SDNN-PPI: Self-attention with deep neural networks effect on protein-protein interaction prediction

Xue Li
China University of Petroleum (East China)
Peifu Han
China University of Petroleum (East China)
Gan Wang
China University of Petroleum (East China)
Wenqi Chen
China University of Petroleum (East China)
Shuang Wang
China University of Petroleum (East China)
Tao Song (t.song@upm.es)
China University of Petroleum (East China)

Research Article

Keywords: Protein-protein interactions, Deep learning, Deep neural network, Self-attention

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1632165/v1

RESEARCH

SDNN-PPI: self-attention with deep neural networks effect on protein-protein interaction prediction

Xue Li, Peifu Han, Gan Wang, Wenqi Chen, Shuang Wang and Tao Song*

*Correspondence: t.song@upm.es
College of Computer Science and technology, China University of Petroleum (East China), Qingdao, China
Full list of author information is available at the end of the article

Abstract

Background: Protein-protein interactions (PPIs) dominate intracellular molecules to perform a series of tasks such as transcriptional regulation, information transduction, and drug signalling. The traditional wet experiment method to obtain PPIs information is costly and time-consuming.

Result: In this paper, SDNN-PPI, a PPI prediction method based on self-attention and deep learning is proposed. The method adopts amino acid composition (AAC), conjoint triad (CT) and auto covariance (AC) to extract global and local features of protein sequences, and leverages self-attention to enhance DNN feature extraction to more effectively accomplish the prediction of PPIs. In order to verify the generalization ability of SDNN-PPI, a 5-fold cross-validation on the intraspecific interactions dataset of Saccharomyces cerevisiae (core subset) and human is used to measure our model in which the accuracy rates reach to 95.4% and 98.94% respectively. The accuracy rates of 93.15% and 88.33% are obtained in the interspecific interactions dataset of human-Bacillus Anthracis and Human-Yersinia pestis, respectively. In the independent data set Caenorhabditis elegans, Escherichia coli, Homo sapiens and Mus musculus, the prediction accuracy is 100% respectively, which is higher than the previous PPIs prediction methods. To further evaluate the advantages and disadvantages of the model, the one-core and crossover network are conducted to predict PPIs, and the data show that the model correctly predicts the interaction pairs in the network.

Conclusion: In this paper, AAC, CT and AC methods are used to encode the sequence, and SDNN-PPI method is proposed to predict PPIs based on self-attention deep learning neural network. Satisfactory results are obtained on interspecific and intraspecific data sets, and good performance is also achieved in cross-species prediction. It can also correctly predict the protein interaction of cell and tumor information contained in one-core network and crossover network.The SDNN-PPI proposed in this paper not only explores the mechanism of protein-protein interaction, but also provides new ideas for drug design and disease prevention.

Keywords: Protein-protein interactions; Deep learning; Deep neural network; Self-attention

Introduction

Text and results for this section, as per the individual journal's instructions for authors.Proteins are organic macromolecules made up of amino acids, which are essential components of cells and sustain life activities. They play an important role

in biology by linking various important physiological activities of cells to PPIs [1], enabling a range of life activities such as apoptosis and immune response. In recent years, a large number of high-throughput experimental methods have emerged to study PPIs, such as yeast two-hybrid screening [2], mass spectrometry [3], hybridization methods [4], immunoprecipitation [5] and protein microarrays [6]. However, all of these are based on biological and chemical experiments, which require a lot of manpower, financial and time resources. Therefore, artificial intelligence-based computational methods have emerged in bioinformatics and become quite prevalent in PPIs studies. Currently there is abundant of data related to amino acid sequence information, which are the prerequisites to build computational models for PPIs prediction [7]. A growing number of researchers have been attracted by the aforementioned methods. The basic steps of PPIs prediction based on protein sequence consist of two parts: protein coding method and machine learning model.

With the rapid development of machine learning techniques [8, 9, 10] and the refinement of neural networks [11, 12, 13, 14], some machine learning-based and sequence-based models have been presented for PPIs prediction. Shen et al.[15] first employed CT (conjoint triad) to extract features from protein sequences and predicted PPIs through support vector machine model incorporating kernel function with 83.9% accuracy. Guo et al. [16] proposed auto covariance (AC) to extract information from protein sequences and used support vector machine model to predict PPIs in the Saccharomyces cerevisiae dataset with 86.55% accuracy. Yang et al. [17] proposed local descriptors (LD) to represent protein sequences and successfully predicted potential PPIs on Saccharomyces cerevisiae dataset by implementing K-neighbor model. You et al. [18] utilized four categories of protein sequence information (AC, CT, LD, MAC) to encode proteins as feature vectors focusing on dimensionality reduction and proposed a new hierarchical PCA-EELM (principal component analysis-integrated extreme learning machine) model to predict protein interactions. In 2014, Barman et al. [19] used support vector machine, Naive Bayes and random forest based on 5-fold cross-validation to complete the host-pathogen interaction prediction. In 2016, An et al. [20] jointly proposed a new computational method called RVM-BiGP, combining the relevance vector machine (RVM) model and Bi-gram probabilities (BiGP), to efficiently handle imbalanced protein interaction datasets. In 2018, Goktepe et al. [21] adopted PCA to fuse PSSM, Bi-gram, AAC, pseudo-amino acid (PseAAC) and weighted jump-order joint triple to obtain approximate features, then used SVM to complete PPIs prediction. Song et al. [22] used PSSM to obtain evolutionary information and proposed a new feature fusion algorithm, which could combine discrete cosine transform (DCT), fast Fourier transform (FFT) and singular value decomposition (SVD). In 2019, Chen et al. [23] extracted features from PseAAC, autocorrelation descriptor (AD), CT and LD by elastic network, and predicted PPI in several datasets with the help of LightGBM network. In 2020, Yu et al. [24] proposed a combination of PseAAC, pseudo-position-specific scoring Matrix (PsePSSM), reduced sequence and index-vectors (RSIV), and AD to encode protein sequences for potential PPIs on Saccharomyces cerevisiae dataset through GTB-PPI model.

Although machine learning methods can make predictions based on best fitting models, it is still open to some limitations on effectively learning the eigenvalues

at a deep level. In recent years, deep learning architectures [25, 26, 27, 28] provide strong support for solving relevant problems in bioinformatics. In 2017, Wang et al. [29] extracted protein sequence features from PSSM, and reconstructed them through stacked auto-encoder. After that, prediction was completed with the help of a new probabilistic classification vector machine (PCVM). Du et al. [30] proposed a deep neural network model, DeepPPI, to improve the performance of PPIs prediction using AAC, DC, LD and other protein transformations where demonstrated the superiority of the model on several datasets. Wang et al. [31] combined Deep Neural Networks (DNNs) with a new local composition ternary description (LCTD) feature representation, and proposed DNN-LCTD method to predict the PPIs on Saccharomyces cerevisiae dataset with the accuracy of 93.12%. In 2018, Hashemifar et al. [32] efficiently combined deep Siamese-like convolutional neural networks and random projection to construct DPPI model for predicting PPIs by associating with protein evolutionary information. In 2019, Zhang et al. [33] proposed a deep model called EnsDNN, which extracted protein interaction information from AC, LD and multi-scale continuous and discontinuous local descriptors (MCD) which achieved 95.29% accuracy in Saccharomyces cerevisiae dataset. You et al. [34] proposed a highly efficient method to detect PPIs by integrating a new protein sequence substitution matrix feature representation and ensemble weighted sparse representation model classifier. Yao et al [35] designed a new protein sequence representation method, Res2vec, and combined effective feature embedding with deep learning techniques to develop the DeepFE-PPI framework, which achieved good performance in PPIs prediction. In 2020, Li et al [36] represented proteins using AC, CT, LD, PseAAC, and built Ensemble model to complete PPIs prediction work. In 2021, Yu et al [37] used PseAAC, AD, multivariate mutual information (MMI), composition-transition-distribution (CTD), amino acid composition PSSM (AAC-PSSM), and dipeptide composition PSSM (DPC-PSSM) to construct the pattern of GcForest-PPI.

Inspired by the above discussion, this paper proposes a protein-protein interaction prediction method, SDNN-PPI. Firstly, protein sequence information is encoded with AAC, CT, and AC. Second of all, in order to carry out effective feature extraction, the deep neural network combined with self-attention method is conducted to adjust the weight of the sequence and further emphasize the key features, so as to establish a network model to fully extract protein sequence information. Eventually, 5-fold cross-validation approach is applied in 2 intraspecies, 2 interspecies, and 4 independent datasets. All of which achieved high accuracy rates. To further evaluate the merits of the model, the effectiveness of the method is tested on one-core network and crossover network. The experimental results show that SDNN-PPI outperforms other state-of-the-art methods and is highly competitive.

Materials and methods

Data sets

In this study, multiple high-confidence PPI datasets were used to measure the performance of SDNN-PPI, including the intraspecific datasets Saccharomyces cerevisiae core subset (S.cerevisiae core subset) [16] and Human [34], the interspecific dataset Human-Bacillus Anthracis (Human-B.Anthracis) [38] and the Human-Yersinia pestis (Human-Y.pestis) [38]. The composition of the four datasets is shown

in Table 1. In addition, four independent datasets [23] including Caenorhabditis elegans (C.elegans), Escherichia coli (E.coli), Homo sapiens (H.sapiens) and Mus musculus (M.musculus) are tested for PPIs. And the predictive performance of the method is further validated on two significant PPI networks [37]. One is the one-core CD9 network, which contains 16 PPIs, and the other is crossover network, which consists of 96 PPIs. In addition, to ensure the balance of positive and negative samples in the dataset, the same number of randomly selected negative samples is in the same amount as positive samples meaning the ratio of positive to negative samples was 1:1.

Feature extraction techniques

Since the length of the protein sequence is different, the input to the neural network used in the experiment is fixed. The protein sequences of different lengths have to be transformed into feature vectors of fixed length when they are input into network layers. In this paper, the feature fusion strategy is used to convert protein sequences into feature vectors based on AAC, CT and AC. AAC has the advantage of obtaining the proportion of each amino acid in the entire protein sequence from a global perspective. CT regards any continuous three amino acids as a unit, and puts the characteristics of amino acids and their adjacent amino acids into consideration, but ignores the information of amino acid discontinuous fragments. In terms of physicochemical properties, AC extracts not only discontinuous fragment information, but also the interaction features of long-distance amino acids by considering the adjacent effects of amino acids. In summary, this method extracts amino acid global features through ACC, and then uses CT to reduce the defect of few short-range amino acid interactions in ACC. And through the AC, which is based on the physicochemical properties, the local features of amino acids with adjacent effects were extracted, and more comprehensive protein information was obtained, which provided strong support for the downstream feature extraction.

Amino acid composition (AAC)

The amino acid composition method [30] normalizes the frequency of occurrence of each amino acid in the protein, which is a concise protein feature extraction method. Specifically, the frequency of twenty amino acids in protein sequences is counted, and each protein sequence is converted into a 1×20-dimensional feature vector. The feature extraction formula is as follows:

$$\mathrm{P}(x) = \frac{n}{N} \tag{1}$$

Where n represents the number of amino acid x in the protein sequence and N represents the number of all amino acids in the protein.

Conjoint triad (CT)

The combined triplet method [23] takes an amino acid and its left and right amino acids as a unit, and divides 20 amino acids into 7 different clusters [15] according to the volume of amino acid side chains and dipoles (as shown in Table 2). Among them, different amino acids belonging to a certain cluster are considered to be the

same. Therefore, the obtained feature is a 343-dimensional feature vector, which is the normalized results of triples (7*7*7). The formula is:

$$\mathrm{P}(C) = \frac{N_C}{N-2} \tag{2}$$

Among them, C represents a triplet, N_C represents the number of occurrences of this triplet, N represents the number of all amino acids in the protein, and the denominator represents that a protein sequence has $N-2$ triplets.

Auto covariance (AC)

The autocovariance method [33] mainly considers the proximity effect of amino acids. The interaction between an amino acid and a certain number of surrounding amino acids is in Hydrophobicity (H1) Hydrophilicity (H2), Net Charge Index (NCI). Polarity (P1), Polarizability (P2), solvent-accessible Surface Area (SASA), and other seven physicochemical properties. The amino acid sequence is replaced by the initial values of the seven physical and chemical properties, and normalized to zero mean and unit standard deviation (SD), as shown in Formula (3).

$$F_{ij} = \frac{f_{i,j} - f_j}{S_i} \tag{3}$$

Where $f_{(i,j)}$ represents the value of the j-th property of the i-th amino acid, f_j represents the average value of the j-th property of 20 amino acids, and S_i represents the corresponding standard deviation. The formula for calculating AC is as follows.

$$\begin{aligned} AC_{\mathrm{lag,j}} = \frac{1}{N - \mathrm{lag}} \sum_{i=1}^{N-lag} \left(F_{ij} - \frac{1}{N}\sum_{i=1}^{N} F_{ij} \right) \times \\ \left(F_{(i+lag),j} - \frac{1}{N}\sum_{i=1}^{N} F_{ij} \right) \end{aligned} \tag{4}$$

Among them, lag represents the distance between the residuals, and N represents the length of the protein sequence. In this paper, j is taken as 7, and lag is taken as 30, and finally a protein sequence is encoded as a 210-dimensional feature vector.

PPIs model based on self-attention combined with deep neural network

The simple neural network receives data at the input layer, transforms the data through multiple hidden layers, and finally computes the result at the output layer. Neurons in the hidden or output layer are connected to all neurons in the previous layers, as shown in Figure 1A. Each neuron computes a weighted sum of its inputs and applies a nonlinear activation function to compute its output $f(x)$ (Figure 1B). The most commonly used activation function is the Rectified linear unit (ReLU), which sets the negative signal threshold to 0 and allows positive signals to pass normally. The deep neural network (DNN) proposed in recent years is an artificial neural network inspired by the neural network of the brain, which consists of multiple interconnected computing units (neurons) and extracts high-level abstractions

from data. DNN is widely used in speech recognition, PPIs [30], and other fields with its powerful feature extraction ability. DNN takes the received data as input, then transforms it in a non-linear way, and the last layer outputs [39, 40]. With regard to avoid over fitting, a dropout layer is also added to drop some neurons during training, as shown in Figure 2.

Self-attention mechanism (Figure 3) is a model framework proposed by the Google team [41] in 2017, which can reduce the dependence on external information and be better at capturing the internal correlation of data or features, especially long-distance dependency. As shown in Figure 3, the weight is obtained by calculating the similarity of Q and K after linear transformation, then the softmax function is used to normalize the weight, and finally attention is obtained by the weight and V. More specifically, let $U = [u_1, u_2, ..., u_N]$ represents the output vector of the embedding layer, which contains the input of N amino acids, and U_i represents the feature vector of the i-th amino acid. Then, the output of the self-attention module is the weighted sum of feature vectors on all the amino acids, and its core formula is [42]:

$$s_i = \sum_{j=1}^{N} \text{softmax}\left(\frac{q_j k_j^T}{\sqrt{d_k}}\right) v_j \tag{5}$$

Where d_k square root represents the scaling factor to control the magnitude of the dot product. q_i, k_i and v_i represent the query, key and value of the i-th amino acid, respectively.

Based on the excellent performance of deep neural network and Attention mechanism, this paper proposes a DNN network that applies multi-layer fully connected layers and self-attention to predict PPIs, named SDNN-PPI. Deep networks have the characteristics of synthesizing various information, but as the number of layers increases, the risk of overfitting will increase, and the focus on key data will also be reduced. Therefore, this paper dynamically pays attention to the key residues in the sequence through the self-attention in the feature extraction layer, adjusts the weights, captures the feature of single residue, promotes the prediction process, and avoids falling into local optimum caused by DNN overfitting. In addition, since self-attention has a strong ability to extract internal features, it is widely used to capture long-range dependencies between tokens in sequential data. Therefore, in the prediction stage, self-attention mechanism is used to enhance the feature extraction of protein pairs, and further exploits the potential relationship of residues to obtain more accurate information. The SDNN-PPI model is shown in Figure 4. It mainly includes three modules, namely the feature extraction layer, the feature fusion layer, and the PPIs prediction layer.

(1) Input layer: The model is based on two proteins (P1, P2) as input, and converts the protein sequences into feature vectors through the three encoding methods of AAC, CT, and AC. Finally, each protein sequence is encoded into a vector with dimension of 573, which consists of 20 AAC features, 343 CT features, and 210 AC features respectively.

(2) Feature extraction layer: SDNN-PPI is composed of two channels, which extract the hidden information of proteins respectively. Each channel is composed of

six fully connected layers (1024-512-256-128-64-32) by adding a self-attention layer that adjusts the global weight of the sequence. To avoid gradient vanishing and over fitting, Batch Normalization and Dropout layers are added after each dense layer. The formula is expressed as:

$$f = \text{Dropout}(BN(\text{Dense}(P))) \tag{6}$$

Where P represents the feature vector of protein sequence, and f represents the output through the full connection layer.

(3) Feature fusion layer: The feature fusion layer connects the protein information (F1', F2 ') obtained by the two channels from the feature extraction layer. The formula is expressed as:

$$F = \text{cat}\,(f1', f2') \tag{7}$$

(4) Prediction layer: The prediction layer is composed of three fully connected layers (31-16-8) and a self-attention layer. Self-attention layer is conducive to increasing the exploration of protein pairs, which is put after the first dense layer. Then there is a single neuron with a Sigmoid activation function that converts the input from the previous layer into an output score. The formula is as follows:

$$\text{P}(P1, P2) = s(\text{Dens}(F)) \tag{8}$$

where s denotes dense layer with one unit activated by sigmoid function.

Evaluation metrics

The following assessments are used for this article: Accuracy (ACC), Sensitivity (Sens), Specificity (Spec), Precision (Prec), Matthews Correlation Coefficient (MCC), and AUC. These assessments are used to calculate accuracy and bias to assess the feasibility and robustness of PPI forecasting methodologies. The definition formula is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{13}$$

Among them, TP (True Positive) is the number of correctly predicted protein pair interactions in the sample data set, TN (True Negative) is the number of correctly predicted protein pairs that do not interact, FP (False Positive) is the number of non-interacting protein pairs predicted as interacting, while FN (False Negative) is the number of interacting protein pairs predicted as non-interacting.

In order to prove the statistical significance of SDNN-PPI, kappa coefficient[43] is also added. Kappa coefficient is an indicator to measure the consistency of two variables[44], which can be used to evaluate the classification accuracy. The results of Kappa coefficient are usually between 0 and 1[45]. When the result is in the range of 0.0 to 0.20, the classification result is considered to be slight, kappa=0.21-0.40 means fair, kappa=0.41-0.60 is moderate, kappa=0.61-0.80 describes substantial, and kappa $>$ 0.81 represents almost perfect. Its calculation formula[43] is as follows:

$$\mathrm{pa} = \frac{p_0 - p_e}{1 - p_e} \tag{14}$$

Where p_0 means accuracy,

$$p_e = \frac{(TN+FP)(TN+FN)}{(TP+FP+TN+FN)^2} + \frac{(FN+TP)(FP+TP)}{(TP+FP+TN+FN)^2} \tag{15}$$

Results and discussion

This part mainly evaluates and discusses the performance of the model. 3.1 describes that the encoding method used in this work can achieve ideal results. After discussion in 3.2, the framework of the model is settled. 3.3 shows the results on two intraspecific and two interspecific datasets. In order to evaluate the validity of the model, 3.4 compares SDNN-PPI with the current advanced algorithms on the intraspecies and interspecific data sets in 3.4. In part 3.5, four independent data sets are used to prove the robustness of the model. In the last part, PPI Network further proves the potential capability of the model in predicting disease development.

Encoding method selection

In this paper, encoding methods of ACC, CT and AC were used to construct 573-dimensional feature vectors to encode proteins, which can extract global and local features. In addition, LD was also used to encode local characteristics of proteins [23]. LD can encode each protein sequence into a 630-dimensional vector. In order to verify the encoding scheme, LD was also originally used in our experiments as another optional encoding method for protein pairs, and S. Cerevisiae (Core Subset) data set was selected to search for best encoding combination scheme based on the experimental results of the model. In order to avoid the dependence of the encoding method on SDNN-PPI model, standard two-channel self-attention model was selected to verify the encoding scheme. As shown in Table 3, compared with the

other 10 combination schemes, the ACC+CT+AC encoding combination scheme achieved the optimal results on 6 evaluation indicators. However, after the addition of LD in encoding scheme ACC+CT+AC, the results did not improve effectively, which may be due to the fact that LD was not accurate enough to extract the features of the encoding of excessively long protein sequences, resulting in poor effects.

Model ablation experiment

To verify the effect of different network structures on the performance of SDNN-PPI, two different network structures were first designed: (a) using a dual-channel network to extract protein information (DNN-PPI a) and (b) directly connect two proteins in a single channel network (DNN-PPI b). As can be seen from the first two lines of Table 4, the dual-channel model was superior to the single-channel model, and the ACC, Spec, Sens, Prec, MCC, and AUC values of DNN-PPI a were 3.12%, 2.79%, 5.84%, 2.92%, 6.22%, and 1.49% higher than those of DNN-PPI b, respectively. Secondly, after setting up the dual-channel model, the meaning of Self-attention was studied: (c) self-attention was added in feature extraction layer (SDNN-PPI a), (d) self-attention was added in prediction layer (SDNN-PPI b), (e) self-attention was added in both feature extraction layer and prediction layer (SDNN-PPI), (f) dual-channel network without self-attention (DNN-PPI a). After building different networks, the S.cerevisiae (core subset) dataset was used to evaluate the model results. As shown in Table 4, the SDNN-PPI performed better, so this model was chosen as the final framework.

Performance of the SDNN-PPI

When training a model with dataset, it is easy to overfit due to unreasonable division of the dataset. Compared with the division technique of traditional models (dividing fixed training sets and test sets), cross-validation can avoid such problems, so this paper uses the 5-fold cross-validation method to evaluate the model. The experimental data is randomly divided into 5 parts, samples of 4 parts are randomly taken as the training set, the other part is used as the test set, and finally the average of the 5 test sets is calculated. The performance of this method was compared with several advanced methods, and the results were shown in Table 5-8.

As can be seen from Table 5, SDNN-PPI had an excellent prediction performance for intraspecific data sets. The average prediction results of S.cerevisiae (core subset) in ACC, Spec, Sens, Prec, MCC and AUC were 95.48%, 97.23%, 93.80%, 97.13%, 91.02% and 98.63, respectively. Similarly, the average results of the Human dataset were ACC 98.94%, Spec 99.02%, Sens 98.54%, Prec 99.02%, MCC 97.57%, and AUC 99.60%, as shown in Table 6. Meanwhile, for the interspecific data set, as shown in Table 7-8, SDNN-PPI achieved 93.15% and 88.33% accuracy in Human-B.anthracis and Human-Y.pestis, respectively. The above experimental results show that the prediction of PPIs by SDNN-PPI is effective and robust. Table 9 presented the statistical significance of SDNN-PPI in four data sets. According to the above description, kappa between 0.61-0.80 indicates that the classification results were substantial, and when kappa > 0.81, the classification results were almost perfect. The kappa values of the 4 data sets in Table 9 were all greater than 0.61 and 3 were greater than 0.81, indicating that the results were statistically significant.

Compared with other methods

To predict protein-protein interactions, various prediction methods have been continuously proposed. In order to more objectively evaluate the predictive performance of the constructed model, the prediction results were compared with other models in the same data set. The comparison results of the intraspecific datasets S.cerevisiae (core subset) and Human were shown in Table 10 and Table 11. The interspecific datasets Human-B.Anthracis and Human-Y.pestis results were shown in Table 11 and Table 12. For comparison methods, the data in the table were extracted from the original text, and N/A means that the data is not available in the original text.

As can be seen from Table 10, ACC, Spec, Sens, Prec, MCC, and AUC of SDNN-PPI were 95.48%, 97.23%, 93.80%, 97.13%, 91.02% and 98.63%, respectively. Compared with other methods, its ACC increased by 0.04% 2.18%. According to Table 11, ACC, Spec, Sens, Prec, MCC, and AUC of SDNN-PPI in Human data set were 98.94%, 99.02%, 98.54%, 99.02%, 97.57% and 99.60, respectively. Compared with other methods, the accuracy of this method is obviously improved. Although SDNN-PPI was not the best on individual indicators, the individual optimality rate was 5/6 on S. cerevisiae and human datasets, indicating that the method was still competitive. For this, the predictive performance of SDNN-PPI method became significantly better than other methods in multiple indicators.

As can be seen from Table 12, ACC, Sens, Prec, MCC and AUC of SDNN-PPI in Human-B.anthracis data set are 93.15%, 96.61%, 90.44%, 86.57% and 98.23%, respectively. The ACC of SDNN-PPI is 93.15%, which is significantly higher than other methods. According to Table 13, the ACC, Sens, Prec, MCC and AUC of SDNN-PPI in Human-Y.pestis data set were 88.33%, 93.92%, 84.63%, 77.26% and 95.74%, respectively. In comparison to other methods, its ACC value was 1.03% 12.23% higher than other methods. Therefore, the SDNN-PPI method achieves better results on interspecies datasets. It was worth noting that the two tables do not display Spec columns because the models being compared did not have Spec values.

Performance on independent data sets

In order to further verify the generalization ability of SDNN-PPI, Saccharomyces Cerevisiae citeDu2017 was selected as the training set, and C.legans, E.coli, H.sapiens and M.musculus were selected as independent test sets. The number of interaction pairs of the independent test set was shown in the test pairs in the first row of Table 14. In addition, the results were evaluated by ACC. Saccharomyces Cerevisiae set consists of 17257 positive pairs and 48594 negative pairs, from which the same number of positive and negative samples are randomly selected to train the model. The prediction results were shown in Table 14. As can be seen from Table 14, the accuracy of SDNN-PPI in these four independent data sets was 100%. This can show that SDNN-PPI achieved good predictive performance on four independent test sets, indicating that the proposed model can characterize important PPIs information and make cross-species predictions. In other words, PPIs prediction models generated by one species can be migrated to other species.

Performance on PPI Networks

Studying the network of PPIs [24] is also of great significance to understanding other information about proteins, and the corresponding biological topological properties can be studied. In this paper, SDNN-PPI detected two important PPIs networks, namely the one-core network and crossover network of Wnt-related pathway. The mononuclear PPIs network is a network of PPIs composed of a core protein, CD9 [37], and interacts with many other proteins. CD9 is a tetrameric protein that plays an important role in cell viability and tumor suppression. The network is composed of CD9 as the core protein and 17 other genes.

The second is a typical crossover and multicore network [46] constructed by 78 genes. This pathway network plays a crucial role in tumor growth and tumor formation. AAC, CT, and AC were used to encode proteins to obtain a 573-dimensional feature vector. The Saccharomyces cerevisiae dataset was used as the training set, and the one-core network and crossover network of the wnt-related paths were used as the test set. The one-core network prediction results of the wnt-related paths were shown in Figure 5, and the other in Figure 6. Solid lines represent true predictions and dashed lines represent false predictions. It can be obtained from the graph that all interacting proteins are correctly identified. Table 15 showed the prediction results of various methods on the two network datasets. The results shown that the proposed method produces comparable or better results in comparison to existing models. After the above discussion, SDNN-PPI was a model with high generalization ability, which can obtain competitive results in multiple data sets and effectively improve the prediction accuracy of PPIs.

Conclusion

The study of PPIs is of great significance for understanding cellular regulation and signal transduction, as well as for exploring and elucidating the mechanism of protein interactions in cells. In this paper, we proposed SDNN-PPI, a self-attention-based deep learning neural network prediction method for PPIs. The protein sequences were encoded by AAC, CT and AC methods, and excellent accuracy was obtained in the intraspecific data sets (S. cerevisiae and Human) and interspecies data sets (Human-B.anthracis and Human-Y.pestis). In order to further verify the universality of SDNN-PPI, the evaluation of C.legans, E.Coli, H.sapiens and M.mesculus data sets also achieved competitive accuracy, indicating that the method can also achieve good performance in cross-species prediction. The PPI network prediction based on one-core and crossover network correctly predicted the protein interaction containing cell and tumor information on the network. Therefore, comprehensive evaluations demonstrated that SDNN-PPI method could provide a new way to solve problems in signaling pathway research, drug-target prediction and disease pathogenesis research [47, 48, 49]. Although protein sequences are transformed into vectors through various encoding methods, the acquisition of comprehensive protein characteristic information is still insufficient. How to better mine the structural information, evolutionary information set of protein pairs and the relationship between protein residues is leading us to the next research direction. At the same time, DNA computing and DNA storage [50, 51] have been applied in more fields [52, 53], and the storage of known protein information and structure may also play a role in promoting biological evolution.

Appendix

Acknowledgements
We thank our partners who provided all the help during the research process and the team for their great support.

Funding
This work was supported by National Key Research and Development Project of China (2021YFA1000102, 2021YFA1000103), Natural Science Foundation of China (Grant Nos. 61873280, 61972416), Taishan Scholarship (tsqn201812029), Foundation of Science and Technology Development of Jinan (201907116), Shandong Provincial Natural Science Foundation(ZR2021QF023), Fundamental Research Funds for the Central Universities (21CX06018A), Spanish project PID2019-106960GB-I00, Juan de la Cierva IJC2018-038539-I.

Abbreviations
PPI: Protein-protein interactions; AAC: amino acid composition; CT: conjoint triad; AC: auto covariance; LD: local descriptor; TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; NB: Naive Bayes; DNN: Deep neural network.

Availability of data and materials
The data and code underlying this article are available in https://github.com/xueleecs/SDNN-PPI.The article all data set on the https://github.com/xueleecs/SDNN-PPI/tree/main/Data

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
XL designed the study, drafted the manuscript. PH, GW, WC helped with ablation experiments. SW helped design the model.TS participated in revise the manuscript.

Author details
College of Computer Science and technology, China University of Petroleum (East China), Qingdao, China.

References
1. Humphreys, I.R., Pei, J.M., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., Stancheva, V.G., Li, X.H., Liu, K.X., Zheng, Z., Barrero, D.J., Roy, U., Kuper, J., Fernandez, I.S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E.C., Biggins, S., Keeney, S., Miller, E.A., Fromme, J.C., Hendrickson, T.L., Cong, Q., Baker, D.: Computed structures of core eukaryotic protein complexes. Science **374**(6573), 1340 (2021). doi:10.1126/science.abm4805
2. Bacon, K., Blain, A., Bowen, J., Burroughs, M., McArthur, N., Menegatti, S., Rao, B.M.: Quantitative yeast-yeast two hybrid for the discovery and binding affinity estimation of protein-protein interactions. Acs Synthetic Biology **10**(3), 505–514 (2021). doi:10.1021/acssynbio.0c00472
3. Woodall, D.W., Dillon, T.M., Kalenian, K., Padaki, R., Kuhns, S., Semin, D.J., Bondarenko, P.V.: Non-targeted characterization of attributes affecting antibody-fc gamma riiia v158 (cd16a) binding via online affinity chromatography-mass spectrometry. Mabs **14**(1) (2022). doi:10.1080/19420862.2021.2004982
4. Hu, L., Wang, X.J., Huang, Y.A., Hu, P.W., You, Z.H.: A survey on computational models for predicting protein-protein interactions. Briefings in Bioinformatics **22**(5) (2021). doi:10.1093/bib/bbab036
5. Susila, H., Nasim, Z., Jin, S., Youn, G., Jeong, H., Jung, J.-Y., Ahn, J.H.: Profiling protein-dna interactions by chromatin immunoprecipitation in arabidopsis. Methods in molecular biology (Clifton, N.J.) **2261**, 345–356 (2021). doi:10.1007/978-1-0716-1186-9$_2$1
6. Ma, J.F., Wu, C., Hart, G.W.: Analytical and biochemical perspectives of protein o-glcnacylation. Chemical Reviews **121**(3), 1513–1581 (2021). doi:10.1021/acs.chemrev.0c00884
7. Chou, K.C., Cai, Y.D.: Predicting protein-protein interactions from sequences in a hybridization space. Journal of Proteome Research **5**(2), 316–322 (2006). doi:10.1021/pr050331g
8. Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J.: Next-generation machine learning for biological networks. Cell **173**(7), 1581–1592 (2018). doi:10.1016/j.cell.2018.05.015
9. Fang, W.W., Yao, X.N., Zhao, X.J., Yin, J.W., Xiong, N.X.: A stochastic control approach to maximize profit on service provisioning for mobile cloudlet platforms. Ieee Transactions on Systems Man Cybernetics-Systems **48**(4), 522–534 (2018). doi:10.1109/tsmc.2016.2606400
10. Li, H.H., Liu, J.X., Liu, R.W., Xiong, N.X., Wu, K.F., Kim, T.H.: A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis. Sensors **17**(8) (2017). doi:10.3390/s17081792
11. Song, T., Pang, S., Hao, S., Rodriguezpaton, A., Zheng, P.: A parallel image skeletonizing method using spiking neural p systems with weights. Neural Processing Letters **50**(2), 1485–1502 (2019)
12. Song, T., Zeng, X., Zheng, P., Jiang, M., Rodriguezpaton, A.: A parallel workflow pattern modeling using spiking neural p systems with colored spikes. IEEE Transactions on Nanobioscience **17**(4), 474–484 (2018)
13. Song, T., Zheng, P., Wong, M.L.D., Wang, X.: Design of logic gates using spiking neural p systems with homogeneous neurons and astrocytes-like control. Information Sciences **372**, 380–391 (2016). doi:10.1016/j.ins.2016.08.055

14. Song, T., Rodriguez-Paion, A., Zheng, P., Zeng, X.X.: Spiking neural p systems with colored spikes. IEEE Transactions on Cognitive and Developmental Systems **10**(4), 1106–1115 (2018). doi:10.1109/tcds.2017.2785332
15. Shen, J.W., Zhang, J., Luo, X.M., Zhu, W.L., Yu, K.Q., Chen, K.X., Li, Y.X., Jiang, H.L.: Predictina protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America **104**(11), 4337–4341 (2007). doi:10.1073/pnas.0607879104
16. Guo, Y.Z., Yu, L.Z., Wen, Z.N., Li, M.L.: Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. Nucleic Acids Research **36**(9), 3025–3030 (2008). doi:10.1093/nar/gkn159
17. Yang, L., Xia, J.F., Gui, J.: Prediction of protein-protein interactions from protein sequence using local descriptors. Protein and Peptide Letters **17**(9), 1085–1090 (2010). doi:10.2174/092986610791760306
18. You, Z.H., Lei, Y.K., Zhu, L., Xia, J.F., Wang, B.: Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. Bmc Bioinformatics **14** (2013). doi:10.1186/1471-2105-14-s8-s10
19. Barman, R.K., Saha, S., Das, S.: Prediction of interactions between viral and host proteins using supervised machine learning methods. Plos One **9**(11) (2014). doi:10.1371/journal.pone.0112034
20. An, J.Y., Meng, F.R., You, Z.H., Chen, X., Yan, G.Y., Hu, J.P.: Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein Science **25**(10), 1825–1833 (2016). doi:10.1002/pro.2991
21. Goktepe, Y.E., Kodaz, H.: Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing **303**, 68–74 (2018). doi:10.1016/j.neucom.2018.03.062
22. Song, X.Y., Chen, Z.H., Sun, X.Y., You, Z.H., Li, L.P., Zhao, Y.: An ensemble classifier with random projection for predicting protein-protein interactions using sequence and evolutionary information. Applied Sciences-Basel **8**(1) (2018). doi:10.3390/app8010089
23. Chen, C., Zhang, Q.M., Ma, Q., Yu, B.: Lightgbm-ppi: Predicting protein-protein interactions through lightgbm with multi-information fusion. Chemometrics and Intelligent Laboratory Systems **191**, 54–64 (2019). doi:10.1016/j.chemolab.2019.06.003
24. Yu, B., Chen, C., Zhou, H.Y., Liu, B.Q., Ma, Q.: Gtb-ppi: Predict protein-protein interactions based on l1-regularized logistic regression and gradient tree boosting. Genomics Proteomics Bioinformatics **18**(5), 582–592 (2020). doi:10.1016/j.gpb.2021.01.001
25. Quang, D., Xie, X.H.: Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic Acids Research **44**(11) (2016). doi:10.1093/nar/gkw226
26. Pang, S.C., Zhang, Y., Song, T., Zhang, X.D., Wang, X., Rodriguez-Paton, A.: Amde: a novel attention-mechanism-based multidimensional feature encoder for drug-drug interaction prediction. Briefings in Bioinformatics **23**(1) (2022). doi:10.1093/bib/bbab545
27. Wang, S., Jiang, M.J., Zhang, S.G., Wang, X.F., Yuan, Q., Wei, Z.Q., Li, Z.: Mcn-cpi: Multiscale convolutional network for compound-protein interaction prediction. Biomolecules **11**(8) (2021). doi:10.3390/biom11081119
28. Wang, S., Song, T., Zhang, S., Jiang, M., Wei, Z., Li, Z.: Molecular substructure tree generative model for de novo drug design. Briefings in bioinformatics (2022). doi:10.1093/bib/bbab592
29. Wang, Y.B., You, Z.H., Li, X., Jiang, T.H., Chen, X., Zhou, X., Wang, L.: Predicting protein-protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Molecular Biosystems **13**(7), 1336–1344 (2017). doi:10.1039/c7mb00188f
30. Du, X.Q., Sun, S.W., Hu, C.L., Yao, Y., Yan, Y.T., Zhang, Y.P.: Deepppi: Boosting prediction of protein-protein interactions with deep neural networks. Journal of Chemical Information and Modeling **57**(6), 1499–1510 (2017). doi:10.1021/acs.jcim.7b00028
31. Wang, J., Zhang, L., Jia, L.Y., Ren, Y.Z., Yu, G.X.: Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. International Journal of Molecular Sciences **18**(11) (2017). doi:10.3390/ijms18112373
32. Hashemifar, S., Neyshabur, B., Khan, A.A., Xu, J.B.: Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics **34**(17), 802–810 (2018). doi:10.1093/bioinformatics/bty573
33. Zhang, L., Yu, G.X., Xia, D.W., Wang, J.: Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing **324**, 10–19 (2019). doi:10.1016/j.neucom.2018.02.097
34. You, Z.H., Huang, W.Z., Zhang, S.W., Huang, Y.A., Yu, C.Q., Li, L.P.: An efficient ensemble learning approach for predicting protein-protein interactions by integrating protein primary sequence and evolutionary information. Ieee-Acm Transactions on Computational Biology and Bioinformatics **16**(3), 809–817 (2019). doi:10.1109/tcbb.2018.2882423
35. Yao, Y., Du, X.Q., Diao, Y.Y., Zhu, H.X.: An integration of deep learning with feature embedding for protein-protein interaction prediction. Peerj **7** (2019). doi:10.7717/peerj.7126
36. Li, F.F., Zhu, F., Ling, X.H., Liu, Q.: Protein interaction network reconstruction through ensemble deep learning with attention mechanism. Frontiers in Bioengineering and Biotechnology **8** (2020). doi:10.3389/fbioe.2020.00390
37. Yu, B., Chen, C., Wang, X.L., Yu, Z.M., Ma, A.J., Liu, B.Q.: Prediction of protein-protein interactions based on elastic net and deep forest. Expert Systems with Applications **176** (2021). doi:10.1016/j.eswa.2021.114876
38. Kosesoy, I., Gok, M., Oz, C.: A new sequence based encoding for prediction of host-pathogen protein interactions. Computational Biology and Chemistry **78**, 170–177 (2019). doi:10.1016/j.compbiolchem.2018.12.001
39. Angermueller, C., Parnamaa, T., Parts, L., Stegle, O.: Deep learning for computational biology. Molecular Systems Biology **12**(7) (2016). doi:10.15252/msb.20156651
40. Webb, S.: Deep learning for biology. Nature **554**(7693), 555–557 (2018). doi:10.1038/d41586-018-02174-z
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: 31st Annual Conference on Neural Information Processing Systems (NIPS). Advances in Neural Information Processing Systems, vol. 30 (2017). ¡Go to ISI¿://WOS:000452649406008

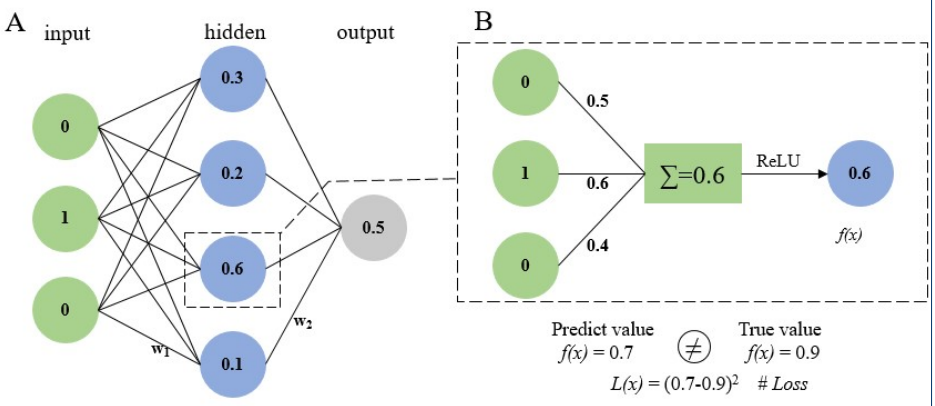

Figure 1 Neural network procedure

42. Lei, Y.P., Li, S.Y., Liu, Z.Y., Wan, F.P., Tian, T.Z., Li, S., Zhao, D., Zeng, J.Y.: A deep-learning framework for multi-level peptide-protein interaction prediction. Nature Communications **12**(1) (2021). doi:10.1038/s41467-021-25772-4
43. Dey, L., Mukhopadhyay, A.: Compact genetic algorithm-based feature selection for sequence-based prediction of dengue-human protein interactions. IEEE/ACM transactions on computational biology and bioinformatics **PP** (2021). doi:10.1109/tcbb.2021.3066597
44. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–74 (1977). doi:10.2307/2529310
45. Tang, W., Hu, J., Zhang, H., Wu, P., He, H.: Kappa coefficient: a popular measure of rater agreement. Shanghai archives of psychiatry **27**(1), 62–7 (2015). doi:10.11919/j.issn.1002-0829.215010
46. Chen, C., Zhang, Q.M., Yu, B., Yu, Z.M., Lawrence, P.J., Ma, Q., Zhang, Y.: Improving protein-protein interactions prediction accuracy using xgboost feature selection and stacked ensemble classifier. Computers in Biology and Medicine **123** (2020). doi:10.1016/j.compbiomed.2020.103899
47. Li, L., Gao, Z., Wang, Y.T., Zhang, M.W., Ni, J.C., Zheng, C.H.: Scmfmda: Predicting microrna-disease associations based on similarity constrained matrix factorization. Plos Computational Biology **17**(7) (2021). doi:10.1371/journal.pcbi.1009165
48. Su, Y.S., Liu, C.L., Niu, Y.Y., Cheng, F., Zhang, X.Y.: A community structure enhancement-based community detection algorithm for complex networks. IEEE Transactions on Systems Man Cybernetics-Systems **51**(5), 2833–2846 (2021). doi:10.1109/tsmc.2019.2917215
49. Tian, Y., Su, X.C., Su, Y.S., Zhang, X.Y.: Emodmi: A multi-objective optimization based method to identify disease modules. IEEE Transactions on Emerging Topics in Computational Intelligence **5**(4), 570–582 (2021). doi:10.1109/tetci.2020.3014923
50. Cao, B., Li, X., Zhang, X., Wang, B., Zhang, Q., Wei, X.: Designing uncorrelated address constrain for dna storage by dmvo algorithm. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020). doi:10.1109/TCBB.2020.3011582
51. Wu, J., Zheng, Y., Wang, B., Zhang, Q.: Enhancing physical and thermodynamic properties of dna storage sets with end-constraint. IEEE transactions on nanobioscience **PP** (2021). doi:10.1109/tnb.2021.3121278
52. Zhou, S.H.: A real-time one-time pad dna-chaos image encryption algorithm based on multiple keys. Optics and Laser Technology **143** (2021). doi:10.1016/j.optlastec.2021.107359
53. Song, T., Wang, X., Li, X., Zheng, P.J.O.: A programming triangular dna origami for doxorubicin loading and delivering to target ovarian cancer cells. Oncotarget **5** (2017)
54. Wang, Y.B., You, Z.H., Yang, S., Li, X., Jiang, T.H., Zhou, X.: A high efficient biological language model for predicting protein-protein interactions. Cells **8**(2) (2019). doi:10.3390/cells8020122
55. Sharma, A., Singh, B.: Ae-lgbm: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and lightgbm. Computers in Biology and Medicine **125** (2020). doi:10.1016/j.compbiomed.2020.103964
56. An, J.Y., You, Z.H., Zhou, Y., Wang, D.F.: Sequence-based prediction of protein-protein interactions using gray wolf optimizer-based relevance vector machine. Evolutionary Bioinformatics **15** (2019). doi:10.1177/1176934319844522

Figures

Tables

Table 1 Compositions of the Four Benchmark Data Sets

Data sets	Interaction pairs	Noninteraction pairs	Protein pairs
S.cerevisiae(core subset)	5594	5594	11188
Human	3899	4262	8161
Human-B.Anthracis	3094	9500	12594
Human-Y.pestis	4097	12500	16597

Table 2 Classification of amino acids based on amino acid side chains and dipole volume

Cluster	Amino acid
1	A, G, V
2	I, L, F, P
3	Y, M, T, S
4	H, N, Q, W
5	R, K
6	D, E
7	C

Table 3 Performance of different coding methods

Encoding methods	Length	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
ACC+CT+LD+AC	1203	92.31±0.66	94.37±0.25	90.26±1.28	94.13±0.25	84.70±1.27	97.00
ACC+CT+LD	993	92.00±0.66	93.31±1.04	90.69±0.57	93.14±1.01	84.03±1.33	97.03
ACC+CT+AC	573	**95.19±0.68**	**97.05±0.65**	**93.33±1.20**	**96.94±0.65**	**90.45±1.34**	**98.60**
ACC+LD+AC	860	91.41±0.52	93.06±0.87	89.76±0.64	92.83±0.86	82.87±1.06	96.58
CT+LD+AC	1183	89.50±0.68	90.97±0.58	88.02±1.47	90.70±0.50	79.04±1.33	95.62
ACC+CT	363	89.79±0.65	90.79±0.81	88.79±0.87	90.61±0.76	79.61±1.29	95.78
ACC+LD	650	88.93±0.43	89.67±1.37	88.20±1.53	89.54±1.14	77.91±0.88	95.05
ACC+AC	230	85.74±1.80	87.20±1.74	84.29±2.66	86.82±1.74	71.54±3.60	92.59
CT+LD	973	90.47±0.52	92.76±0.57	88.18±0.93	92.42±0.55	81.03±1.02	95.97
CT+AC	553	89.06±0.55	90.70±0.87	87.43±1.34	90.40±0.73	78.19±1.07	94.74
LD+AC	840	91.44±0.44	92.72±0.72	90.17±0.56	92.54±0.69	82.92±0.89	96.60

Table 4 Comparison among different layer architectures for SDNN-PPI on S.cerevisiae(core subset)

Architectures	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
DNN-PPI a	94.9	96.35	95.83	96.25	89.84	98.54
DNN-PPI b	91.78	93.56	89.99	93.33	83.62	97.05
SDNN-PPI a	95.16	96.96	93.37	96.86	90.4	98.53
SDNN-PPI b	95.21	96.98	93.44	96.87	90.48	98.56
SDNN-PPI	**95.48**	**97.23**	**93.80**	**97.13**	**91.02**	**98.63**

Table 5 Prediction results of S.cerevisiae (core subset) under five-fold cross-validation

testing set	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
1	95.44	97.59	93.48	97.48	90.97	98.42
2	95.17	96.69	94.1	96.59	90.39	98.92
3	95.13	97.32	93.3	97.20	90.35	98.28
4	95.49	96.37	94.36	96.27	91.98	98.74
5	96.16	98.21	93.74	98.14	92.39	98.80
average	95.48±0.37	97.23±0.66	93.80±0.39	97.13±0.66	91.02±0.74	98.63

Table 6 Prediction results of Human data set under five-fold cross-validation

testing set	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
1	98.78	99.49	98.08	98.97	97.06	99.67
2	99.29	98.85	99.49	99.23	98.46	99.63
3	98.85	99.10	98.46	98.97	97.56	99.51
4	98.78	99.36	98.72	99.1	97.95	99.72
5	98.97	98.84	99.10	98.83	96.8	99.46
average	98.94±0.19	99.10±0.24	98.77±0.49	99.02±0.13	97.57±0.60	99.60

Table 7 Prediction results of Human-B.Anthracis data set under five-fold cross-validation

testing set	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
1	91.44	85.14	97.74	86.8	83.54	97.93
2	93.78	93.05	94.51	93.15	87.57	98.65
3	92.49	87.72	97.25	88.79	85.36	98.03
4	94.26	90.78	97.74	91.39	88.73	98.22
5	93.78	91.76	95.79	92.07	87.62	98.32
average	93.15±1.03	89.69±2.88	96.61±1.27	90.44±2.32	86.57±1.87	98.23

Table 8 Prediction results of Human-Y.pestis data set under five-fold cross-validation

testing set	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
1	85.91	75.98	95.85	79.94	73.28	95.73
2	88.16	83.9	92.43	85.15	76.61	95.39
3	88.16	80.83	95.49	83.3	77.16	95.73
4	91.27	89.26	93.29	89.68	82.62	96.31
5	88.16	83.76	92.55	85.07	76.61	95.55
average	88.33±1.71	82.74±4.34	93.92±6.06	84.63±1.85	77.26±3.79	95.74

Table 9 Prediction results of four data sets in kappa coefficient

data sets	S.cerevisiae(core subset)	Human	Human-B.Anthracis	Human-Y.pestis
kappa	0.91	0.98	0.85	0.76

Table 10 Comparison results of different PPIs prediction methods on S. cerevisiae (core subset)

Methods	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
DeepPPI[30]	94.43±0.30	N/A	92.06±0.36	96.65±0.59	88.97±0.62	N/A
DeepFE-PPI[35]	94.78±0.61	N/A	92.99±0.66	96.45±0.87	89.62±1.23	N/A
LightGBM-PPI[23]	95.07	**97.94**	92.21	97.82	0.903	0.903
Bio2Vec[54]	93.30	N/A	92.70	93.55	87.49	97.20
StackPPI[46]	94.64	96.46	92.81	96.33	89.34	N/A
GTB-PPI[24]	95.15±0.25	N/A	92.21±0.36	97.97±0.60	90.45±0.53	N/A
AE-LGBM[55]	95.40±0.20	98.70±0.20	92.10±0.30	N/A	91.00±0.40	N/A
GcForest-PPI[37]	95.44±0.18	N/A	92.72±0.44	**98.05±0.25**	91.02±0.35	N/A
SDNN-PPI	**95.48±0.37**	97.23±0.66	**93.80±0.39**	97.13±0.66	**91.02±0.74**	**98.63**

Table 11 Comparison results of different PPIs prediction methods on Human

Methods	ACC(%)	Spec(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
RPEC[22]	96.59	N/A	96.72	96.18	93.18	N/A
Bio2Vec[54]	97.31	N/A	96.28	98.48	94.76	99.61
GWOSVM[56]	94.56	N/A	95.55	93.08	89.51	N/A
DeepFE-PPI[30]	98.71±0.30	N/A	98.54±0.55	98.77±0.53	97.43±0.61	N/A
AE-LGBM[55]	98.70±0.10	**99.20±0.20**	98.10±0.20	N/A	97.30±0.30	N/A
AE-AC[55]	97.19	98.06	96.34	N/A	N/A	N/A
SDNN-PPI	**98.94±0.19**	99.10±0.24	**98.77±0.49**	**99.02±0.13**	**97.57±0.60**	**99.60**

Table 12 Comparison results of different PPIs prediction methods on Human-B.Anthracis

Methods	ACC(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
LBE-BN[56]	78.70	73.00	42.00	43.4	83.70
LBE-NB[56]	82.50	53.80	47.80	39.70	82.10
LBE-RF[56]	85.40	24.0	67.00	34.00	86.80
ACC-BN[56]	77.40	51.70	37.30	30.30	79.00
LBE-j48[56]	80.06	31.20	39.60	23.90	54.10
LD-DNN[38]	91.70	89.50	93.90	83.50	96.37
SDNN-PPI	**93.15**	**96.61**	**90.44**	**86.57**	**98.23**

Table 13 Comparison results of different PPIs prediction methods on Human-Y.pestis

Methods	ACC(%)	Sens(%)	Prec(%)	MCC(%)	AUC(%)
LBE-BN[56]	76.10	73.50	38.60	40.10	81.30
LBE-NB[56]	80.90	45.50	43.2	32.80	78.60
LBE-RF[56]	84.6	16.00	66.30	27.30	83.50
ACC-BN[56]	80.00	52.40	42.10	34.90	75.60
LBE-j48[56]	80.10	27.90	37.10	20.80	51.70
LD-DNN[38]	87.30	84.20	90.40	74.90	94.99
SDNN-PPI	**88.33**	**93.92**	**84.63**	**77.26**	**95.74**

Table 14 Comparison of ACC of different PPIs prediction methods on independent test sets

Species/Methods	C.elegans	E.coli	H.sapiens	M.musculus
test pairs	4013	6984	1412	313
DeepPPI[30]	94.84	92.19	93.77	91.37
DeepFE-PPI[35]	**100**	**100**	**100**	**100**
LightGBM-PPI[23]	90.16	92.16	94.83	94.57
StackPPI[46]	97.11	98.71	97.66	98.40
GcForest-PPI[37]	96.01	96.3	98.58	99.04
GTB-PPI[24]	92.42	94.06	97.38	98.08
AE-LGBM[55]	90.10	92.10	94.80	94.50
SDNN-PPI	**100**	**100**	**100**	**100**

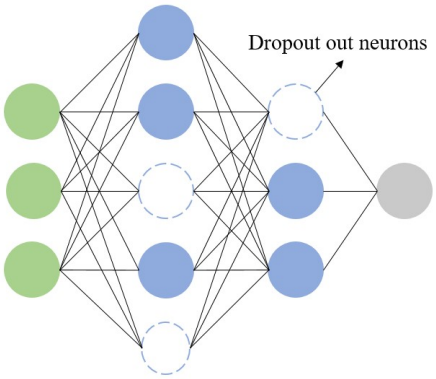

Figure 2 Networks with dropout

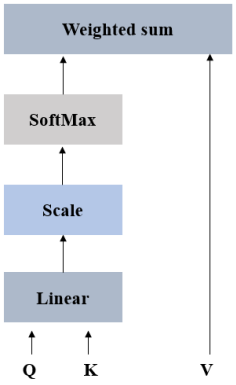

Figure 3 Model of Self-attention

Table 15 Performance of different methods on PPI network

	LightGBM-PPI[23]	StackPPI[46]	GTB-PPI[24]	AE-LGBM[55]	GcForest-PPI[37]	SDNN-PPI
CD9	15/16	N/A	15/16	16/16	16/16	**16/16**
Wnt	89/96	93/96	92/96	95/96	94/96	**96/96**

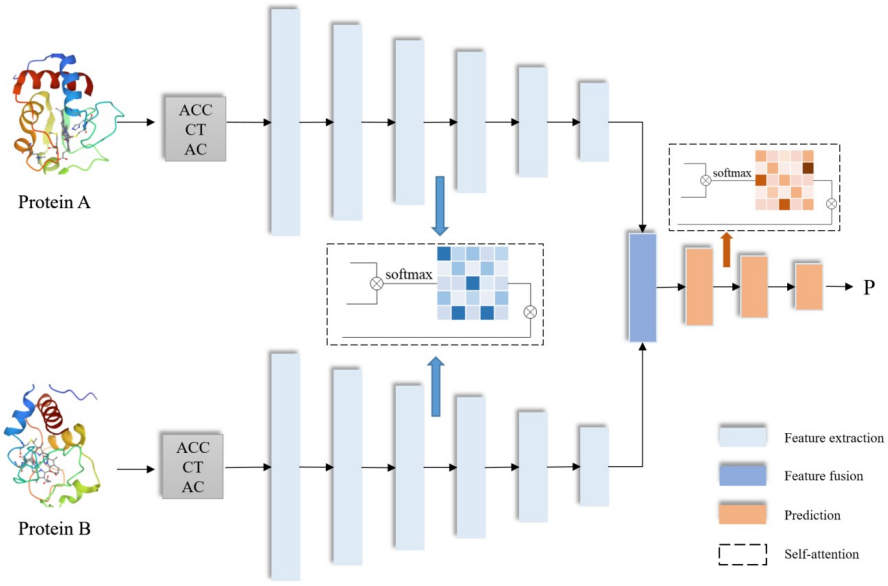

Figure 4 SDNN-PPI for prediction of protein-protein interaction

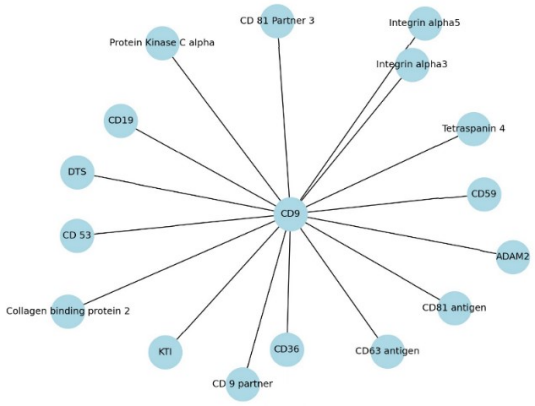

Figure 5 The predicted results of PPIs networks of a one-core network for CD9

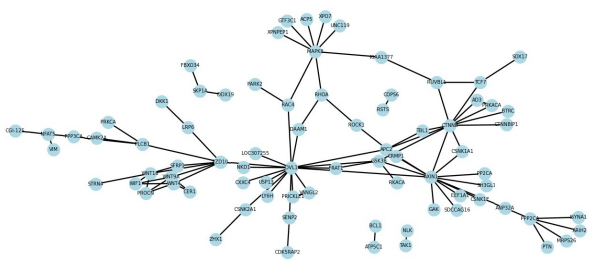

Figure 6 The predicted results of a crossover network for the Wnt-related pathway