260
MaSIF-site	0.304	0.372	0.217	0.299	0.141	0.225
GraphPPIS	**0.336**	**0.423**	**0.328**	**0.395**	**0.280**	**0.323**

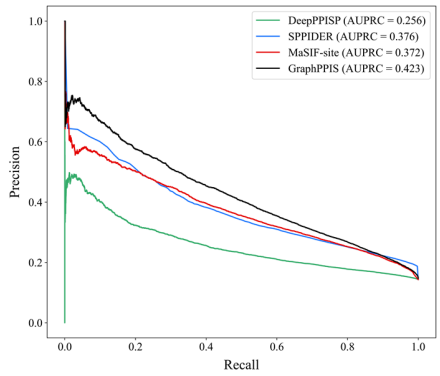

Figure 3. Precision-recall curves of GraphPPIS and other structure-based methods as labeled on Test_315.

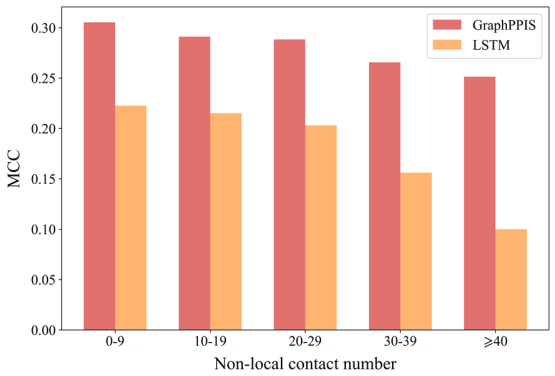

Figure 4. The MCC of GraphPPIS and LSTM on amino acids with different number of non-local contacts.

precision-recall curves (**Figure 3**) and the ROC curves (**Supplementary Figure S4**), where the curves of GraphPPIS largely locate on the upper portion of the figures.

Since our training set was built on native complex structures, it would be interesting to examine the impact on method performance when unbound structures are used for prediction. We compared the performance of GraphPPIS with other structure-based methods on a subset of Test_60 (Btest_31) and their corresponding unbound structures (UBtest_31). As shown in **Table 4**, the performance of these four structure-based methods both drops when predicting unbound monomeric structures, because they were all trained with bound structures. Specifically, MCC decreases by 35.0% for MaSIF-site but only 14.6% for GraphPPIS. In general, GraphPPIS continues to outperform other structure-based methods by more than 26.1% and 24.2% on MCC and AUPRC, respectively.

3.4 Impact of the long-range contacts

The performance improvement of GraphPPIS over LSTM is likely due to its better capability in capturing long-range contact information. To illustrate this, we compared the performance of GraphPPIS and LSTM on amino acids with different number of non-local contacts, defined as the contacts from the residues that are more than or equal to 20 residues away in sequence positions, but spatially adjacent according to the adjacency matrix of the protein graph. **Figure 4** shows that GraphPPIS consistently surpasses LSTM on Test_60 and more importantly, the performance gap between them enlarges as the non-local contact number of the amino acids increases. Specifically, the performance of GraphPPIS surpasses LSTM by 37.2% in MCC on the amino acids with 0 to 9 non-local contacts, and the gap widens to as much as 151.6% on the amino acids with ≥ 40 non-local contacts. Similar results were observed with other measures such as F1 (**Supplementary Figure S5**). These results highlight the importance of the non-local contact information and the superior ability of GraphPPIS in harnessing non-local contacts for PPI site detection.

3.5 Spatial distribution of the false-positive sites

Native protein interacting sites should be clustered together to form connected patches. Here, we analyzed the spatial distribution of the false-positive sites predicted by GraphPPIS and LSTM using the measures described in **Section 2.6**. For each protein in Test_60, we used $Dev(\boldsymbol{X})$ to evaluate the spatial dispersion of the false positives predicted by GraphPPIS and LSTM, and the distributions over 60 proteins are shown in **Figure 5A**. The average $Dev(\boldsymbol{X})$ of GraphPPIS and LSTM are 273.0 Å^2 and 368.7 Å^2, respectively, indicating that the false positives predicted by GraphPPIS are significantly more clustered (P-value = 3.6×10^{-9}) according to the Wilcoxon Signed Rank Test (Wilcoxon, 1945). On the other hand, we analyzed the distances between the false positives and the native PPI sites by calculating $Dis(\boldsymbol{X},\boldsymbol{Y})$ (**Figure 5B**). The average $Dis(\boldsymbol{X},\boldsymbol{Y})$ of GraphPPIS and LSTM are 11.2 Å and 13.3 Å, respectively, suggesting that the false positives predicted by GraphPPIS are significantly closer to the native PPI sites (P-value = 5.0×10^{-10}). These results demonstrate that with the spatial information and deep graph convolution technique, the false-positive interacting sites predicted by GraphPPIS are more clustered and closer to the true positives, and therefore are more likely to be potential PPI sites.

Figure 6 shows the PPI site prediction results of GraphPPIS (**A**) and LSTM (**B**) for guanine-nucleotide-exchange factors ARNO from *Homo sapiens* (PDB ID: 1R8S, chain E). In this example, there are 40 protein-binding residues over a total of 187 residues. GraphPPIS predicts 47 binding residues in which 32 are true positives, leading to an MCC of 0.660. By comparison, LSTM predicts 52 binding residues with only 25 residues as true positives, leading to 80% more false positives than GraphPPIS and

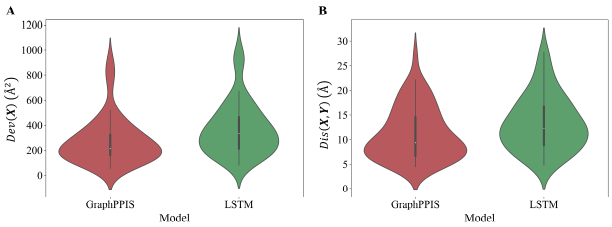

Figure 5. The distributions of $Dev(\boldsymbol{X})$ (**A**) and $Dis(\boldsymbol{X},\boldsymbol{Y})$ (**B**) by GraphPPIS and LSTM on Test_60. Each boxplot indicates the median and quartiles with whiskers reaching up to 1.5 times the interquartile range. Each violin plot illustrates the kernel probability density, where the shaded area represents the proportion of the samples located there.

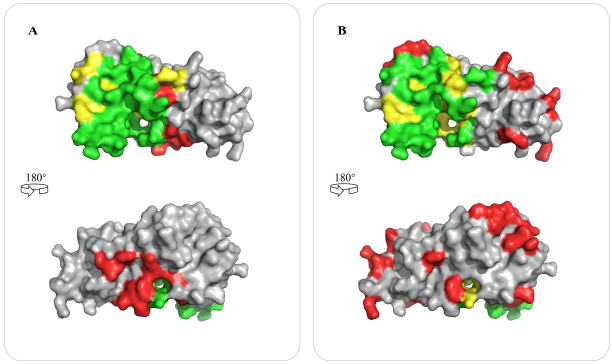

Figure 6. One example (PDB ID: 1R8S, chain E) of protein-binding residue predictions by GraphPPIS (**A**) and LSTM (**B**). True positives, false positives and false negatives are colored in green, red and yellow, respectively. Top and bottom panels show the front and back views of the protein

a lower MCC of 0.404. As shown in **Figure 6A**, false-positive binding residues (colored in red) predicted by GraphPPIS are mostly around the interface of protein-protein interaction, while there are many scattered and isolated false positives predicted by LSTM far away from the native PPI sites (**Figure 6B**). **Supplementary Figure S6** shows another representative example (Carboxypeptidase A1, PDB ID: 3FJU, chain A) where LSTM predicts spatially scattered false positives while GraphPPIS makes more clustered predictions.

4 Conclusion

In this study, we propose a deep graph-based method called GraphPPIS for PPI site prediction, where we construct the protein graphs using distance-based adjacency matrices and employ deep graph convolution technology to learn the node representations. GraphPPIS shows preferable performance than existing sequence-based and structure-based methods in comprehensive evaluations.

However, there is still room for further improvement on this task, as the AUROC on the test set is less than 0.8. For instance, enlarging the dataset or using advanced sampling techniques for imbalanced problems may help train a better model. Adding other informative features to the existing feature groups may facilitate better representation of the proteins, such as pre-trained amino acid embedding, relative amino acid propensity, and physicochemical properties. In the future, we would further extend our framework to a purely sequence-based method based on predicted distance maps through SPOT-Contact (Hanson, et al., 2018) or predicted protein structural models through AlphaFold2 (Jumper, et al., 2021). We also expect that our graph-based architecture can be easily applied to various fields, including predicting protein binding sites with other molecules, such as DNA, RNA and small ligands.

Funding

This study has been supported by the National Key R&D Program of China (2020YFB020003), National Natural Science Foundation of China (61772566, 62041209), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), Introducing Innovative and Entrepreneurial Teams (2016ZT06D211), Guangzhou S&T Research Plan (202007030010), Shenzhen Science and Technology Program (Grant No. KQTD20170330155106581) and the Major Program of Shenzhen Bay Laboratory S201101001.

Conflict of Interest: none declared.

References

Altschul, S.F., *et al.* Basic local alignment search tool. *Journal of molecular biology* 1990;215(3):403-410.

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997;25(17):3389-3402.

Berman, H.M., *et al.* The protein data bank. *Nucleic acids research* 2000;28(1):235-242.

Buchfink, B., Xie, C. and Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* 2015;12(1):59-60.

Capra, J.A. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007;23(15):1875-1882.

Chen, J., *et al.* Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *Journal of cheminformatics* 2021;13(1):1-10.

Chen, M., *et al.* Simple and Deep Graph Convolutional Networks. In, *Proceedings of the 37th International Conference on Machine Learning*. 2020.

Chen, S., *et al.* To Improve Protein Sequence Profile Prediction through Image Captioning on Pairwise Residue Distance Map. *Journal of Chemical Information and Modeling* 2019;60(1):391-399.

De Las Rivas, J. and Fontanillo, C. Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. *Briefings in functional genomics* 2012;11(6):489-496.

Deng, A., *et al.* Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *International journal of molecular sciences* 2020;21(7):2274.

Dhole, K., *et al.* Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. *Journal of theoretical biology* 2014;348:47-54.

Esmaielbeiki, R., *et al.* Progress and challenges in predicting protein interfaces. *Briefings in Bioinformatics* 2015;17(1):117-131.

Gainza, P., *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* 2020;17(2):184-192.

Guharoy, M. and Chakrabarti, P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* 2007;23(15):1909-1918.

Hammes, G.G., Chang, Y.-C. and Oas, T.G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences* 2009;106(33):13737-13741.

GraphPPIS

Hanson, J.*, et al.* Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* 2018;34(23):4039-4045.

Hwang, H.*, et al.* Protein–protein docking benchmark version 3.0. *Proteins: Structure, Function, and Bioinformatics* 2008;73(3):705-709.

Jumper, J.*, et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021:1-11.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen‐bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* 1983;22(12):2577-2637.

Kingma, D.P. and Ba, J. Adam: A Method for Stochastic Optimization. In, *3rd International Conference on Learning Representations (Poster)*. 2015.

Kipf, T.N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In, *5th International Conference on Learning Representations*. 2017.

Kuzmanov, U. and Emili, A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome medicine* 2013;5(4):1-12.

Li, Q., Han, Z. and Wu, X.-M. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. In, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence; 2018. p. 3538-3545.

Li, Y., Golding, G.B. and Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* 2020.

Li, Y. and Ilie, L. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC bioinformatics* 2017;18(1):485.

Liang, S.*, et al.* Protein binding site prediction using an empirical scoring function. *Nucleic Acids Research* 2006;34(13):3698-3707.

Mirdita, M.*, et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* 2017;45(D1):D170-D176.

Murakami, Y. and Mizuguchi, K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 2010;26(15):1841-1848.

Northey, T.C., Barešić, A. and Martin, A.C.R. IntPred: a structure-based predictor of protein–protein interaction sites. *Bioinformatics* 2017;34(2):223-229.

Orii, N. and Ganapathiraju, M.K. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PloS one* 2012;7(11):e49029.

Paszke, A.*, et al.* Pytorch: An imperative style, high-performance deep learning library. In, *Advances in neural information processing systems*. 2019. p. 8026-8037.

Porollo, A. and Meller, J. Prediction‐based fingerprints of protein‐protein interactions. *Proteins: Structure, Function, and Bioinformatics* 2007;66(3):630-645.

Qiu, J.*, et al.* ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *Journal of molecular biology* 2020;432(7):2428-2443.

Rao, J.*, et al.* Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. *Iscience* 2021;24(5):102393.

Remmert, M.*, et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 2012;9(2):173-175.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 2015;10(3):e0118432.

Shoemaker, B.A. and Panchenko, A.R. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* 2007;3(3):e42.

Song, Y.*, et al.* Communicative representation learning on attributed molecular graphs. In, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,(IJCAI 2020)*. 2020. p. 2831-2838.

Suzek, B.E.*, et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;23(10):1282-1288.

Wang, B.*, et al.* Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM transactions on computational biology and bioinformatics* 2019.

Wells, J.A. and McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 2007;450(7172):1001-1009.

Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics* 1945;1(6):80-83.

Wu, F.*, et al.* Simplifying Graph Convolutional Networks. In, *International Conference on Machine Learning*. 2019. p. 6861-6871.

Xie, Z., Deng, X. and Shu, K. Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. *International journal of molecular sciences* 2020;21(2):467.

Zeng, M.*, et al.* Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 2020;36(4):1114-1120.

Zhang, B.*, et al.* Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* 2019;357:86-100.

Zhang, J. and Kurgan, L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Briefings in bioinformatics* 2018;19(5):821-837.

Zhang, J. and Kurgan, L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 2019;35(14):i343-i353.

Zhu, H., Du, X. and Yao, Y. ConvsPPIS: identifying protein-protein interaction sites by an ensemble convolutional neural network with feature graph. *Current Bioinformatics* 2020;15(4):368-378.