Briefings in Bioinformatics, 2025, **26(2)**, bbaf079
https://doi.org/10.1093/bib/bbaf079
Problem Solving Protocol

HSSPPI: hierarchical and spatial-sequential modeling for PPIs prediction

Yuguang Li[1], Zhen Tian[1,2], Xiaofei Nan[1], Shoutao Zhang[3,4], Qinglei Zhou[1], Shuai Lu[1,5,*]

[1]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, Henan, China
[2]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324003, Zhejiang, China
[3]School of Life Sciences, Zhengzhou University, Zhengzhou 450001, Henan, China
[4]Zhongyuan Intelligent Medical Laboratory, Zhengzhou 450001, Henan, China
[5]National Supercomputing Center in Zhengzhou, Zhengzhou University, Zhengzhou 450001, Henan, China
*Corresponding author. School of Computer and Artificial Intelligence and National Supercomputing Center in Zhengzhou, Zhengzhou University, Zhengzhou 450001, Henan, China. E-mail: ieslu@zzu.edu.cn

Abstract

Motivation: Protein–protein interactions play a fundamental role in biological systems. Accurate detection of protein–protein interaction sites (PPIs) remains a challenge. And, the methods of PPIs prediction based on biological experiments are expensive. Recently, a lot of computation-based methods have been developed and made great progress. However, current computational methods only focus on one form of protein, using only protein spatial conformation or primary sequence. And, the protein's natural hierarchical structure is ignored. **Results:** In this study, we propose a novel network architecture, HSSPPI, through **h**ierarchical and **s**patial-**s**equential modeling of protein for PPIs prediction. In this network, we represent protein as a hierarchical graph, in which a node in the protein is a residue (residue-level graph) and a node in the residue is an atom (atom-level graph). Moreover, we design a spatial-sequential block for capturing complex interaction relationships from spatial and sequential forms of protein. We evaluate HSSPPI on public benchmark datasets and the predicting results outperform the comparative models. This indicates the effectiveness of hierarchical protein modeling and also illustrates that HSSPPI has a strong feature extraction ability by considering spatial and sequential information simultaneously. **Availability and implementation:** The code of HSSPPI is available at https://github.com/biolushuai/Hierarchical-Spatial-Sequential-Modeling-of-Protein.

Keywords: protein–protein interaction sites; hierarchical graph; spatial form; sequential form; feature fuse

Introduction

Proteins are involved in nearly all biological processes and cellular functions necessary for life, such as gene expression, RNA transcription, DNA synthesis, and immune reaction [1]. Biological functions are performed through interactions and chemical reactions among proteins and proteins or other molecules [2, 3]. The detection of protein–protein interaction (PPI) and the recognition of protein-protein interaction sites (PPIs) are important research topics in proteomics [4, 5]. The accurate identification of PPIs is of great significance for the design of novel drugs [6, 7], the annotation of protein function [8], the understanding of molecular mechanism of diseases [9, 10], and construction of PPI networks [11, 12]. The determination of PPIs through traditional biological experiments is time-consuming and high-cost [13–15]. In recent years, with the rapid improvement of protein databases [16, 17] and the urgent demand for disease prevention, various computational PPIs prediction methods have been developed.

Existing computational PPIs prediction methods can be roughly classified into two types, sequence-based and structure-based methods, according to recent reviews [18–20]. Almost all those methods rely on machine learning or deep learning models. The traditional machine learning models like naive Bayes [21], support vector machine [22], random forests [23, 24], and XGBoost [25] have widely been used for predicting PPIs. Recently, many studies utilize deep learning models such as convolutional neural networks (CNN) [26], recurrent neural networks [27, 28], and graph neural networks (GNN) [29, 30] to predict PPIs. Moreover, inspired by the natural language processing (NLP) tools, some methods regard the protein sequence as the natural language which has the concept of forward and backward order, and introduce the bidirectional operation through bidirectional gated recurrent unit (Bi-GRU) and Bi-LSTM. With the advancement of language models and transfer learning, some protein language models are proposed for many bioinformatics tasks and outperform other methods at embedding protein sequences. Although the computational methods mentioned above have made great achievements, there still remain some limitations.

Firstly, those methods model proteins simply through one form: sequential or spatial. Sequence-based methods treat protein as a type of sequence data consisting of a set of residues like natural language. Structure-based methods treat protein as a kind of graph data by extracting nodes and edges. However, a protein performs functions by its special conformation

Received: September 23, 2024. **Revised:** February 10, 2025. **Accepted:** February 13, 2025

folding from its primary sequence. Therefore, residues binding or not are influenced by their spatial and sequential neighbors which should be analyzed from both protein sequence and structure. Models focus on sequential or spatial form couldn't take advantage of both data and only utilize one type of neighbors for making predicting. Some methods make a combination of features extracted from protein sequence and structure and then feed them into a predictive model [26]. Although this combination can achieve better performance, the prediction performance also relies on the selection of predictive models that model proteins in only one form.

Secondly, those methods ignore the natural hierarchical structure of proteins. Hierarchy information is widely used in a lot of bioinformatics tasks such as drug–drug interaction [31], drug-target affinity prediction [32], molecule property prediction [33], protein function prediction [34], and protein binding residues prediction [12]. Conformation information at different levels such as atom, residue, motif, protein, and interaction network are fused for capturing molecule or protein structure adequately. As for PPIs prediction, from a high-level view, interactions between proteins are driven by a few key residues. From a low-level view, the interactions between proteins are ultimately the interactions between atoms. Moreover, HIGH-PPI [12] utilizes structure information at a higher level: the PPI network. However, current computational methods for PPIs prediction don't consider the feature information at the atom level. Those methods usually extract features from protein graphs at the residue level and don't capture more detailed information such as the geometric arrangement of atoms from side chains of an amino acid residue. Therefore, the prediction performance is limited by the expressive power of the features used.

These limitations motivate us to develop a new approach to fully explore the protein features to predict the PPIs. Here, we propose a novel deep-learning architecture, HSSPPI, through modeling protein as a **h**ierarchical graph and considering **s**patial and **s**equential forms for PPIs prediction. In this study, HSSPPI takes protein as a hierarchical graph in which a set of atoms serve as a residue and the residues constitute a protein. More specifically, a residue-level graph is composed of residues in a protein, in which a residue node corresponds to an atom-level graph consisting of atoms. Moreover, we model protein through spatial and sequential forms and design a spatial-sequential (S-S) block containing a graph convolutional network (GCN) layer linked with a Bi-GRU layer. This block can capture sequential and spatial neighbor information of target residue for fully exploring protein structure and extracting long-distance interaction between residues and atoms. By providing the learned atom-level information to the residue-level feature, the fused feature embedding of the protein can be enriched so that the interaction relationship can be accurately learned. Therefore, HSSPPI can predict PPI sites more accurately. We test and evaluate HSSPPI on two public PPIs prediction tasks, and the results show that our proposed method achieves better performance. The main contributions of our work are summarized as follows:

- We propose a novel network architecture, HSSPPI, to predict PPIs through hierarchical and S-S modeling protein for fully capturing protein structure information.
- We represent protein as a hierarchical graph in which the residues are nodes in the protein and the atoms are the nodes in corresponding residues resulting in performance improvement.
- We design a S-S block consisting of a GCN layer linked with a Bi-GRU layer for extracting spatial and sequential neighbor information at the same time.
- HSSPPI outperforms the comparative methods on two public benchmark tasks which indicates the effectiveness of hierarchical and S-S modeling of protein for PPIs prediction.

Materials and methods

Benchmark datasets

In this study, we evaluate our proposed method HSSPPI on two public PPIs prediction tasks including the DeepPPISP task [26] and GraphPPIS task [35]. The specific information of both tasks is described as follows and the statistics information of datasets is provided in Table S1 in Supplementary Section 1.

DeepPPISP task

On the DeepPPISP task, we evaluate and compare HSSPPI with competing models on Train352 and Test70 datasets which are extracted by DeepPPISP. Three benchmark datasets including Dset_186 [36], Dset_72 [36],and PDBset_164 [37] are combined in the DeepPPISP task. Proteins in these datasets are downloaded from the RCSB PDB database and selected according to the criteria of resolution <3.0 Å and sequence homology <25%. In this study, an interaction site of protein is defined if the surface residue of a protein has a reduction of at least 1 Å squared in its absolute solvent accessibility area after the formation of a complex. Dset_186, Dset_72, and PDBset_164 contain 186, 72, and 164 proteins, respectively. To ensure that the training set and the test set are from an identical distribution, Zeng et al. fuse the three datasets into one resulting in a fused dataset consisting of 422 proteins [26]. Then, they split it into a training set containing 352 proteins (Train352) and a test set containing 70 proteins (Test70).

GraphPPIS task

On the GraphPPIS task, models are trained and evaluated on the same datasets as AGAT-PPIS [38] including one training set named Train335 along with four testing sets named Test60, Test287, TestB25, and TestUB25. Train335 and Test60 datasets are also developed from the Dset_186, Dset_72, and PDBset_164 datasets mentioned above. Unlike the DeepPPISP task, the GraphPPI task removes proteins with more than 25% sequence similarity and 90% of overlapping proteins through BLASTClust after combining the three baseline datasets. This operation results in 395 non-redundant proteins. And then, 335 proteins are randomly selected as the training set (Train335) leaving 60 proteins performed as the testing set (Test60). Test287, TestB25, and TestUB25 are obtained by deleting protein sequences that could not align with their corresponding structures from Test_315 and UBtest_31 which are used for testing GraphPPIS [35].In this work, Train335 and Test60 are applied as benchmark datasets on the GraphPPIS task and the remaining testing sets are used for evaluating the generalization performance of models. It should be noted that proteins in the TestUB25 are corresponding unbound structures in the TestB25 set which contain the same proteins and are all derived from Test60. The results on TestB25 and TestUB25 can reflect the robustness of PPIs prediction methods and evaluate the impact of conformational changes on the method performance.

Table S1 shows the number of interaction and non-interaction sites on both DeepPPISP task and GraphPPIS task. The datasets

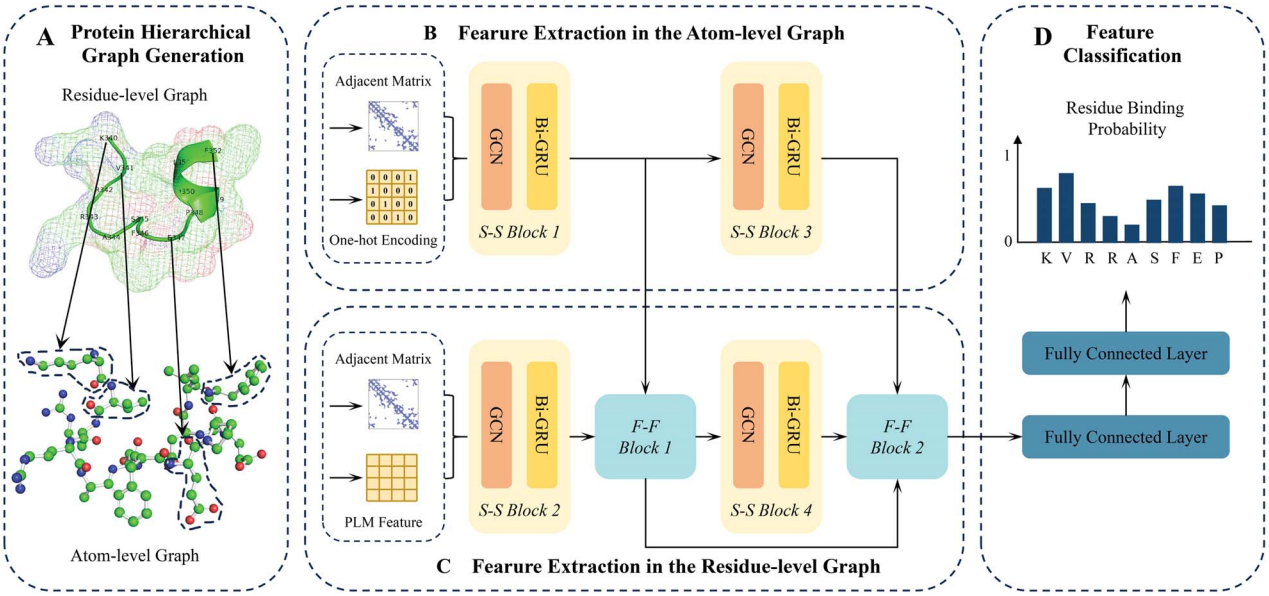

Figure 1. The detailed architecture of HSSPPI. (A) The part of protein hierarchical graph generation in which the nodes of the protein are residues and the nodes of the residues are corresponding atoms. Therefore, the hierarchical graph of a protein is a residue-level graph and an atom-level graph. (B) The part of feature extraction in the residue-level graph consists of two S-S blocks and two F-F blocks. A GCN layer and a Bi-GRU layer constitute the S-S block used for capturing neighbor information comprehensively from spatial and sequential forms of the protein. The F-F blocks are applied for fusing features from the residue-level graph and the atom-level graph. (C) The part of feature extraction in the atom-level graph only consists of two S-S blocks. (D) The part of feature classification in which two fully connected layers output the probability of a residue interacting or not.

are unbalanced and each set represents a considerable imbalance between the interaction and non-interaction sites. Therefore, evaluation metrics sensitive to unbalanced data are taken as more important.

Architecture of HSSPPI

As shown in Fig. 1, the main architecture of HSSPPI consists of four parts, which are the protein hierarchical graph generation, the feature extraction in the residue-level network, the feature extraction in the atom-level graph, and the feature classification. We'll explain each component in detail below.

Protein hierarchical graph generation

It can be seen from Fig. 1A that we represent protein as a hierarchical graph in which a node in a protein is a residue (residue-level graph) and a node in a residue is an atom (atom graph). The edges in the residue-level and the atom-level graph can be calculated by the angles and distances between residues and atoms. Usually, the information at the residue-level can be directly used for predicting whether a residue interacts with others or not. Considering the protein structure is naturally hierarchical and residue-level prediction is always made through a few atoms in the target residue which are inadequate and limited. Therefore, we utilize the atom-level graph containing all atoms helping to make residue-level judgments in this study.

Usually, a protein G is defined as the residue-level graph which contains a set of nodes and edges: $G_r = \{V_r, E_r\}$ and $V_r = \{r_1, r_2, \ldots, r_i, \ldots, r_{n-1}, r_n\} \in R^{N*F_r}$, where N is the number of residues in the protein G and F_r is the feature dim of the residue, and E_r are the set of edges between all residues. Here, the distance between two residues in G_r is calculated by the mean distance of heavy atoms of each residue. Therefore, the existence of an edge is determined by the distance and we test a set of distance thresholds for finding the best performance of HSSPPI.

Similarly, the atom-level graph can be shown as $G_a = \{V_a, E_a\}$ and $V_a = \{a_1, a_2, \ldots, a_j, \ldots, a_{m-1}, a_m\} \in R^{M*F_a}$, where M is the number of atoms in the protein G and F_a is the feature dim of the atom, and E^a are the set of edges between all atoms. The edges in the atom-level graph are also determined by the distance threshold between all heavy atoms in the protein.

Through this operation, the feature representation of the protein G can be extracted from the residue-level (G_r) and atom-level (G_a), which can refine the dimension of the manually extracted feature and enrich the feature representation.

Feature extraction in the atom-level graph

The feature extraction part in the atom-level graph in HSSPPI contains two S-S blocks (S-S Block 1 and S-S Block 3). As shown in Fig. 1B, we get the atom feature representation and adjacent matrix of the atom-level graph after the protein hierarchical generation. In this study, the S-S block is designed to extract the sequential and spatial neighbor features through the linked GCN layer and Bi-GRU layer.

Atom feature representation As for the atom-level graph $G_a = \{V_a, E_a\}$, we utilize one-hot encoding of the atom types to obtain the atom feature embedding. According to a related study [39], 37 kinds of atoms are labeled for 20 types of residues. As defined above, the nodes in the atom-level graph are $V_a = \{a_1, a_2, \ldots, a_j, \ldots a_m\} \in R^{M*F_a}$, where a_j represents the *jth* atom and M is the number of atoms in the protein. After one-hot encoding of the atoms, the protein is represented as a feature matrix which has a dimension of $M * 37$, i.e.,$F_a = 37$ in this study.

Edges E_a stands for the connection between atoms which is determined by the distance threshold. In this study, the connection relationships are contained in an adjacency matrix $A_a \in R^{(M*M)}$. If the distance between two atoms is less than the threshold, there is an edge between them which is labeled as 1 in the adjacency matrix. Otherwise, it is labeled as 0. That is to say, the *ith* row and *jth* column element in the adjacency matrix is 1 standing

for that there is an edge between the *ith* and *jth* atom which is regarded as connected. Here, we select the distance threshold from a set of {1.3 , 1.5 , 1.7, 1.9, 2.1, 2.3 , 2.5 Å}.

spatial-sequential block For capturing the spatial and sequential dependency from protein data at the same time, we propose a S-S block based in the GCN and a Bi-GRU. GCN can learn the representation of a local neighborhood around each node in a graph. In this block, we represent a 3D structure of a protein as a graph in which each atom is a node and edges represent the connections between nodes. Given a protein G_a with M residues, the atom-level graph is represented by the node feature matrix V_a and the adjacency matrix A_a. The graph convolutional operation is computed as follows:

$$V_a^{l+1} = \sigma(D_a^{-1}A_aV_a^lW_a^l + b_a^l) \tag{1}$$

where D_a is the diagonal degree matrix of the adjacency matrix A_a, the normalization of A_a is achieved through $D_a^{-1}A_a$. V_a^l is the node feature of layer l, W_a^l is weight matrix of layer l, b_a^l is a biases vector of layer l, and σ is a nonlinear activation function. Here, we use $f(V_a, A_a)$ to represent the output of GCN. The GCN model is used to learn and integrate the spatial neighbor information of the target residue.

Similar to the semantic and syntactic rules in NLP, proteins have special restrictions on their sequences to perform certain functions. To further explore the sequential neighbor information, a Bi-GRU [40] model is adopted to capture the contextual information. Bi-GRU consists of a forward neural network and a backward neural network, which capture data information from both directions. GRU has two gates and defines the update gate as u_t and the reset gate as r_t.

$$u_t = \sigma_u(W_u * [f(V_a, A_a), h_{t-1}] + b_u) \tag{2}$$

$$u_r = \sigma_r(W_r * [f(V_a, A_a), h_{t-1}] + b_r) \tag{3}$$

where t is the time step, W_u and W_r are weight matrixes, V_t is the input for current time t, h_{t-1} is previous hidden states, b is bias, σ a nonlinear activation function.

The hidden layer $\widetilde{h_t}$ is the new internal memory, which is calculated as follows:

$$\widetilde{h_t} = tanh(W_h * [f(V_a, A_a), r_t * h_{t-1}]) \tag{4}$$

where r_t is the data obtained through the reset gate. Then the new memory state of the output h_t is calculated as follows:

$$h_t = u_t * h_{t-1} + (1 - u_t) * \widetilde{h_t} \tag{5}$$

The Bi-GRU applies two GRUs learning forward and backward sequential information of the input protein. And, the output of Bi-GRU is the concatenation of the forward result $\overrightarrow{h_t}$ and the backward result $\overleftarrow{h_t}$, which is calculated as follows:

$$H_a = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{6}$$

After the above calculation, the sequential and spatial information of protein can be fully utilized.

Feature extraction in the residue-level graph

As shown in Fig. 1C, the feature extraction part in the residue-level graph in HSSPPI includes two S-S blocks (S-S Block 2 and S-S Block 4) and two feature-fusion (F-F) blocks (F-F Block 1 and F-F Block 2). We get the residue feature representation and adjacent matrix of the residue-level graph after the protein hierarchical generation. In this study, the S-S blocks are designed to extract the sequential and spatial neighbor features through the linked GCN layer and Bi-GRU layer. The F-F blocks are used for fusing features from the atom-level and the residue-level.

Residue feature representation As for the residue-level graph $G_r = \{V_r, E_r\}$, we employ a protein language model [41], ProtT5, to obtain sequence embedding of the protein. As previous definition $V_r = \{r_1, r_2, \ldots, r_i, \ldots r_n\} \in R^{N*F_r}$, r_i stands for the *ith* residue and N is the number of residues in the protein. As suggested by [42], we utilize the residue feature obtained by ProtT5 which performs better than the concatenated feature and other protein language models on PPIs prediction. After the embedding generation of ProtT5, the protein sequence is represented as a feature matrix and its dimension is $N * 1024$ in which a 1024-dimensional vector stands for a residue at the same position, i.e. $F_r = 1024$ in this study.

We use the distance between two residues to extract the edges E_r which indicates the connections between the two residues. Here, we utilize an adjacency matrix $A_r \in R^{(N*N)}$ representing the connections between all residues. It is worth emphasizing that the distance is calculated by taking the average of the distances between their heavy atoms. If the distance between two residues is less than the threshold, there is an edge between them which is labeled as 1 in the adjacency matrix. Otherwise, it is labeled as 0. That is to say, the *ith* row and *jth* column element in the adjacency matrix is 1 standing for that there is an edge between the *ith* and *jth* residue which is regarded as connected. In this study, we select the distance threshold from a set of {4.0 Å, 4.5 Å, 5.0 Å, 5.5 Å, 6.0 Å, 6.5 Å, 7.0 Å}.

spatial-sequential block As shown in Fig. 1C, the S-S blocks for feature extraction in the residue-level graph share the same architecture as the S-S block for the atom-level graph. As described earlier, the input feature matrix in this block is $G_r = \{V_r, E_r\}$, where $V_r \in R^{N*F_r}$ which is different from the input embedding in the atom-level graph, which is $G_a = \{V_a, E_a\}$, where $V_a \in R^{M*F_a}$. To distinguish, we label S-S blocks with different numbers (S-S blocks are labeled 1 and 3 in the atom-level graph, and 2 and 4 in the residue-level graph). Besides, the main difference is that the output feature of S-S block in the atom-level graph is combined with the output feature in the residue-level graph. Therefore, the feature extraction operation in the residue-level graph will be affected by the operation in the atom-level graph which considers the natural hierarchical structure of the protein for improving PPIs prediction. And, this improvement is obtained through the F-F block that we will describe in detail in the following subsection.

feature-fusion block In order to realize the F-F between the atom-level graph and the residue-level graph, we propose a F-F block. There are two F-F blocks (F-F block 1 and 2) of the feature extraction part in the residue-level graph. F-F Block 1 is used to fuse the output features of S-S Block 1 and S-S Block 2. F-F Block 2 is used to fuse the output features of F-F Block 1, S-S block 3 and S-S block 4. First, we construct a map function to form the combined features, in which if the atom i belongs to the residue j, we concatenate the feature of atom i and residue j. The map matrix is defined as follows:

$$map_{ij} = \begin{cases} 1 & \text{if } a_i \in r_j, \\ 0 & \text{else.} \end{cases} \tag{7}$$

Then, the F-F Block 1 takes in the output feature matrix $H_{a_1} \in R^{M*F_{a_1}}$ of S-S Block 1 in the atom-level graph, and the output

feature matrix $H_{r_1} \in R^{N*F_{r_1}}$ of S-S Block 2 in the residue-level graph, where F_{a_1} and F_{r_1} are the dimensions of the hidden matrixes in the two graphs. The F-F operation can be calculated as follows:

$$H_{a_1}^{map} = map(H_{a_1}) \tag{8}$$

$$H_{out_1} = H_{r_1} \oplus H_{a_1}^{map} \tag{9}$$

where $H_{a_1}^{map} \in R^{N*F_{r_1}}$ and the output feature matrix of the F-F Block 1 $H_{out_1} \in R^{N*(2*F_{r_1})}$. And so on, the output feature matrix of the F-F Block 2 can be shown as $H_{out_2} \in R^{N*(2*F_{r_2})}$, which takes $H_{a_2} \in R^{M*F_{a_2}}$ of S-S Block 2 and $H_{r_1} \in R^{N*F_{r_1}}$ of S-S Block 4 as input.

Feature classification

We apply the fused feature of two graphs to make the PPIs prediction. To capture more information on features derived from different levels, we utilize a skip-connection operation to combine these features. Then, the output of the skip-connection operation is concentrated and fed into the networks in the feature classification part. As shown in Fig. 1D, this part consists of two fully connected layers using the *Relu* activation function. The process for the feature classification can be shown as follows:

$$H_{skip_1} = H_{out_1} S_1, H_{out_1} \in R^{N*(2*F_{r_1})} \tag{10}$$

$$H_{skip_2} = H_{out_2} S_2, H_{out_2} \in R^{N*(2*F_{r_2})} \tag{11}$$

$$H_{o_1} = \sigma_{o_1}(W_{o_1} * (H_{skip_1} \oplus H_{skip_2}) + b_{o_1}) \tag{12}$$

$$H_{out} = \sigma_{o_2}(W_{o_2} * H_{o_1} + b_{o_2}) \tag{13}$$

where $S_1 \in R^{(2*F_{r_1})*F_{o_1}}$, $S_2 \in R^{(2*F_{r_2})*F_{o_2}}$, W_{o_1} and W_{o_2} are the weight matrixes, b_{o_1} and b_{o_2} are the bias, and H_{out} is the final output which contains the predicted binding probability of each residue.

Implementation

We implement HSSPPI using PyTorch. The loss function is weighted cross-entropy loss and the optimization is Adaptive Momentum (Adam). We select the learning rate from (0.1, 0.01, 0.001, 0.0001). We choose the dropout rate from (0.2, 0.5, 0.8). Finally, we take the learning rate of 0.1, and the dropout rate of 0.2. 20% of the training set is randomly selected as an independent validation set to tune parameters. For each combination, networks are trained until the performance on the validation set stops improving or for a maximum of 200 epochs. The training time of each epoch varies roughly from 20 to 30 min depending on the model depth, using a single NVIDIA RTX4090 GPU.

Results

Comparison with competing methods

Performance evaluation on the DeepPPISP task

On the DeepPPISP task, we compared our proposed HSSPPI with eight state-of-the-art methods, which are SPPIDER [43], SCRIBER [44], DeepPPISP [26], Attention-CNN [45] DELPHI [46], HN-PPIS [30], EGRET [29], and EnsemPPIS [47]. Among these methods, SPPIDER, HN-PPIS, DeepPPISP, and EGRET use structural features of the protein, while other methods are based on protein sequence features. To make a fair comparison, the results shown in this section are generated through training and testing models using the same dataset. Table 1 shows the results of HSSPPI compared with the other models. The results shown in Table 1 are all tested on Test70 dataset of the DeepPPISP task. Some of the results shown in Table 1 are obtained from the web server or by rerunning the source code, and some are derived directly from the corresponding published studies. The evaluation metrics have been displayed in Supplementary Section 2.

From Table 1 we can see that HSSPPI has a significant improvement on F1, MCC, and AUPRC, which are important evaluation metrics for imbalanced data such as PPIs prediction task [26]. Especially, HSSPPI achieves a 50.23% improvement on F1, 15.16% improvement on MCC, and 7.16% improvement in AUPRC when compared to the state-of-the-art method EnsemPPIS. It should be noted that the progress in F1 value is quite remarkable. In addition, HSSPPI achieves 68% improvement over EnsemPPIS on Precision, which indicates that our proposed model effectively controls the false-positive rate. HSSPPI achieves a 30.83% improvement in Recall to EnsemPPIS, which reflects that our model has a good ability to identify positive cases. However, HSSPPI obtains significant improvements on most evaluation metrics except ACC, which is mainly due to the imbalance of datasets. Compared with the structure-based methods EGRET, HSSPPI performs better on most metrics and demonstrates its effectiveness by considering the protein's natural hierarchical structure and modeling protein through spatial and sequential forms. The excellent performance of HSSPPI also indicates that the deep-learning model can explore the potential connection between protein sequence and structure, thereby improving the performance of protein binding sites prediction.

Performance evaluation on the GraphPPIS task

We also evaluate HSSPPI on the GraphPPIS task, and the results are shown in Table 2. Compared with the DeepPPISP task, the other 11 additional competing methods are tested and compared on the Test60 dataset of the GraphPPIS task, PSIVER [21], ProNA2020 [48], DLPred [49], MaSIF-site [50], GraphPPIS [35], RGN [51], ProB-site [52], AGAT-PPIS [38], DeepProSite [53], GHGPR-PPIS [54], and GACT-PPIS [55]. All results listed in Table 2 are obtained from the original published studies.

As shown in Table 2, it can be found that HSSPPI achieves the best results on Precision, Recall, and F1. And, it should be noted that HSSPPI performs better than all sequence-based methods and achieves better results than some structure-based methods. Specifically, compared with the best existing method GACT-PPIS, HSSPPI achieves 9.9, 1.1, and 7% improvement on Precision, Recall, and F1, respectively. And, HSSPPI only performs worse than the four methods on ACC. Compared with GHGPR-PPIS which has the highest MCC, HSSPPI obtains significant improvement on AUROC. Although HSSPPI is inferior to GHGPR-PPIS and GACT-PPIS in some indicators on the GraphPPIS task, it is more in line with the development trend in this post-AlphaFold era, because it employs atomic characteristic information and shows generalization ability on the other independent test sets as shown in Tables 3, 4, and 5.

Table 3 shows the model performance on the Test287 which consists of protein complexes solved from January 2014 to May 2021 and stored in the Protein Data Bank (PDB) [35, 38]. The results show that HSSPPI achieves the best result across the critical evaluation metrics of F1. As shown in Tables 4 and 5, to illustrate the generalization ability of HSSPPI, we run experiments on other independent testing sets, TestB25 and TestUB25. TestB25 consists of protein chains from bound protein complex structures, while TestUB25 contains protein chains from bound protein complex structures. The combined comparison of TestB25 and TestUB25 reflects the generalization performance of models and the ability

Table 1. Performance comparison with HSSPPI and other state-of-the-art methods on the Test70 of the DeepPPISP task

Methods	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
SPPIDER[a,1]	0.667	0.240	0.315	0.273	0.063	0.510	0.235
SCRIBER[a,1]	0.616	0.274	0.569	0.370	0.159	0.635	0.307
DeepPPISP[b,2]	0.655	0.303	0.577	0.397	0.206	0.671	0.320
Attention-CNN[b,2]	0.657	0.313	0.611	0.414	0.229	NA	0.359
DELPHI[c,1]	0.667	0.320	0.604	0.418	0.236	0.690	NA
HN-PPISP[b,2]	0.667	0.324	0.632	0.427	0.244	NA	0.360
EGRET[b,2]	0.715	0.358	0.516	0.438	0.270	0.719	0.405
EnsemPPIS[b,1]	**0.732**	0.375	0.532	0.440	0.277	0.719	0.405
HSSPPI	0.690	**0.630**	**0.696**	**0.661**	**0.319**	**0.759**	**0.434**

Notes: [a]Results obtained by utilizing the web server. [b]Results reported from corresponding studies. [c]Results obtained by reproducing the source code. [1]Methods use protein sequences as input. [2]Methods take protein structural information as input. The best results are shown in bold. NA means not available.

Table 2. Performance comparison with HSSPPI and other competing methods on the Test60 of the GraphPPIS task

Methods	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
PSIVER[a,1]	0.561	0.188	0.534	0.278	0.074	0.573	0.190
ProNA2020[a,1]	0.738	0.275	0.402	0.326	0.176	NA	NA
SCRIBER[a,1]	0.667	0.253	0.568	0.350	0.193	0.665	0.278
DLPred[a,1]	0.682	0.264	0.565	0.360	0.208	0.677	0.294
DELPHI[a,1]	0.697	0.276	0.568	0.372	0.225	0.699	0.319
DeepPPISP[a,2]	0.657	0.243	0.539	0.335	0.167	0.653	0.276
SPPIDER[a,1]	0.752	0.331	0.557	0.415	0.285	0.755	0.373
EnsemPPIS[b,1]	0.716	NA	NA	0.401	0.266	0.726	0.352
MaSIF-site[a,2]	0.780	0.370	0.561	0.446	0.326	0.775	0.439
GraphPPIS[a,2]	0.776	0.368	0.584	0.451	0.333	0.786	0.429
RGN[a,2]	0.785	0.382	0.587	0.463	0.349	0.791	0.441
ProB-site[b,2]	0.799	0.407	0.612	0.517	0.368	0.844	0.467
DeepProSite[b,2]	0.842	0.443	0.501	0.470	0.379	0.813	0.490
AGAT-PPIS[b,2]	0.856	0.539	0.603	0.569	0.484	0.867	0.574
GHGPR-PPIS[b,2]	**0.860**	0.551	0.620	0.583	**0.501**	0.596	NA
GACT-PPIS[b,2]	0.854	0.530	0.646	0.583	0.499	**0.876**	**0.595**
HSSPPI	0.811	**0.650**	**0.657**	**0.653**	0.307	0.756	0.434

Notes: [a]Results obtained from RGN. [b]Results reported from corresponding studies. [1]Methods use protein sequences as input. [2]Methods take protein structural information as input. The best results are shown in bold. NA means not available.

Table 3. Performance comparison with HSSPPI and other competing methods on the Test287 of the GraphPPIS task

Methods	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
GraphPPIS[a,2]	–	–	–	0.349	0.432	–	–
AGAT-PPIS[a,2]	**0.859**	0.501	0.633	0.559	0.481	**0.874**	**0.572**
GHGPR-PPIS[b,2]	–	–	–	0.564	**0.486**	–	–
HSSPPI	0.768	**0.612**	**0.667**	**0.638**	0.273	0.743	0.323

Notes: [a]Results obtained from AGAT-PPIS. [b]Results reported from corresponding studies. [2]Methods take protein structural information as input. The best results are shown in bold. - means not provided.

Table 4. Performance comparison with HSSPPI and other competing methods on the TestB25 of the GraphPPIS task

Methods	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
GraphPPIS[a,2]	–	–	–	0.352	0.394	–	–
AGAT-PPIS[a,2]	**0.878**	0.528	0.601	0.554	**0.485**	**0.880**	**0.583**
GACT-PPIS[b,2]	0.839	0.491	0.636	0.554	0.463	0.863	0.568
HSSPPI	0.863	**0.681**	**0.657**	**0.669**	0.338	0.800	0.392

Notes: [a]Results obtained from AGAT-PPIS. [b]Results reported from corresponding studies. [2]Methods take protein structural information as input. The best results are shown in bold. - means not provided.

to capture the impact of conformational changes. The results state that HSSPPI outperforms GraphPPIS, AGAT-PPIS, GHGPR-PPIS, and GACT-PPIS on Precision, Recall, and F1. Meanwhile, as the comparison results on the TestUB25 show, HSSPPI obtains the highest ACC and is only lower than GACT-PPIS on AUROC and AUPRC. It is worth mentioning that TestUB25 comprises unbound protein structures. The better performance shows that HSSPPI captures more specific information on conformational changes by representing protein as a hierarchical graph fused atom-level and residue-level information. In conclusion, the results shown in Tables 3, 4, and 5 indicate that HSSPPI has much better adaptability and generality.

Table 5. Performance comparison with HSSPPI and other competing methods on the TestUB25 of the GraphPPIS task

Methods	ACC	Precision	Recall	F1	MCC	AUROC	AUPRC
GraphPPIS[b,2]	0.843	0.376	0.461	0.414	0.327	0.780	0.365
AGAT-PPIS[b,2]	0.837	0.375	0.534	0.441	0.356	0.796	0.368
GACT-PPIS[b,2]	0.852	0.409	0.522	0.459	**0.378**	**0.813**	**0.397**
GHGPR-PPIS[b,2]	0.837	0.375	0.534	0.441	0.356	0.796	0.368
HSSPPI	**0.863**	**0.681**	**0.657**	**0.669**	0.338	0.800	0.392

Notes: [b]Results reported from corresponding studies. [2]Methods take protein structural information as input. The best results are shown in bold.

Comparison on different model architectures

To identify the necessity of each part in HSSPPI, the following ablation experiments are implemented on the Test70 dataset of the DeepPPISP task in this section. As shown in Fig. 2A, four experiments are performed to demonstrate the necessity of each part in HSSPPI. In the first experiment (Only-Atom), only the atom-level graph of protein is used and the model architecture consists of two S-S blocks and one F-F block. In the second experiment (Only-Residue), only the residue-level graph of protein is used and the model architecture is comprised of two S-S blocks and one F-F block. As for the third experiment (Fusion-Once), features from the atom-level graph and the residue-level graph are fused only once, i.e. one S-S block is used for feature extraction and then one F-F block is used for F-F. As for the fourth experiment (Fusion-Thrice), features from two graphs are fused thrice, i.e. three S-S blocks are utilized for feature extraction in each graph and the output of each S-S block will be fused in an F-F block. It should be noted that our proposed HSSPPI adopts two graphs and features from them are fused twice.

The results of the four experiments are shown in Fig. 2B. Through the Only-Atom and Only-Residue experiments, it can be seen that the feature extracted from the residue-level graph outperforms the feature extracted from the atom-level graph. As the PPIs prediction task is residue-level, the residue-level graph provides more specific information. It can be observed that the results of Fusion-Once are worse than the results of Only-Residue which may be caused by the information from the atom-level graph introducing noise. However, our proposed HSSPPI outperforms the results of Only-Residue which indicates that the atom-level graph provides useful information through more times feature extraction and fusion. Whereas, too many fusions can cause performance reduction as the problem of excess smoothness from graph convolution operation just like the results of Fusion-Thrice show.

Comparison on different distance thresholds

As mentioned in the feature representation section, we use an adjacency matrix to characterize the edge features which show the connections between atoms and residues. And, if the distance between two atoms or residues is less than a threshold, there is an edge between them, and the corresponding element in the adjacency matrix is 1. To find the distance threshold that can make HSPPISP achieve the best performance, we set seven residue distance thresholds and seven atom distance thresholds. The residue distance threshold is selected from 4.0 Å, 4.5 Å, 5.0 Å, 5.5 Å, 6 Å, 6.5 Å, and 7Å. The atom distance thresholds are chosen from 1.3 Å, 1.5 Å,1.7 Å, 1.9 Å,2.1 Å, 2.3 Å, and 2.5 Å. They are then cross-combined to form 49 sets of experimental data and verified on the Test70 dataset of the DeepPPISP task.

Because AUPRC is a threshold-independent metric and more sensitive than AUROC on imbalanced datasets, we take AUPRC to select the best combination of distance thresholds. As shown in Fig. 2C, the best performance of HSSPPI is achieved when the residue distance threshold is 5.5 Å and the atom distance threshold is 2.3 Å. Therefore, our proposed method HSSPPI in this study utilizes the two thresholds. And, it should be noted that all the experimental results of HSSPPI are based on these distance thresholds.

Comparison on different GNN in S-S block

To learn spatial and sequential information of proteins at the same time, GCN and Bi-GRU are applied in the S-S block. For illustrating the effect of GNN model selection, we build HSSPPI using six different GNN architectures including SAGEConv [56], ChebConv [57], GATConv [58], GATv2Conv [59], GCNConv [60], and GraphConv [61] in S-S block. It should be mentioned that the other parts of HSSPPI share the same feature representation and model structure. All these models introduce innovations in GNNs that are suitable for different tasks and datasets. ChebConv [57] is based on Chebyshev polynomial filtering and has low time complexity. GATConv [58] and GATv2Conv [59] add different attention mechanisms that can capture complexity between nodes better. The nodes in GCNConv [60] share convolution weights. GraphConv [61] can customize the size and shape of the convolution kernel based on specific tasks and data sets. SAGEConv [56] is suitable for large-scale graph-structured data and reduces computational complexity through a sampling mechanism.

Figure 2D describes the comparison of HSSPPI results using different GNN architectures on the DeepPPIS dataset. From the results, we see that HSSPPI achieves the best performance using GraphConv. Therefore, the S-S block used in our proposed HSSPPI consists of a GCN layer implemented by SAGEConv. In this section, we run all experiments on the Test70 dataset of the DeepPPISP task.

Case study

To more directly show the advantage of HSSPPI on the PPIs prediction of individual proteins. We select a protein (PDB ID: 1B6C, Chain: A) in the Test70 dataset for visualization. Figure 3 shows the interacting sites of 1B6C_A obtained by the experiment and predicted by HSSPPI, Ensemble, and Seq-Insite, respectively. We use PyMol [62] to visualize proteins, where the interacting sites are in blue and the non-interacting sites are in gray. It can be seen from Fig. 3 that HSSPPI has a more accurate prediction than other methods. Especially in the display after the conformation of 1B6C_A is rotated 180°, the predicted interact sites of HSSPPI are more accurate, which indicates that HSSPPI significantly reduces the false-positive rate, and the MCC value achieves 15.5% improvement than EnsemPPIS.

Discussion

Our research proposes a novel deep-learning architecture named HSPPISP that aims to improve the ability of feature representation

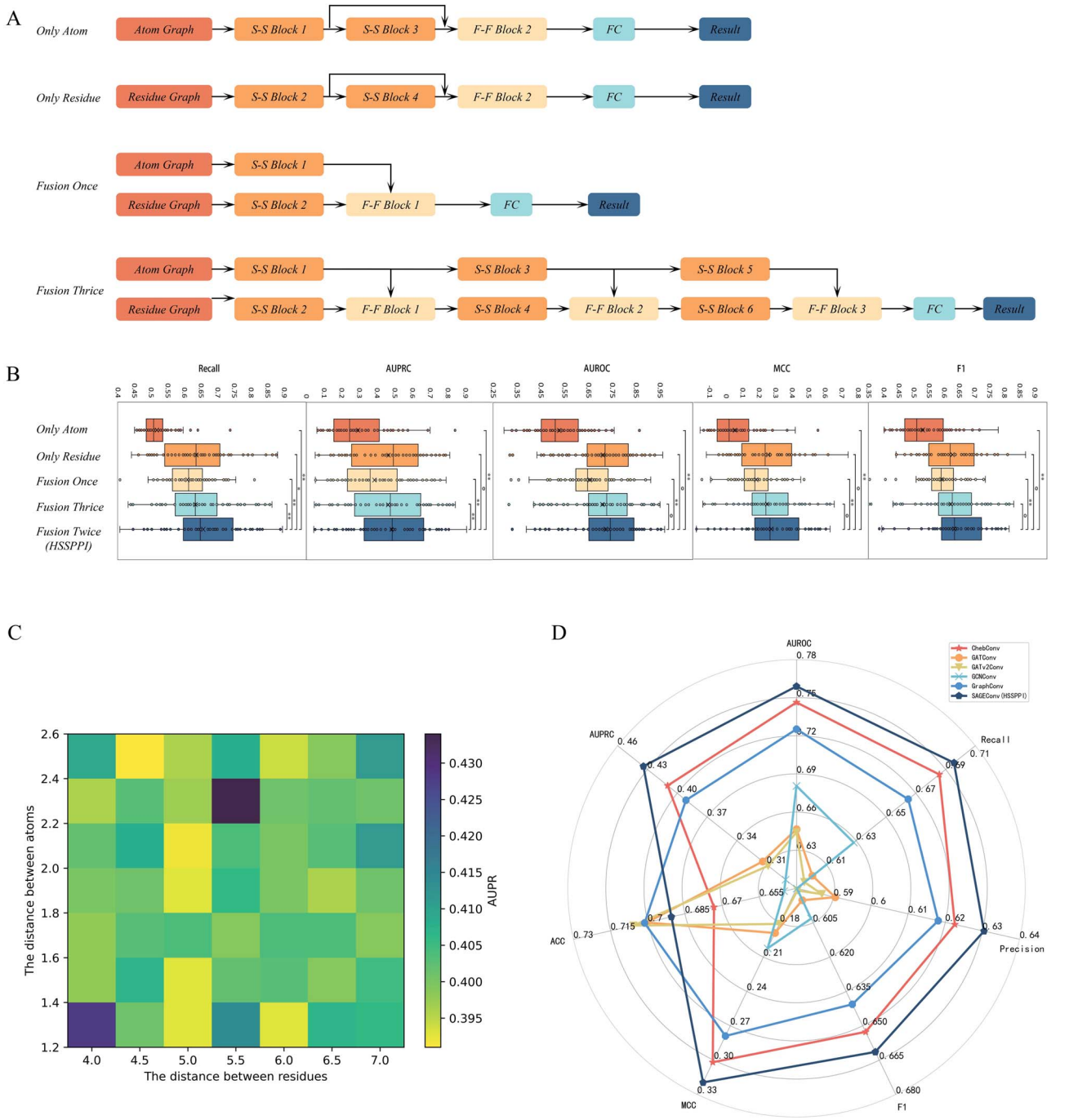

Figure 2. Evaluating HSSPPI on Test70 dataset (A) Four ablation experiments on model architecture, respectively Only Atom, Only Residue, Fusion Once, and Fusion Thrice. **Only-Atom** means that only the atom-level graph is utilized, and two S-S blocks and one F-F block are used. **Only-Residue** means that only the residue-level graph is utilized, and two S-S blocks and one F-F block are used. **Fusion-Once** means that features from two graphs are fused but only one S-S block in each graph and one F-F block are used. **Fusion-Thrice** means that features from two graphs are fused, and three S-S blocks in each graph and three F-F blocks are used. (B) Results comparison of ablation experiments on Recall, AUPRC AUROC, MCC, and F1. ** stands for $P \leq 0.005$. * stands for $0.05 \geq P > 0.005$. ∘ stands for $P > 0.05$. (C) Performance of HSSPPI using different distance thresholds. The unit of the axis is Å. (D) Prediction results using different graph convolution operations in S-S block.

of protein for PPIs prediction. Current computational PPIs prediction methods can be divided into two classes: sequence-based and structure-based.

One of the limitations of PPIs prediction methods is that these methods focus on the usage of sequence or structure form of protein. The protein embedding fed into the predictive model is usually returned by running other tools processing protein sequence or structure. Although the combination of sequence-based and structure-based features can achieve performance improvement. The feature-based deep-learning method only focuses on one form of protein: sequence or structure. To overcome this limitation, we design a S-S block that is comprised of a GCN layer linked with a Bi-GRU layer so that our proposed method HSSPPI can capture spatial and sequential neighbor information at the same time through the two forms of protein. This operation allows us to aggregate complex protein embedding from local

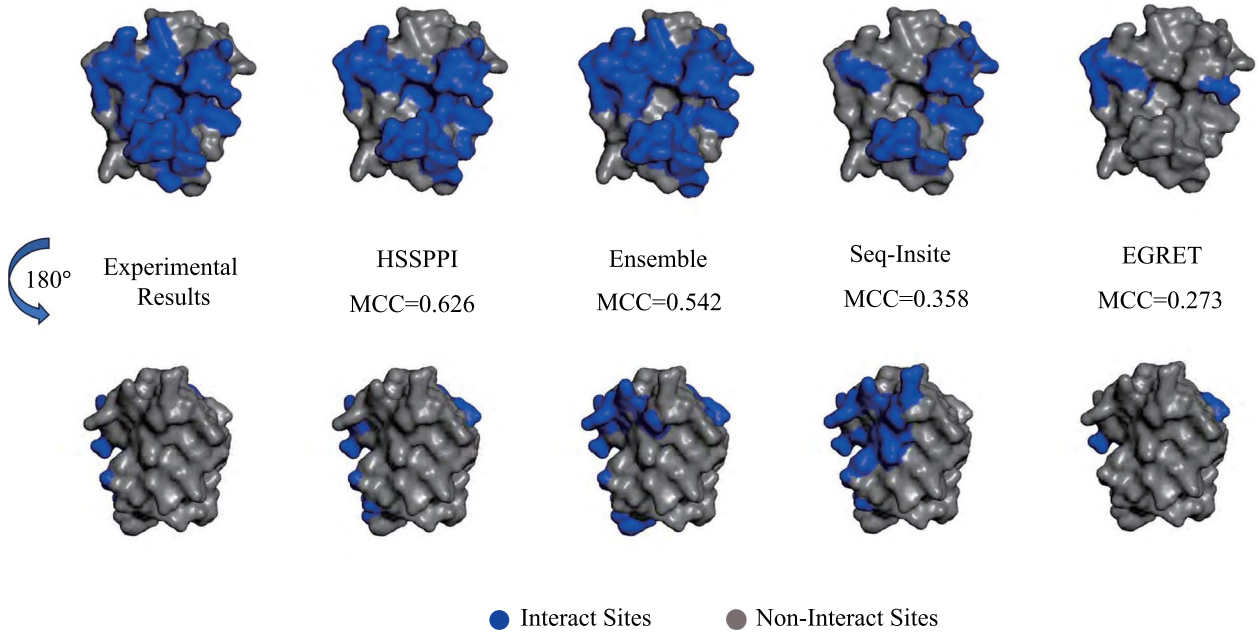

Figure 3. Visualization of a representative protein (PDB ID: 1B6C, Chain ID: A), in which the interaction sites are in blue color and the non-interaction sites are in gray color.

and long-distance neighbors. Another limitation is that recent models ignore the protein's natural hierarchical structure. The hierarchical information has been widely used in various bioinformatics tasks including DTI, DTA, and PPI prediction. Protein structure information at different levels such as atom, residue, motif, protein, and the interaction network is used for capturing conformation features adequately. To further improve the feature representation quality, we represent protein as a hierarchical graph, in which a node in protein is a residue (residue-level graph), and the corresponding atoms in the residue constitute another graph (atom-level graph). We can provide a more comprehensive protein representation through the hierarchical modeling of protein.

Overall, our study introduces HSSPPI, a new model that aims to improve the ability of protein feature representation for PPISP prediction. To demonstrate the effectiveness of HSSPPI, we compare it with eight state-of-the-art competing methods on the DeepPPISP task and GraphPPIS task. The prediction results of HSSPPI are better than these methods on most metrics. The results indicate the effectiveness of F-F from the atom-level and residue-level graphs and the feature extraction through the S-S Bolck.

In the post-Alphafold era, protein complex structure prediction methods could provide more specific results such as AlphaFold-Multimer [63] and DMFold-Multimer [64]. The PPIs are easy to obtain from the predicted complex structure. However, the usage of AlphaFold-Multimer and DMFold-Multimer is long-time consuming [65]. Therefore, the PPIs prediction models including HSSPPI could be one essential useful complement. And, one of the developing trends in protein structure prediction is the usage of atom coordinates such as AlphaFold-Multimer [63] and RoseTTAFold All-Atom [66]. Moreover, HSSPPI tries another way of fusing the structure information of residues and atoms and achieves better performance for protein–protein binding sits prediction.

In conclusion, HSSPPI is a highly competitive PPIs prediction method. Its hierarchical architecture provides a new option for protein feature representation. And, HSSPPI is helpful for biologists to effectively predict PPIs. In future work, we will consider incorporating the partner's features in the feature extraction part.

Key Points

- HSSPPI could fully capture protein conformation information through hierarchical and S-S modeling for accurate PPIs prediction.
- HSSPPI represents the protein as a hierarchical graph for aggregating residue-level and atom-level features.
- HSSPPI adopts the spatial-sequential block for extracting spatial and sequential neighbor information at the same time.
- The results on the two benchmark tasks demonstrate that HSSPPI outperforms the baseline methods on the evaluation metrics.

Author contributions

Yuguang Li (Data curation, Writing—original draft preparation), Xiaofei Nan and Zhen Tian (Writing—review && editing), Shoutao Zhang (Visualization, Investigation), Qinglei Zhou (Visualization, Investigation), and Shuai Lu (Conceptualization, Methodology, Software).

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research was funded by the National Natural Science Foundation of China (Grant No.62371423 and 61572444), Natural Science Foundation of Henan Province of China (Grant No. 232300421117), Major Science and Technology Projects in Henan Province (Grant No. 201400210400), Key Research and Development Projects in Henan Province (Grant No. 241111210500), Key Scientific and Technological Project of Henan Province (Grant NO. 252102211042).

References

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 1998;**92**:291–4. https://doi.org/10.1016/S0092-8674(00)80922-8
2. Ahmad M, Imran A, Movileanu L. Overlapping characteristics of weak interactions of two transcriptional regulators with WDR5. *Int J Biol Macromol* 2024;**258**:128969. https://doi.org/10.1016/j.ijbiomac.2023.128969
3. Kundu S, Lin C, Jaiswal M. *et al*. Profiling glycosylphosphatidylinositol (GPI)-interacting proteins in the cell membrane using a bifunctional GPI analogue as the probe. *J Proteome Res* 2023;**22**:919–30. https://doi.org/10.1021/acs.jproteome.2c00728
4. Xiao Y, Chen Y-M, Zou Z. *et al*. Profiling of RNA-binding protein binding sites by in situ reverse transcription-based sequencing. *Nat Methods* 2024;**21**:247–58. https://doi.org/10.1038/s41592-023-02146-w
5. Engelberg K, Bechtel T, Michaud C. *et al*. Proteomic characterization of the Toxoplasma gondii cytokinesis machinery portrays an expanded hierarchy of its assembly and function. *Nat Commun* 2022;**13**:4644. https://doi.org/10.1038/s41467-022-32151-0
6. Qin Y, Li D, Qi C. *et al*. Structure-based development of potent and selective type-II kinase inhibitors of RIPK1. *Acta Pharm Sin B* 2024;**14**:319–34. https://doi.org/10.1016/j.apsb.2023.10.021
7. Qiu J, Liu Q, Li P. *et al*. Ligand-directed photodegradation of interacting proteins: oxidative HER2/HER3 heterodimer degradation with a lapatinib-derived photosensitizer. *J Med Chem* 2023;**66**:10265–72. https://doi.org/10.1021/acs.jmedchem.3c00252
8. Basu S, Zhao B, Biró B. *et al*. Describeprot in 2023: more, higher-quality and experimental annotations and improved data download options. *Nucleic Acids Res* 2024;**52**:D426–33. https://doi.org/10.1093/nar/gkad985
9. Jiang Y-J, Xia Y, Yi-Xuan H. *et al*. Identification of P21 (CDKN1A) activated kinase 4 as a susceptibility gene for familial non-medullary thyroid carcinoma. *Thyroid* 2024;**34**:583–97. https://doi.org/10.1089/thy.2023.0564
10. Zeng AGX, Bansal S, Jin L. *et al*. A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. *Nat Med* 2022;**28**:1212–23. https://doi.org/10.1038/s41591-022-01819-x
11. Li Y, Wang H-B, Cao J-L. *et al*. Proteomic analysis of mitochondria associated membranes in renal ischemic reperfusion injury. *J Transl Med* 2024;**22**:1–23. https://doi.org/10.1186/s12967-024-05021-0
12. Gao Z, Jiang C, Zhang J. *et al*. Hierarchical graph learning for protein–protein interaction. *Nat Commun* 2023;**14**:1093. https://doi.org/10.1038/s41467-023-36736-1
13. Mehlman TS, Biel JT, Azeem SM. *et al*. Room-temperature crystallography reveals altered binding of small-molecule fragments to PTP1B. *Elife* 2023;**12**:e84632. https://doi.org/10.7554/eLife.84632
14. Wang M, Song X, Chen J. *et al*. Intracellular environment can change protein conformational dynamics in cells through weak interactions . *Sci Adv* 2023;**9**:eadg9141. https://doi.org/10.1126/sciadv.adg9141
15. Nardella C, Pagano L, Pennacchietti V. *et al*. An intramolecular energetic network regulates ligand recognition in a SH2 domain. *Protein Sci* 2023;**32**:e4729. https://doi.org/10.1002/pro.4729
16. Yang J, Roy A, Zhang Y. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res* 2012;**41**:D1096–103. https://doi.org/10.1093/nar/gks966
17. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052
18. Li S, Sanan W, Wang L. *et al*. Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. *Curr Opin Struct Biol* 2022;**73**:102344. https://doi.org/10.1016/j.sbi.2022.102344
19. Soleymani F, Paquet E, Viktor H. *et al*. Protein–protein interaction prediction with deep learning: a comprehensive review. *Comput Struct Biotechnol J* 2022;**20**:5316–41. https://doi.org/10.1016/j.csbj.2022.08.070
20. Tang T, Zhang X, Liu Y. *et al*. Machine learning on protein–protein interaction prediction: models, challenges and trends. *Brief Bioinform* 2023;**24**:bbad076.
21. Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 2010;**26**:1841–8. https://doi.org/10.1093/bioinformatics/btq302
22. Taherzadeh G, Yang Y, Zhang T. *et al*. Sequence–based prediction of protein–peptide binding sites using support vector machine. *J Comput Chem* 2016;**37**:1223–9. https://doi.org/10.1002/jcc.24314
23. Wei Z-S, Yang J-Y, Shen H-B. *et al*. A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE Trans Nanobioscience* 2015;**14**:746–60. https://doi.org/10.1109/TNB.2015.2475359
24. Hou Q, De Geest P, Vranken WF. *et al*. Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics* 2017;**33**:1479–87. https://doi.org/10.1093/bioinformatics/btx005.
25. Wang X, Zhang Y, Bin Y. *et al*. Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis. *Comput Biol Med* 2021;**134**:104516. https://doi.org/10.1016/j.compbiomed.2021.104516
26. Zeng M, Zhang F, Fang-Xiang W. *et al*. Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 2019;**36**:1114–20. https://doi.org/10.1093/bioinformatics/btz699
27. Yiwei Li G, Golding B, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* 2021;**37**:896–904.
28. Stringer B, de Ferrante H, Abeln S. *et al*. PIPENN: protein interface prediction from sequence with an ensemble of neural nets. *Bioinformatics* 2022;**38**:2111–8. https://doi.org/10.1093/bioinformatics/btac071
29. Mahbub S, Bayzid MS. EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Brief Bioinform* 2022;**23**:bbab578.
30. Kang Y, Yulong X, Wang X. *et al*. HN-PPISP: a hybrid network based on MLP-mixer for protein–protein interaction site prediction. *Brief Bioinform* 2023;**24**:bbac480.

31. Wang Y, Min Y, Chen X. *et al*. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In: *WWW'21: The Web Conference 2021, Virtual Event / Ljubljana*, Slovenia: ACM IW3C2, 2021, 2921–33.
32. Chu Z, Huang F, Haitao F. *et al*. Hierarchical graph representation learning for the prediction of drug-target binding affinity. *Inform Sci* 2022;**613**:507–23. https://doi.org/10.1016/j.ins.2022.09.043
33. Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. *Commun Chem* 2023;**6**:34. https://doi.org/10.1038/s42004-023-00825-5
34. Zhonghui G, Luo X, Chen J. *et al*. Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics* 2023;**39**:btad410. https://doi.org/10.1093/bioinformatics/btad410
35. 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. https://doi.org/10.1093/bioinformatics/btab643
36. Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 2010;**26**:1841–8. https://doi.org/10.1093/bioinformatics/btq302
37. Singh G, Dhole K, Pai PP. *et al*. Springs: prediction of protein-protein interaction sites using artificial neural networks Technical report, *PeerJ PrePrints*, 2014;**2**:e266v2. https://doi.org/10.7287/peerj.preprints.266v2.
38. Zhou Y, Jiang Y, Yang Y. AGAT-PPIS: a novel protein–protein interaction site predictor based on augmented graph attention network with initial residual and identity mapping. *Brief Bioinform* 2023;**24**:1–8. https://doi.org/10.1093/bib/bbad122.
39. Jumper J, Evans R, Pritzel A. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
40. She D, Jia M. A BiGRU method for remaining useful life prediction of machinery. *Measurement* 2021;**167**:108277. https://doi.org/10.1016/j.measurement.2020.108277
41. Elnaggar A, Heinzinger M, Dallago C. *et al*. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:7112–27.
42. Hou Z, Yang Y, Ma Z. *et al*. Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning. *Commun Biol* 2023;**6**:73. https://doi.org/10.1038/s42003-023-04462-5
43. Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions. *Proteins* 2007;**66**:630–45. https://doi.org/10.1002/prot.21248
44. Zhang J, Kurgan L. Scriber: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 2019;**35**:i343–53. https://doi.org/10.1093/bioinformatics/btz324
45. Shuai L, Li Y, Nan X. *et al*. Attention-based convolutional neural networks for protein-protein interaction site prediction. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX, USA: IEEE, 2021, 141–44.
46. Yiwei Li G, Golding B, Ilie L. Delphi: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* 2021;**37**:896–904.
47. Mou M, Pan Z, Zhou Z. *et al*. A transformer-based ensemble framework for the prediction of protein–protein interaction sites. *Research* 2023;**6**:0240. https://doi.org/10.34133/research.0240
48. Qiu J, Bernhofer M, Heinzinger M. *et al*. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J Mol Biol* 2020;**432**:2428–43. https://doi.org/10.1016/j.jmb.2020.02.026
49. Zhang B, Li J, Lijun Quan Y. *et al*. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* 2019;**357**:86–100. https://doi.org/10.1016/j.neucom.2019.05.013
50. Gainza P, Sverrisson F, Monti F. *et al*. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92. https://doi.org/10.1038/s41592-019-0666-6
51. Wang S, Chen W, Han P. *et al*. RGN: residue-based graph attention and convolutional network for protein–protein interaction site prediction. *J Chem Inf Model* 2022;**62**:5961–74. https://doi.org/10.1021/acs.jcim.2c01092
52. Khan SH, Tayara H, Chong KT. ProB-site: protein binding site prediction using local features. *Cells* 2022;**11**:2117.
53. Fang Y, Jiang Y, Wei L. *et al*. DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. *Bioinformatics* 2023;**39**:btad718.
54. Zeng X, Meng F-F, Li X. *et al*. GHGPR-PPIS: a graph convolutional network for identifying protein-protein interaction site using heat kernel with generalized PageRank techniques and edge self-attention feature processing block. *Comput Biol Med* 2024;**168**:107683. https://doi.org/10.1016/j.compbiomed.2023.107683
55. Meng L, Zhang H. GACT-PPIS: prediction of protein-protein interaction sites based on graph structure and transformer network. *Int J Biol Macromol* 2024;**283**:137272. https://doi.org/10.1016/j.ijbiomac.2024.137272
56. Hamilton WL, Ying Z, Leskovec J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, Long Beach, CA, USA: Curran Associates, Inc., 2017, 1024–34.
57. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, Barcelona, Spain: Curran Associates, Inc., 2016, 3837–45.
58. Veličković P, Cucurull G, Casanova A. *et al*. Graph attention networks. In *The 6th International Conference on Learning Representations*, Vancouver, BC, Canada: OpenReview.net, 2018.
59. Brody S, Alon U, Yahav E. How attentive are graph attention networks. In *The Tenth International Conference on Learning Representations*, Virtual Event: OpenReview.net, 2022.
60. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In the *5th International Conference on Learning Representations*, Toulon, France: OpenReview.net, 2017.
61. Morris C, Ritzert M, Fey M. *et al*. Weisfeiler and leman go neural: higher-order graph neural networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA: AAAI Press, 2019.
62. DeLano WL. *et al*. Pymol: an open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr* 2002;**40**:82–92.
63. Evans R, O'Neill M, Pritzel A. *et al*. Protein complex prediction with AlphaFold-multimer bioRxiv. 2021.
64. Zheng W, Wuyun Q, Li Y. *et al*. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* 2024;**21**:279–89. https://doi.org/10.1038/s41592-023-02130-4

65. Krapp LF, Abriata LA, Rodriguez FC. *et al*. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* 2023;**14**:2175. https://doi.org/10.1038/s41467-023-37701-8

66. Krishna R, Wang J, Ahern W. *et al*. Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science* 2024;**384**:eadl2528. https://doi.org/10.1126/science.adl2528

Briefings in Bioinformatics, 2025, **26(2)**, bbaf079
https://doi.org/10.1093/bib/bbaf079
Problem Solving Protocol