


SGPPI: structure-aware prediction of protein–protein interactions in rigorous conditions with graph convolutional network

Yan Huang, Stefan Wuchty, Yuan Zhou and Ziding Zhang 

Corresponding authors. Dr Ziding Zhang, State Key Laboratory of Livestock and Poultry Biotechnology Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China. E-mail: zidingzhang@cau.edu.cn; Dr Yuan Zhou, Department of Biomedical Informatics, Ministry of Education Key Laboratory of Molecular Cardiovascular Sciences, Center for Non-Coding RNA Medicine, School of Basic Medical Sciences, Peking University, Beijing 100191, China. E-mail: zhouyuanbioinfo@hsc.pku.edu.cn

Abstract

While deep learning (DL)-based models have emerged as powerful approaches to predict protein–protein interactions (PPIs), the reliance on explicit similarity measures (e.g. sequence similarity and network neighborhood) to known interacting proteins makes these methods ineffective in dealing with novel proteins. The advent of AlphaFold2 presents a significant opportunity and also a challenge to predict PPIs in a straightforward way based on monomer structures while controlling bias from protein sequences. In this work, we established Structure and Graph-based Predictions of Protein Interactions (SGPPI), a structure-based DL framework for predicting PPIs, using the graph convolutional network. In particular, SGPPI focused on protein patches on the protein–protein binding interfaces and extracted the structural, geometric and evolutionary features from the residue contact map to predict PPIs. We demonstrated that our model outperforms traditional machine learning methods and state-of-the-art DL-based methods using non-representation-bias benchmark datasets. Moreover, our model trained on human dataset can be reliably transferred to predict yeast PPIs, indicating that SGPPI can capture converging structural features of protein interactions across various species. The implementation of SGPPI is available at <https://github.com/emerson106/SGPPI>.

Keywords: protein–protein interaction, deep learning, graph convolutional neural network, structure, prediction

Introduction

The accumulation of protein–protein interaction (PPI) network data provides the mechanistic understanding of complex cellular events in biological systems [1, 2], playing a vital role in drug discovery and therapy development [3, 4]. Experimental approaches for the large-scale identification of PPIs in model organisms have been in progress over the past few decades [5]. Although often regarded as the golden standard, the experimental determination of PPIs depends on specific experiment conditions, leaving their coverage often limited. As experimental methods are often time-consuming and labor-intensive, *in silico* approaches, such as machine learning (ML)-based methods, have become increasingly popular to provide testable hypotheses and as the supplement [6, 7].

ML models predict novel interactions by learning hidden features of known PPIs and are often based on the similarity criteria, which assume that proteins sharing a common interaction partner should contain common characteristics. Such properties usually capture the physicochemical properties of amino acid sequences, structural similarity, evolutionary identity, PPI network

partners or topological properties [8–10], with protein sequence features enjoying the greatest popularity. As the primary structures of proteins, amino acid sequences largely encode the functions of proteins. Considering that protein sequences are abundant, many research efforts utilized protein sequence features to predict PPIs. In 2005, Martin *et al.* developed a signature molecular descriptor to encode protein sequences to predict PPIs with a support vector machine (SVM) classifier [11]. Similarly, Shen *et al.* proposed the conjoint triad (CT) descriptors to represent the composition of amino acid sequences in a compact framework [12]. Since CT descriptors are not suitable to capture the long-range interactions of residues which are important for the description of PPIs [13], Guo *et al.* developed an auto covariance encoding strategy to reflect the neighboring effects of residues. Other component frequency-based coding strategies, such as composition-transition-distribution and composition of k-spaced amino acid pairs (CKSAAP), were also widely used in PPI predictions [14, 15]. In 2012, Pitre *et al.* developed PIPE2, which estimates the polypeptide sequence similarity between query proteins and known PPIs to determine whether two proteins interact [16].

Yan Huang is a PhD student at the State Key Laboratory of Livestock and Poultry Biotechnology Breeding, College of Biological Sciences, China Agricultural University. His current research interests include protein bioinformatics and machine learning.

Stefan Wuchty is an Associate Professor at the Department of Computer Science and Biology, and a member of the Institute of Data Science and Computing and Sylvester Comprehensive Cancer Center at the University of Miami. His research interests revolve around systems and network biology.

Yuan Zhou is an Associate Investigator at the Department of Biomedical Informatics, Peking University. His research interest includes transcriptomic and epitranscriptomic bioinformatics.

Ziding Zhang is a Professor at the State Key Laboratory of Livestock and Poultry Biotechnology Breeding, College of Biological Sciences, China Agricultural University. His research interests are protein bioinformatics and computational systems biology.

Received: September 9, 2022. **Revised:** November 17, 2022. **Accepted:** January 5, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Although sequence-based approaches have proven to be effective, encodings of interacting protein sequences alone cannot fully capture all PPI-relevant information. For example, evolutionary profiles of sequences and structures provide additional features beyond sequence composition, which allows more powerful PPI prediction. Specifically, Zahiri *et al.* represented the interacting proteins with position-specific scoring matrices (PSSMs) [17], while Hamp and Rost [18] used evolutionary profiles showing increased prediction performance and robustness.

Deep learning (DL) technologies have recently been applied to predict PPIs [19–21]. In particular, multilayer perceptrons and convolutional neural networks can predict PPIs by embedding protein sequence characteristics [22–24]. Natural language processing methods were also used to effectively convert amino acid sequences into high-dimensional vectors for PPI predictions [25, 26]. DL models can also be jointly used to better leverage their strengths. For instance, Chen *et al.* combined the advantages of both convolutional and recurrent neural networks to predict PPIs, capturing both local significant features and sequence characteristics from the primary protein sequences [27].

Still, such largely sequence-based prediction models discount for protein structural features, but proteins actually exert their functions by folding into three-dimensional (3D) structures that bind other molecules in the 3D structural space. To this end, Zhang *et al.* developed a successful structure-informed PPI prediction method by first searching a complex template for the query protein through sequence and structural alignment and then predicting the interaction probability via a Bayesian classifier [10, 28]. While the most prominent obstacle to include protein structural features for PPI predictions had been the scarcity of accurate large-scale protein structures, the recent development of AlphaFold2 [29] allows the prediction of protein monomer structures from protein sequences with an accuracy comparable to experiment methods, offering an avenue to account for protein structures in the prediction of PPIs.

Compared with linear sequences, protein 3D structures are more challenging in feature extraction due to their complex topologies. To address this issue, a frequently used strategy is to convert the protein structures into residue networks or graphs in which residues can be regarded as nodes, while residue contacts are regarded as edges. Graph convolutional network (GCN) is one of the most popular DL models in capturing structural relations among such graph-structured data. In the field of protein bioinformatics, GCN has been widely used for protein binding interface predictions, protein function annotation and drug discovery [30]. For example, Torng *et al.* trained a GCN model to extract features from protein pocket and ligand graph representations and achieved competitive performance on the common virtual screening benchmark datasets [31]. Gligorijevic *et al.* achieved rapid prediction of protein functions from computationally inferred structures by integrating GCN and a natural language model [32]. Recently, Yuan *et al.* developed a GCN-based model to predict the residues that are likely to be involved in interacting with other proteins [33]. Although the interacting partner proteins of these residues were not specifically predicted in this model, its sound performance at least indicated that GCN could serve as a promising architecture for depicting interacting residues. On the other hand, it is therefore interesting to explore how the GCN representation of residue networks could better describe the interactions between specific protein pairs and predict specific PPIs.

To this end, in this work, we established a GCN-based PPI prediction model, Structure and Graph-based Predictions of Protein Interactions (SGPPI). To learn the structural patterns of PPIs, SGPPI

considered both the global structural characteristics of proteins and the local structural features from the patches at the potential protein interaction interfaces. Moreover, SGPPI also incorporated the evolutionary profiles into the structural representation of PPIs to improve its performance. In the following sections, we will first introduce the construction of benchmark datasets and the implementation of SGPPI. The performance assessment of the proposed model and the corresponding analysis will then be described.

Materials and methods

Construction of benchmarking datasets

We regarded the PPI prediction task as a pair-input problem, indicating that each pair of instances (proteins) is the item of input to the model. Such pair-input problems often have specific requirements for both the dataset and the model architecture. As for the training and testing datasets, as Park and Marcotte have mentioned [34], the predictive performance of pair-input methods may be overestimated due to shared instances (proteins) between training and test sets [34]. In other words, a classical PPI prediction model, which was trained on a dataset with many frequently presented similar proteins, is prone to detect interactions as the consequence of over-representation of proteins that are more likely to be involved in PPIs (e.g. hubs in the PPI network) rather than predicting specific PPIs. To avoid such representation bias, we employed three rigorous benchmarking datasets [Profppikernel dataset, Human Reference Interactome (HuRI) dataset and (filtered) Pan's dataset] to benchmark the performance of different PPI prediction methods, as specified below.

Compiled by Hamp and Rost [18], Profppikernel dataset captured both human and yeast PPI data, which were collected from reliable human and yeast interactions as of the Hippie V1.2 (10/2011) [35] and Database of Interacting Proteins [36], respectively. To limit the influence of sequence similarity of proteins, the sequence redundancy of interacting proteins was removed by setting the sequence identity threshold to 40%. After the application of such rigorous sequence similarity limits, 842 human PPIs and 746 yeast PPIs remained.

HuRI human dataset and Pan's dataset were collected as the two alternative datasets for further benchmarking the performance of SGPPI. HuRI human dataset was based on the HuRI Mapping Project, which detected 64 000 PPIs involving 9000 human proteins by high-throughput yeast two-hybrid screens [37]. This comprehensive PPI map makes it possible to build a larger benchmarking dataset for PPI predictive models. We followed Profppikernel's strategy [18] and constructed a non-representation-bias dataset containing 1706 PPIs. The original Pan's dataset [38] (http://www.csbio.sjtu.edu.cn/bioinf/LR_PPI/Data.htm) was based on the Human Protein References Database (HPRD, 2007 version) and has been frequently used as the benchmarking dataset for PPI predictions [26, 27, 39]. We also followed Profppikernel's strategy to filter Pan's dataset and converted it into a non-representation-bias dataset that covered 1160 PPIs.

We applied 10-fold cross-validation to evaluate the performance of our PPI prediction model. In all three benchmark datasets, the interacting pairs of proteins were divided into 10 subgroups for cross-validation. For each subgroup, we randomly sampled 10 times as many negative interactions as positive ones, ensuring that all these samples were sequence-dissimilar to positive training PPIs in the given subgroup. In this work, we plotted the Precision-Recall (PR) curve to reflect the overall relationship between precision and recall when different

predictive thresholds are applied and mainly used the area under PR curve (AUPRC) to quantify the predictive performance. We also plotted the receiver operating characteristic (ROC) curve to reflect the relationship between sensitivity and specificity and used the area under ROC curve (AUROC) for performance measurement. In addition, we also introduced three common performance metrics (i.e. Precision, Recall and F1-score) for method evaluation and comparison. These three metrics are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (1)$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}, \quad (2)$$

$$F1 = \frac{TP}{TP + FP}, \quad (3)$$

where TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative samples, respectively.

Computational framework of SGPPI

SGPPI used a Siamese network architecture to represent and predict the interacting proteins (Figure 1), where each protein was characterized separately. In the feature representation module, each protein was represented by a contact map of protein residues based on the 3D structure as provided by AlphaFold2. Such an undirected network of residues was further refined by contacts between the patch residues on protein surfaces that are close to each other and likely to contain protein binding hotspots. To capture more local and evolutionary information conducive to protein binding, each residue was further annotated by the corresponding values in the PSSM profile and location in the underlying protein secondary structure. After a series of graph convolutions of such protein information, the resulting feature vectors of interacting proteins were merged to predict the presence (absence) of an interaction between a pair of proteins through fully connected layers.

Graph representation of protein structures

As residue contact maps were used to represent protein 3D structures, we collected the 3D atomic coordinates of all proteins in the benchmark dataset from AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk/>). Considering that not all residues contribute equally to PPIs, we discarded residues buried inside the protein structure by discarding residues with relative solvent accessibility <0.2 [40].

Numerous studies have shown that certain patches on the protein surfaces tend to contain protein binding hotspots, and residues in such patches tend to be more hydrophobic and more conserved than other residues [41–43]. In particular, the surface and patch residues determine the structural properties of the protein, suggesting that such residues can further improve our contact map. Here, we used JET2 [44] to recognize the patches on protein surfaces, where we defined a graph $G=(V, E)$, where V represents all the considered surface residues, and E denotes residue-residue contacts. A contact was identified if the geometrical distance of any two residues' $C\alpha$ atoms is less than a certain threshold (default 10 Å), allowing us to represent a protein structure by an undirected graph of the surface/patch residues. Subsequently, amino acid sequence, protein secondary structures, geometric features and evolutionary information features were encoded and mapped onto every node (i.e. every residue) in

the undirected graph and were finally integrated into our GCN framework.

Feature encodings

PSSM encoding evolutionary information

We used PSSM profiles to reflect the conservation and mutation profile of each amino acid residue, which corresponds to a 20-dimensional vector, reflecting the conservation of 20 amino acids at the corresponding position among a set of homologous sequences. PSSM profiles were generated by PSI-Basic Local Alignment Search Tool [45] search against the NR90 database (version of November 2019) at NCBI with three iterations and E-value cutoff of 0.0001.

JET2 features encoding local and global geometrical information

JET2 is a dedicated protein surface patch identification algorithm, which divides each protein interface into three structural regions, called seed, extension and outer layer, from the core part to the rim part [44]. The seeds of proteins in our datasets were identified by computing the conservation level of every residue in the protein sequences. Then, JET2 extended the interface patches based on these seeds by combining the evolutionary traces of each residue, interface propensities and circular variances (CVs) computed from the protein structure. We introduced the above features as a five-dimensional vector to characterize each node (surface residue) in the residue contact map. First, we used one binary byte to indicate whether the residue belongs to the protein interface calculated by JET2 and used two values to reflect the accessibility at the atomic and residue levels, respectively. Second, CV measures the vectorial distribution of a set of neighboring points around a fixed point in 3D space. Accordingly, the last two values in the five-dimensional vector of each residue were the global CV and local CV calculated by JET2, which described the geometry of the entire protein surface and local residues, respectively. Overall, the five JET2 descriptors well portrayed the geometric features of proteins at various levels, which enrich the node characteristics of the residue contact map from the perspective of structural interactions.

One-hot encoding of protein secondary structural information

DSSP [46, 47] was used to identify the secondary structures of the given proteins. The secondary structures were captured by eight states in DSSP, suggesting an eight-dimensional one-hot vector to characterize the secondary structure state of each residue.

Graph convolutional neural network

In this work, we utilized GCN to propagate and extract the hidden features from protein structures. Given a protein that is represented by a (surface) residue contact map with n nodes, the input of GCN includes two parts: an adjacency matrix $A \in \mathbb{R}^{n \times n}$ and a node feature matrix $X \in \mathbb{R}^{n \times m}$, where m is the length of the feature vector of residues ($m=33$ in this work). We applied two-layer GCN to update the hidden features of residues using the following update rule:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}), \quad (4)$$

where H denotes the hidden state in the convolution process and $l \in \{0, 1\}$, $W^{(l)}$ denotes the weight matrix associated with the

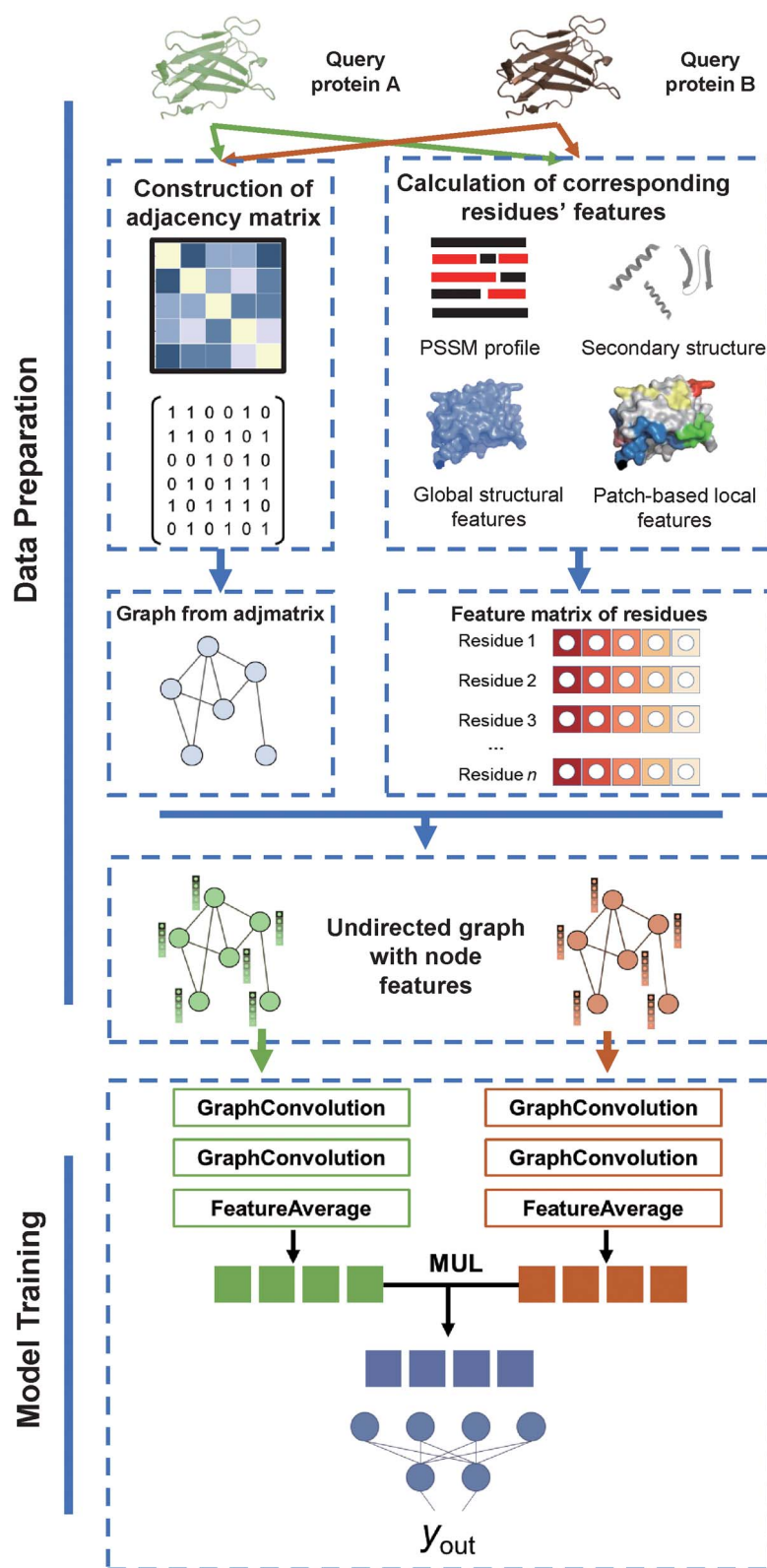


Figure 1. Overview of SGPI workflow. SGPI uses a Siamese network architecture to represent and predict interacting proteins. By identifying the interface patches and nearby surface residues, each protein is characterized by its monomer structures predicted by AlphaFold2 through a contact network map of residues at a certain threshold. We refined such a map with node features, including (i) evolutionary information of the residue through PSSMs; (ii) location in the underlying protein secondary structure and (iii) global and local geometrical descriptors. Merging feature vectors of interacting proteins after a series of graph network convolutions, we predict the interaction probability of a given protein pair through fully connected layers.

lth-layer of the GCN. We applied ReLU as the nonlinear activation function. Finally, \tilde{A} is computed as the following equation:

$$\tilde{A} = A + I, \quad (5)$$

where A denotes the adjacency matrix. \tilde{D} in Equation (1) is the diagonal degree matrix of \tilde{A} . After convolution, the average hidden feature of all residues contains the concerned information of the given protein. For the pair-input proteins, a Siamese-like architecture obtains the well-extracted features s^1 and s^2 through a parameter-sharing GCN. The probability of interaction for the given pair proteins y_{out} was calculated by

$$y_{out} = \text{SoftMax}(f(s^1 \odot s^2)), \quad (6)$$

where \odot denotes the Hadamard product and f denotes the fully connected feedforward neural network.

Baseline methods for predicting PPIs

Baseline encoding strategy

We applied two typical composition-based encoding schemes as the baseline descriptors of protein sequences: (i) amino acid composition (AAC) reflects the percentage of 20 standard amino acids in a given protein sequence, which is formulated as follows:

$$\text{AAC} = (f_1, f_2, \dots, f_{20}), \quad (7)$$

where f_i is the ratio of a certain kind of amino acid over all amino acids in the given protein. (ii) CKSAPP is the rate of k -spaced amino acid pairs normalized by all possible 400 kinds of pair combinations. CKSAAP can be formulated as

$$\text{CKSAAP} = (f_1^0, f_2^0, \dots, f_{400}^0, f_1^1, f_2^1, \dots, f_{400}^1, \dots, f_1^k, f_2^k, \dots, f_{400}^k), \quad (8)$$

where f is the ratio of a specific k -spaced amino acid pair over all possible pairs in the given protein. In this work, we set $k=0, 1, 2, 3$, pointing to the representation of a protein sequence as a 1600-dimensional vector.

Baseline ML methods

Random forest (RF) and SVM are two classical ML methods commonly used for various bioinformatics prediction problems. We compared these two classical methods with SGPPI as the baseline methods in this work. RF and SVM were implemented through the sklearn library in Python. Grid search was used to optimize the parameters in RF and SVM. Specifically, in the RF model with CKSAAP encoding, the optimized `n_estimators` and `max_depth` were 300 and 12, respectively. In the RF model with AAC encoding, the optimized `n_estimators` and `max_depth` were 100 and 9, respectively. In the SVM model, we chose the commonly used 'rbf' as the kernel function, and the optimized `C` and `gamma` were 0.1 and 16, respectively.

Results

The overall performance benchmarking strategy

To measure the prediction performance of SGPPI, we mainly used PR and ROC curves to measure the performance of the predictors through 10-fold cross-validation on the following three PPI datasets. The Profppikernel human dataset has been widely used to test the performance of PPI prediction methods [18, 48],

Table 1. Performance of different state-of-the-art PPI prediction methods on the Profppikernel dataset

Methods	AUPRC ^a
SGPPI	0.422
Profppikernel ^b	0.359
PIPR ^c	0.342
PIPE2 ^b	0.220
SigProd ^b	0.264

^aAUPRC denotes the average AUPRC value of 10-fold cross-validation.

^bResults were retrieved from Hashemifar et al. [48]. ^cWe implemented PIPR using the source code on Github (https://github.com/muhaochen/seq_ppi) and chose the embedding vector that performed the best on the benchmark dataset.

which holds 842 positive and a corresponding set of 8420 negative PPI samples. To test the performance of the predictors more comprehensively, we also collected experimentally verified PPIs from the HuRI Mapping Project [37] and HPRD as two unbiased larger dataset alternatives (i.e. HuRI dataset and Pan's dataset). By adopting the same rigorous dataset preparation strategy against protein instance redundancy, each dataset does not allow proteins with similar sequences to appear in the training and the test samples. While all proteins in the test samples were dissimilar compared to the positive training samples, proteins in the test samples and negative training samples can be similar. As a consequence, any predictive model cannot predict PPIs simply by similarity to known proteins in the training PPIs, suggesting that all these three datasets are non-representation-bias benchmarking datasets.

Performance assessment on Profppikernel human dataset

To assess the predictive power of SGPPI, we compared its performance to methods that were already tested on the human Profppikernel dataset [48], including sequence-based methods such as Profppikernel, PIPE2 and SigProd. Briefly, Profppikernel [18] used evolutionary profiles, while SigProd [11] represented a sequence by 3-mers to classify the interactions between proteins through a SVM. PIPE2 predicted a PPI if subsequences of the interacting proteins occur more frequently in the positive training set [16]. We also compared the performance of SGPPI against a recent state-of-the-art approach, PIPR, which used a deep residual recurrent convolutional neural network (CNN) to capture both local features and the contextualized information of protein sequences. Notably, PIPR was reported to achieve the best performance in comparison to other state-of-the-art methods when predicting binary PPIs [27]. As the positive/negative data are imbalanced, PR curves are preferred over ROC curves as a comprehensive performance measure. In particular, we observed that SGPPI that integrates the sequence, evolutionary, structural and geometric information of proteins through GCN significantly outperforms all four state-of-the-art methods on the Profppikernel human dataset utilizing the AUPRC metric (Table 1).

Performance benchmarking on alternative datasets

We introduced two alternative datasets (i.e. HuRI dataset and Pan's dataset) to further assess the performance of SGPPI. In particular, we compared the performance of SGPPI to more traditional ML approaches, such as RF and SVM, where we represented the sequences of interacting proteins through vectors of AAC and CKSAPP. While traditional ML methods with routine encoding schemes failed to perform well on both HuRI datasets (Figure 2A

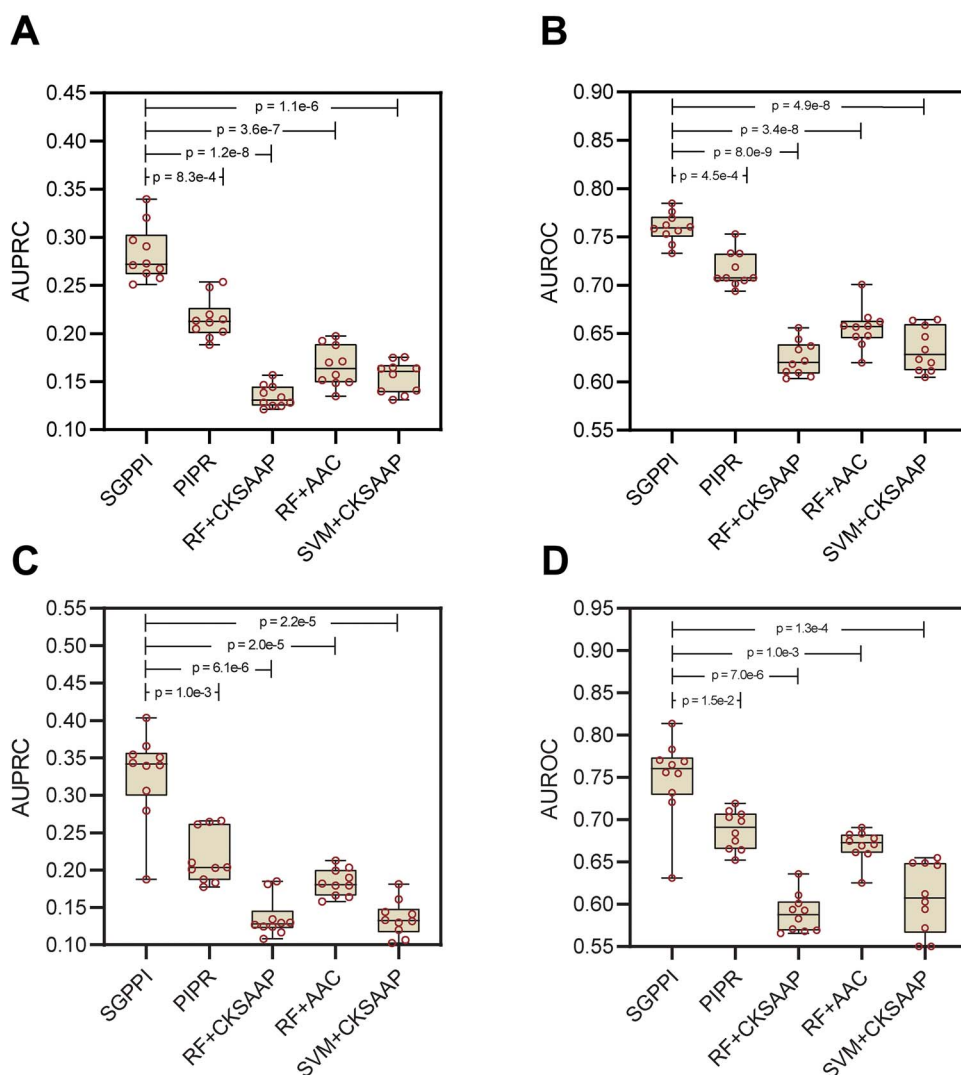


Figure 2. Prediction performance through 10-fold cross-validation on two alternative datasets. (A, B) Performance on the HuRI human dataset. (C, D) Performance on the Pan's dataset. Utilizing (A, C) AUPRC and (B, D) AUROC as prediction performance metrics, we observed that SGPPI clearly outperforms more traditional PPI prediction methods, including random forests (RF) and SVM, where the interacting sequences were represented by vectors of AAC and *k*-spaced amino acid pairs (CKSAAP). Paired *t*-test was used to determine the statistical significance in the performance of two models.

and B) and Pan's dataset (Figure 2C and D), the PR curves and ROC curves suggest that SGPPI generally outperformed the state-of-the-art DL model PIPR in these two more comprehensive datasets. By introducing more performance measurements, Tables 2 and 3 further quantify the performance of SGPPI in comparison to two traditional ML models and the PIPR method, highlighting the substantially better precision and F1-score of SGPPI.

Analyzing the informative characteristics of SGPPI

The construction of the residue contact map as of AlphaFold2 is the key step in representing the underlying protein structure. Two residues were considered to be in contact if the C_{α} atom distance of any two residues is less than a certain threshold. Specifically, we identified 10 Å as the best distance cut-off among several thresholds by cross-validation test on the Profppikernel human datasets (Figure 3A).

We also paid attention to the contribution of structural information to the performance of SGPPI. This structural information was encoded through a residue contact map derived from monomer structures as well as through residue-specific

structural features of each node in the residue contact map. By assessing the importance of per residue structural features by an ablation experiment, we first observed that removal of structural features substantially reduced the model's performance even if the original residue contact maps were totally retained (Figure 3B). Moreover, as intuitively expected, the residue contact map also substantially contributed to the structural representation of interacting proteins. In particular, we kept the edges in the contact map fixed to hold the global degree distribution constant but shuffled all the nodes' positions (i.e. degree preserving shuffling). To ensure sufficient randomization, the new position of each node after shuffling should be different from the original one. We retrained our model on the permuted residue contact map, and the result demonstrates that the permutation of residue contact map could lead to a sharp decline in the performance of SGPPI where the mean AUPRC decreased from 0.4225 to 0.3118 (Figure 3B).

Finally, as for the selection of nodes in the residue contact map, we provided an example of human non-classical major histocompatibility complex to illustrate the importance of the inclusion of protein surface patches. HLA class I histocompatibility antigen,

Table 2. Prediction performance metrics evaluated through 10-fold cross-validation on the HuRI dataset

Methods	AUROC ^a	AUPRC ^a	Precision ^{a,b}	Recall ^{a,b}	F1-score ^{a,b}
SGPPI	0.765	0.293	0.309	0.411	0.355
PIPR	0.720	0.225	0.235	0.417	0.314
RF + CKSAAP	0.624	0.135	0.122	0.358	0.184
RF + AAC	0.656	0.166	0.159	0.363	0.211
SVM + CKSAAP	0.639	0.155	0.147	0.357	0.206

^aAll the metrics shown in the table were the average value of 10-fold cross-validation. ^bWe used the threshold corresponding to the max F1-score to determine the values of TP, FP, TN and FN.

Table 3. Prediction performance metrics evaluated through 10-fold cross-validation on the Pan's dataset

Methods	AUROC ^a	AUPRC ^a	Precision ^{a,b}	Recall ^{a,b}	F1-score ^{a,b}
SGPPI	0.750	0.327	0.319	0.478	0.375
PIPR	0.688	0.216	0.225	0.475	0.302
RF + CKSAAP	0.624	0.135	0.122	0.358	0.187
RF + AAC	0.656	0.166	0.159	0.363	0.223
SVM + CKSAAP	0.639	0.155	0.147	0.357	0.210

^aAll the metrics shown in the table were the average value of 10-fold cross-validation. ^bWe used the threshold corresponding to the max F1-score to determine the values of TP, FP, TN and FN.

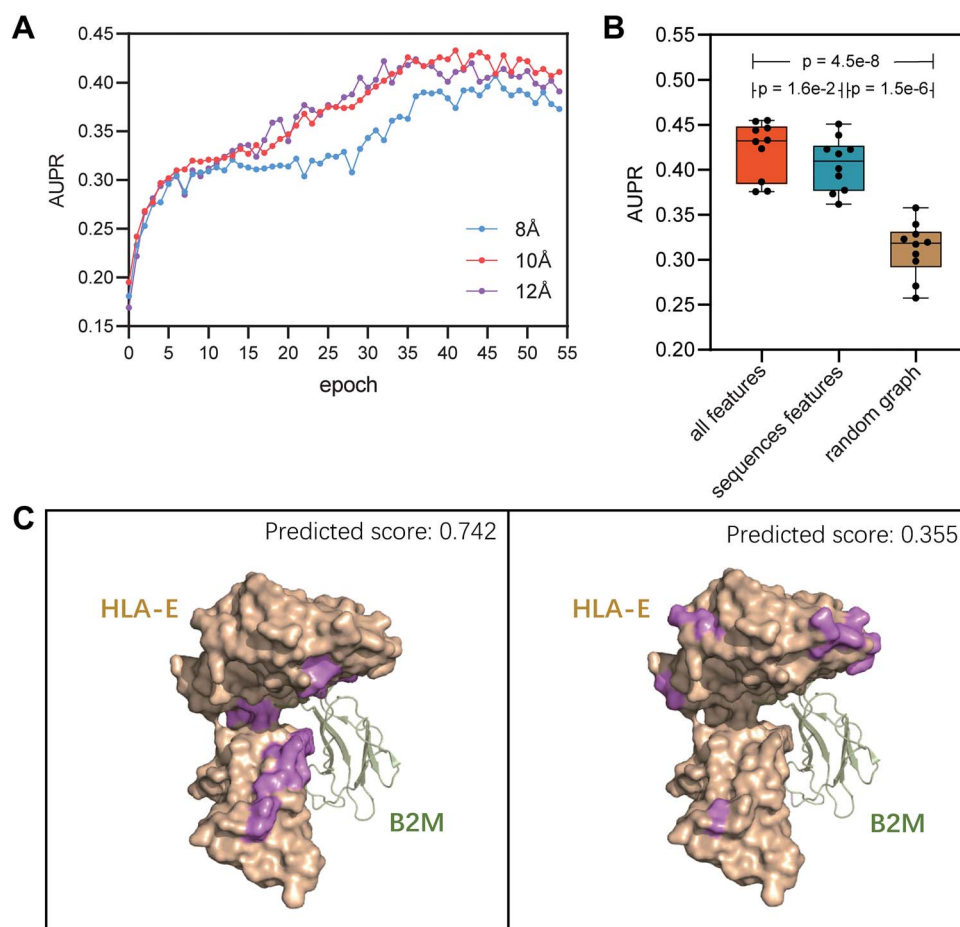


Figure 3. Analysis of informative structural features in SGPPI. **(A)** Performance comparison of the SGPPI model on the Profppikernel dataset using three different thresholds to construct the residue contact map. **(B)** Ablation experiments show the contribution of various structural features to the model performance of SGPPI. Paired t-test was used to determine the statistical significance in the performance of two models. **(C)** Indicating the importance of patch residues, we determined the interaction score between B2M and HLA-E with SGPPI using the actual (left panel) and randomly sampled patches on the protein surfaces (right panel). Notably, the prediction score considerably drops as a consequence of disrupting such residue patch information. The patch residues are highlighted on the protein surface representations.

alpha chain E (HLA-E) participates in the presentation of peptide antigens to the immune system with beta-2-microglobulin (B2M). We randomly selected the partial residues of HLA-E as protein

patches to build a factitious residue contact map. Compared with the original patches, the predicted interaction score based on random patches decreased significantly, which proved the

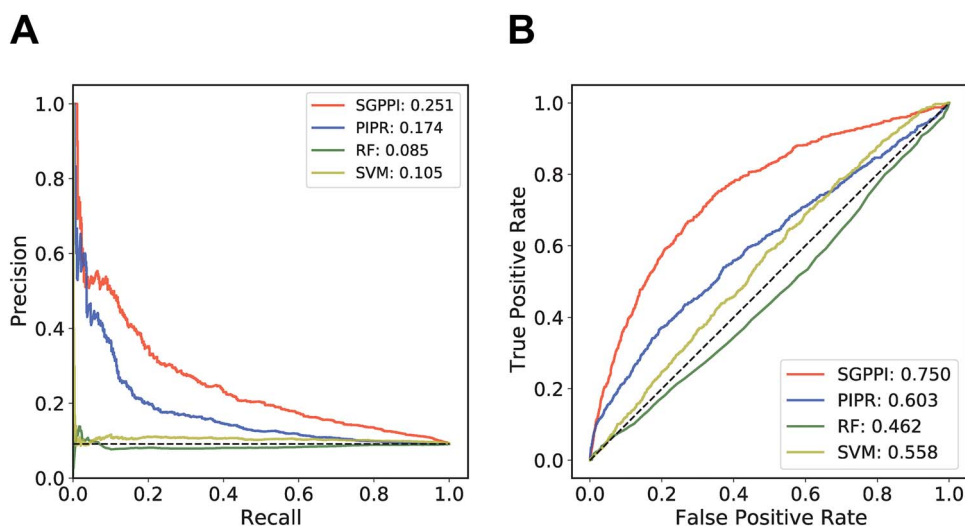


Figure 4. Performance of different prediction methods in the cross-species test. We trained several sequence-based methods and SGPPI on the human interaction data and tested such models for their ability to predict PPIs in yeast. In particular, we utilized RF and SVMs, where we represented interacting proteins through CKSAPP. PR and ROC analysis clearly indicated that the naive transfer of the SGPPI model was less sensitive to the underlying encoded structural information than the sequence information. The dashed lines in panels (A) and (B) show the performance of random prediction, and the corresponding AUPRC and AUROC are 0.091 and 0.500, respectively.

importance of local structural information on the protein surface patches in protein structural representation (Figure 3C). Together, a modest to high performance reduction can be observed with the perturbation of various informative structural features.

Cross-species prediction test

Protein sequence composition could differ substantially in different species, while the protein structures are relatively more conserved. As the prediction model of SGPPI largely relies more on structural than sequence information, we surmised that naively transferring a trained model to predict PPIs in a different species may be more reliable than sequenced-based methods. To check this hypothesis, in particular, we tested the cross-species predictive ability of SGPPI and other sequence-based approaches by training on the human HuRI dataset with a larger amount of non-representation-biased PPIs and predicted PPIs in yeast using the Profppikernel yeast dataset as a testing standard. As proteins from different species were used for training and testing, traditional sequence encoding-based ML methods, such as RF and SVM, that used CKSAPP encoding of protein sequences could hardly cope with such a difficult task (Figure 4 and Supplementary Table S1 available online at <http://bib.oxfordjournals.org/>). In turn, SGPPI performs on a par with human PPI prediction (Figures 2 and 4), suggesting that the encoded structural information is more robust and conducive to find PPIs across species.

Discussion and conclusion

Identification of PPIs is critical for understanding the functional mechanisms of proteins. The prediction of PPIs by computational methods has continuously been an important topic in the field of bioinformatics. However, traditional ML models are susceptible to the bias in the dataset. Specifically, these models are often overestimated when there are certain shared protein components between the training and test sets. Therefore, classical prediction scenarios are not suitable to evaluate the performance in predicting novel PPIs or in cross-species PPI predictions. In this work, we constructed the rigorous datasets where both negative and positive test protein pairs were allowed to be sequence-similar to

negative training protein pairs but were obligate to be sequence-dissimilar to any training PPIs. The prediction in such a rigorous scenario is much more meaningful for dealing with novel proteins predictions and cross-species predictions.

By using the rigorous benchmarking datasets, we introduced a GCN-based framework, SGPPI, to deal with the PPI prediction issue. With the advent of AlphaFold2, protein structures can be easily obtained, which provides more intuitive and robust information to predict PPIs. SGPPI first calculated the residue contact maps according to these protein structures. We regarded the residue contact map as an undirected graph with the residues as nodes and the residue contacts as edges. Instead of simply utilizing the commonly used sequence feature representations, we regard the protein as a collection of protein interface patches and integrated the global and local structural features of each residue in these patches. Besides, a comprehensive set of protein sequence and structural features have been considered, including residue conservation information, protein secondary structure types and global and local geometrical descriptors. By assigning the nodes in the residue graph with rich biological features, the information of protein deposited in the graph is further integrated and enriched. Further, GCN can effectively spread and update the information of the node features through the connection in the graph. More than the basic GCN structure, to deal with the pair input prediction problem, the DL model of SGPPI achieves coupled feature extraction of two input proteins through a Siamese-like GCN architecture and then predicts the interaction probability of the given protein pair through feature merging and full connection layers.

By comparison with previous sequence-based PPI methods under the challenging datasets without representation bias, SGPPI has exhibited strong robustness and a high precision. This result suggested that the introduction of structural information can indeed improve the performance of PPI prediction while considering sequence and evolutionary information. In particular, the contribution of protein patches to the prediction of PPIs has been demonstrated in ablation experiments (Figure 3B and C). The results of cross-species prediction also confirmed that the structural information in predicting PPIs can effectively deal with

novel proteins. It is also noteworthy that the absolute performance metrics of SGPPI (AUPRC or other metrics) are heavily dependent on the rigorous nature of the datasets using the stringent Profppikernel's strategy. If we used the original Pan's dataset preparation where Profppikernel's strategy was not applied, a very striking performance could be achieved (Supplementary Table S2 available online at <http://bib.oxfordjournals.org/>). But as stated above, for the evaluation for novel PPI prediction, rigorous scenario is deemed essential [18, 34] and performance improvement achieved by SGPPI on the rigorous datasets could be considered to be substantial. We believe SGPPI can provide novel approaches for the effective introduction of structural information and hope that the development of SGPPI will further promote the development of PPI prediction.

Key Points

- SGPPI is a structure-based DL framework for predicting PPIs using graph convolutional neural networks.
- SGPPI integrates both global and local features of structures and applies convolutions on residue contact maps to capture the characteristics of proteins.
- SGPPI achieved a competitive performance in rigorous benchmark datasets compared with state-of-the-art DL-based methods.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib/>.

Data availability

The data underlying this article are available in Figshare, at <https://dx.doi.org/10.6084/m9.figshare.20353353>. The source code is available in GitHub, at <https://github.com/emersON106/SGPPI>.

Funding

National Key Research and Development Program of China (2021YFF1201201 to Y.Z.); National Natural Science Foundation of China (32270703 and 31970645 to Z.Z.).

References

1. Bludau I, Aebersold R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat Rev Mol Cell Biol* 2020;**21**:327–40.
2. Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. *Chem Rev* 2016;**116**:4884–909.
3. Jubb H, Higuero AP, Winter A, et al. Structural biology and drug discovery for protein-protein interactions. *Trends Pharmacol Sci* 2012;**33**:241–8.
4. Scott DE, Bayly AR, Abell C, et al. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nat Rev Drug Discov* 2016;**15**:533–50.
5. von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;**417**:399–403.
6. Hu L, Wang X, Huang YA, et al. A survey on computational models for predicting protein-protein interactions. *Brief Bioinform* 2021;**22**:1–18.
7. Skrabanek L, Saini HK, Bader GD, et al. Computational prediction of protein-protein interactions. *Mol Biotechnol* 2008;**38**:1–17.
8. Bitbol AF, Dwyer RS, Colwell LJ, et al. Inferring interaction partners from protein sequences. *Proc Natl Acad Sci U S A* 2016;**113**:12180–5.
9. Kovacs IA, Luck K, Spirohn K, et al. Network-based prediction of protein interactions. *Nat Commun* 2019;**10**:1240.
10. Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 2012;**490**:556–60.
11. Martin S, Roe D, Faulon JL. Predicting protein-protein interactions using signature products. *Bioinformatics* 2005;**21**:218–26.
12. Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* 2007;**104**:4337–41.
13. Guo Y, Yu L, Wen Z, et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* 2008;**36**:3025–30.
14. Lian X, Yang S, Li H, et al. Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. *J Proteome Res* 2019;**18**:2195–205.
15. Xiao N, Cao DS, Zhu MF, et al. Protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 2015;**31**:1857–9.
16. Pitre S, Hooshyar M, Schoenrock A, et al. Short co-occurring polypeptide regions can predict global protein interaction maps. *Sci Rep* 2012;**2**:239.
17. Zahiri J, Yaghoubi O, Mohammad-Noori M, et al. Ppieve: protein-protein interaction prediction from pssm based evolutionary information. *Genomics* 2013;**102**:237–42.
18. Hamp T, Rost B. Evolutionary profiles improve protein-protein interaction prediction from sequence. *Bioinformatics* 2015;**31**:1945–50.
19. Jothi R, Kann MG, Przytycka TM. Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics* 2005;**21**:i241–50.
20. Zhang F, Song H, Zeng M, et al. Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics* 2019;**19**:e1900019.
21. Hu X, Feng C, Ling T, et al. Deep learning frameworks for protein-protein interaction prediction. *Comput Struct Biotechnol J* 2022;**20**:3223–33.
22. Du X, Sun S, Hu C, et al. Deepppi: boosting prediction of protein-protein interactions with deep neural networks. *J Chem Inf Model* 2017;**57**:1499–510.
23. Yang X, Yang S, Lian X, et al. Transfer learning via multi-scale convolutional neural layers for human-virus protein-protein interaction prediction. *Bioinformatics* 2021;**37**:4771–8.
24. Chen Z, Zhao P, Li C, et al. Ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res* 2021;**49**:e60.
25. Sledzieski S, Singh R, Cowen L, et al. D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst* 2021;**12**:969–982 e966.

26. Sun T, Zhou B, Lai L, et al. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 2017;**18**:277.
27. Chen M, Ju CJ, Zhou G, et al. Multifaceted protein-protein interaction prediction based on siamese residual rcnn. *Bioinformatics* 2019;**35**:i305–14.
28. Zhang QC, Petrey D, Garzon JI, et al. Preppi: a structure-informed database of protein-protein interactions. *Nucleic Acids Res* 2013;**41**:D828–33.
29. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
30. Sun M, Zhao S, Gilvary C, et al. Graph convolutional networks for computational drug development and discovery. *Brief Bioinform* 2020;**21**:919–35.
31. Torng W, Altman RB. Graph convolutional neural networks for predicting drug-target interactions. *J Chem Inf Model* 2019;**59**:4131–49.
32. Gligorijevic V, Renfrew PD, Kosciolk T, et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168.
33. Yuan Q, Chen J, Zhao H, et al. Structure-aware protein-protein interaction site prediction using deep graph convolutional network. *Bioinformatics* 2021;**38**:125–32.
34. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;**9**:1134–6.
35. Schaefer MH, Fontaine JF, Vinayagam A, et al. Hippie: integrating protein interaction networks with experiment based quality scores. *PLoS One* 2012;**7**:e31826.
36. Salwinski L, Miller CS, Smith AJ, et al. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:D449–51.
37. Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. *Nature* 2020;**580**:402–8.
38. Pan XY, Zhang YN, Shen HB. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J Proteome Res* 2010;**9**:4992–5001.
39. Yang F, Fan K, Song D, et al. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinform* 2020;**21**:323.
40. Petersen B, Petersen TN, Andersen P, et al. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 2009;**9**:51.
41. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol* 1998;**280**:1–9.
42. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science* 1995;**267**:383–6.
43. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 2007;**450**:1001–9.
44. Laine E, Carbone A. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLoS Comput Biol* 2015;**11**:e1004580.
45. Altschul SF, Madden TL, Schaffer AA, et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
46. Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* 2015;**43**:D364–8.
47. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**:2577–637.
48. Hashemifar S, Neyshabur B, Khan AA, et al. Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics* 2018;**34**:i802–10.