BMC Bioinformatics

RESEARCH Open Access

PCP-GC-LM: single-sequence-based protein contact prediction using dual graph convolutional neural network and convolutional neural network

J. Ouyang[1,2], Y. Gao[1,2*] and Y. Yang[2]

*Correspondence:
oyjq@xtu.edu.cn

[1] Key Laboratory of Intelligent Computing Information Processing, Xiangtan University, Xiangtan, China
[2] School of Computer Science, Xiangtan University, Xiangtan, China

Abstract

Background: Recently, the process of evolution information and the deep learning network has promoted the improvement of protein contact prediction methods. Nevertheless, still remain some bottleneck: (1) One of the bottlenecks is the prediction of orphans and other fewer evolution information proteins. (2) The other bottleneck is the method of predicting single-sequence-based proteins mainly focuses on selecting protein sequence features and tuning the neural network architecture, However, while the deeper neural networks improve prediction accuracy, there is still the problem of increasing the computational burden. Compared with other neural networks in the field of protein prediction, the graph neural network has the following advantages: due to the advantage of revealing the topology structure via graph neural network and being able to take advantage of the hierarchical structure and local connectivity of graph neural networks has certain advantages in capturing the features of different levels of abstraction in protein molecules. When using protein sequence and structure information for joint training, the dependencies between the two kinds of information can be better captured. And it can process protein molecular structures of different lengths and shapes, while traditional neural networks need to convert proteins into fixed-size vectors or matrices for processing.

Results: Here, we propose a single-sequence-based protein contact map predictor PCP-GC-LM, with dual-level graph neural networks and convolution networks. Our method performs better with other single-sequence-based predictors in different independent tests. In addition, to verify the validity of our method against complex protein structures, we will also compare it with other methods in two homodimers protein test sets (DeepHomo test dataset and CASP-CAPRI target dataset). Furthermore, we also perform ablation experiments to demonstrate the necessity of a dual graph network. In all, our framework presents new modules to accurately predict inter-chain contact maps in protein and it's also useful to analyze interactions in other types of protein complexes.

Keywords: Protein contact prediction, Single-sequence-based, Graph neural network, Convolutional neural network

Introduction

Residues contacts provide proteins translationally and rationally invariant topological representation and can provide distance information or residues-residues interaction information as used in many protein-related problems that include drug design, protein design, protein function prediction, and protein structure prediction [1–8]. In the past years, many methods have been developed for protein contact map prediction due to the success of evolutionary- information-based and deep learning architecture in protein contact map prediction. These methods based on evolutionary information mostly rely on the quality of the Multiple sequence alignment (MSA) and infer residue pairs with direct evolutionary couplings from the MSA of homologous proteins. These methods can accurately predict residue pair contacts when providing a large amount of MSA and other evolutionary information [9–14]. However, many proteins have little or even no homology to generate MSA homology profiles. Facing these proteins, the methods based on evolutionary information may have poor performance. Predicting protein contacts for proteins with little or no homology to generate Multiple sequence alignment (MSA) homology profiles remains a challenging task.

Previous protein prediction methods, usually search for similar sequences from protein databases for multiple sequence comparisons to mine the underlying evolution information or co-evolution information. However, these methods face difficulties in protein with less evolution information. Recently, single-sequence-based protein prediction methods have gained attention. These methods use the protein sequence as input to predict the protein structure and function, without relying on homologous information. Therefore, the first difficulty of single-sequence-based protein prediction methods is how to select effective protein sequence features. SSCpred [15] predicts residue pairs contact map using protein sequences initialized as one-hot code and protein structure properties of SPIDER3-Single as input and employs a full convolution model with 30 ResNet blocks. Experimental results demonstrated that compared with several most evolutionary-information-based methods, SSCpred achieves completive performance on nonhomology protein targets. On this basis, to explore other possible features beyond one-hot encoding to improve the performance of single-sequence-based methods facing few homologous information. Influenced by pre-trained models for natural language processing, many large-scale pre-trained language models for protein prediction have also been proposed. The language model processed by this pre-trained language model can obtain information on relevant biological feature. These representational feature are obtained only from the amino acid sequence and have multi-scale organizational learning, which can reflect biochemical constraints from amino acids to the long-range homology of proteins. The relevant information of the protein sequence can be used as the feature input of the downstream protein prediction task, reducing the scale of model training and improving the prediction accuracy. SPOT-Contact-LM [16] is a published single-sequence-based method after SSCpred. In addition to using one-hot as input, the attention map of the protein language model Esm-1b [17] is also used as input. Unlike SSCpred which only uses one-hot encoding as an input feature, Spot-Contact-LM precision is improved by 20% and uses fewer ResNet blocks. In the work of SPOT-Contact-LM, they trained six models with the same architecture but different input feature combinations or different training strategies. Their work showed that the attention map from Esm-1b is as input to improve prediction precision, which indicates that

the protein feature from Esm-1b may contain potential protein connection distribution. Therefore, selecting the feature of the protein sequence and digging information from the protein feature improves single-sequence-based prediction precision by using protein pre-training models.

Usually, the contact map prediction task is treated as an image segmentation task, in which each residue pair corresponds to a pixel. ResNet has an excellent performance in image segmentation [18–21], and methods based on ResNet architecture have beenthe mainstream of protein contact prediction methods. The success of ResNet in contact map prediction has benefited from its blocks of skip-connected convolutional layers. The 2d inflation 1d sequential feature is directly combined with the other residue pairs feature or only takes the residue pairs feature as the input feature. Because the ResNet block focuses on local neighbor residues in the input feature but may not consider those residues which are far apart in the input feature and this can lead to loss of information. To resolve the long-range contact problem, it is usually to build deeper networks or more convolution blocks to combine features of the more residue pairs. There is an increasing computation burden and hard training while improving prediction accuracy via deeper neural networks.

It's a natural way to express a protein with the graph of nodes representing residues and edges feature representing residues-residues pair feature. As a popular deep learning method in recent years, the Graph Neural Network (GNN) is widely used in chemistry, biology, material, and other fields [22–28]. These methods can learn the acceptable representation embedding from both the original data of nodes and edges features, and be used for different node-level and graph-level predictions. The GNN makes use of its advantages of the topological structure to encode context and global information and is wildly used in molecular feature extraction, drug discovery, molecular distance prediction and drug feature representation. Considering the potential information of residues feature embedding and the interaction between residues-residues, we will obtain structural information from the interaction relationship between residues, and use the graph neural network to capture the interaction relationship between the features of different dimensions. In this study, we propose a new model, named PCP-GC-LM, which consists of a graph neural network and a 2d convolution to predict contact maps. Our method applies a graph neural network to obtain sequence embedding and residues pairs features by aggregating protein sequence information. Instead of a single graph to update node or edge features, we design an effective dual graph process modeling for the update feature which updates the association between protein residues and sets two branches to update the information of different graphs. To verify the effect of the dual graph on updated graph representation, we intentionally use fewer convolution blocks to decode and predict protein contact probability. We prove that our method has fewer convolution blocks than other single-sequence-based methods and achieves a more remarkable effect through our graph update module.

Methods

In this study, we propose a new single-sequence-based protein contact prediction model in which the novelty of the proposed model is to construct dual-graph branches to update protein embedding features. We describe protein as a graph, in which residues sequence feature are regarded as nodes and residues pair feature as

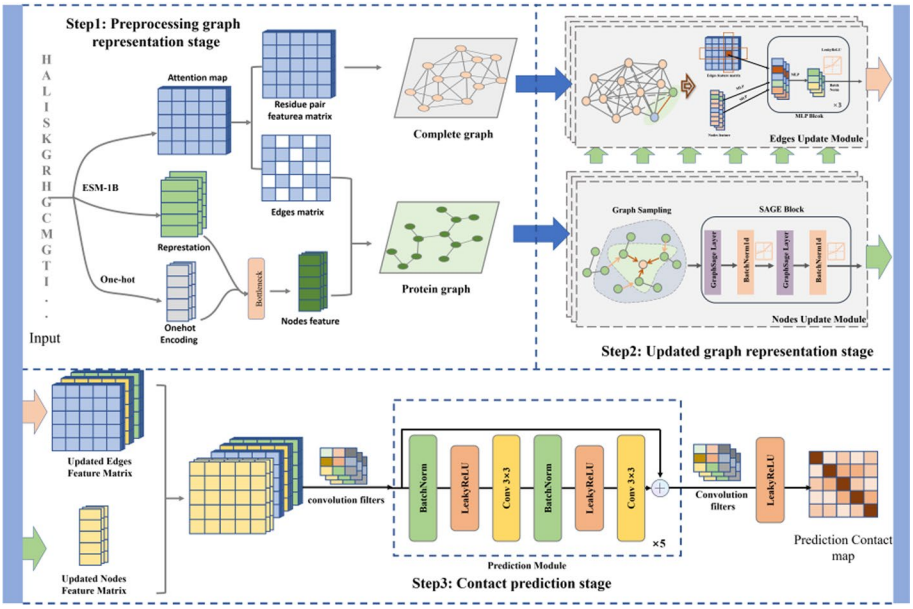

Fig. 1 The frame of PCP-GC-LM. Including preprocessing graph representation stage, updated graph representation and contact prediction layers

edges feature. The structure of the calculation model is dual graph convolutional neural network and convolutional architecture which adopts graph neural network to update graph representation and convolutional neural network to predict contact map. The proposed model takes protein sequence as the input, and the overall frame diagram is shown in Fig. 1. The frame of PCP-GC-LM mainly includes preprocessing graph representation, updated graph representation, and contact prediction layer:

(1) Preprocessing graph representation stage: According to the protein sequence, can split two different dimensional features: one-dimensional residues sequence features including one-hot code(21-dimensional) and representation (from Esm-1B) and two-dimensional residues pairs features (mostly was attention map from Esm-1B). The one-hot encoding and representation are combined through a bottleneck layer to form the nodes feature matrix, while the attention map serves as the edges feature matrix. In the graph conversion stage, the attention map is processed by MLP to obtain an edge matrix, which represents the connection between different nodes in the protein graph.

(2) Updated graph representation stage: Mostly including edges update module and nodes update module. Edges update module including aggregation operation and MLP, the purpose of which is to update the edges feature matrix. In the aggregation operation, each residues pair's edges feature and two residues feature concatenated together are sent into MLP to update the edges feature. The nodes update module includes GraphSage and MLP, the purpose of which is to update the nodes' feature matrix. Each node connects its node feature with those of neighboring nodes through the graph neural network and updates the node feature.

(3) Contact prediction stage: After passing through the update module, the edges feature matrix and the nodes feature matrix are concatenated and sent into the 2d convolution module to predict the contact map.

Preprocessing graph representation

Most of the existing deep learning-based protein contact prediction methods compare the input amino acids to a paragraph of text processed by natural language or a picture in machine vision. However, protein is not one-dimensional or two-dimensional data, it has a specific structure, and its structure largely determines its properties, making them more complex than one- or two-dimensional data. For now, there are few methods to deal with the topological structure information of proteins, so facing this situation, it is necessary to develop methods that can effectively deal with the topological structure information of proteins. To overcome this challenge, we propose to introduce the topological structure of proteins based on existing methods, which will enable us to better understand the complex relationships between protein structure and function.

The input of PCP-GC-LM can be divided into two parts: the protein residues feature and the residues-residues pair feature. In our design, the residues feature includes one-hot encoding and the representation feature of Esm-1b, and the residues-residues pairs feature is mainly are attention map matrix from the last layer of the Esm-1B. In natural language processing (NLP), the attention matrix usually represents the degree of correlation between different words in a statement and other words. Similarly, in the protein language model, the attention matrix can also be expressed as the degree of correlation between different residues. On this basis, we believe that correlation may be related to the interaction information of protein residues. Therefore, we take the attention matrix generated by the pre-training model as input in the preprocessing stage to transform the edge connection features of the graph. In preprocessing graph representation stage of our method, we will change the protein sequence input into different graph data by the dimension of input feature and their representation significance.

Considering the protein sequence feature is sequential and it is difficult for us to get the edge bonds between each residue, we first generate the complete graph, in which every node is connected with all other nodes. Due to the complete graph may contain the possible adverse effect of some nodes-nodes connect and the strong fitting ability of deep learning, it is necessary to construct the edges matrix to dismay adverse effect and fitting real connect, that is, to generate the protein.

In generating the edge matrix of the protein graph, we design two different ways to link different residues pair:

1. The first way is to connect the residues by the position of the residues on the protein sequence, connecting the residue pairs separated by less than d_seq residues on the protein sequence(the d_seq set 10 in our method).

2. Another approach to processing the attention matrix is by using a Multilayer Perceptron (MLP) to obtain the interaction matrix. In this method, the MLP generates an interaction matrix where the top 40 items with the highest probability value in each row are selected as connecting edges, and their corresponding values are set to 1. On the other hand, disconnected edges are assigned a value of 0.

The detail of the protein edges connects was shown in Fig. 2 and formula (1), the orange connections represent residue pairs in the protein sequence that are separated by fewer than d_seq residues, and the blue connections represent residue pairs in the processed interaction matrix that have a value of 1. And i, j represent the subscripts of residues at different positions in the protein sequence, respectively.

$$Connect_{i,j} = \begin{cases} |i-j| = d\_seq \\ Interaction\_map_{i,j} = 1 \end{cases} \tag{1}$$

Model architecture

After passing through the preprocessing graph representation module, two graph data are obtained, namely, the protein graph containing residue sequence features and the complete graph containing residue pair feature matrix. Therefore, to update different features respectively, this module is provided with two branch update- modules: node feature update module and edge feature update module. The protein graph is called g_p and the fully connected graph is called g_c here.

Here, the protein graph can describe a graph G(V_p, E_c),each node $v_i \in V_p$ represents amino acid and amino acid feature $h_i \in R^{1300}$ containing one-hot encoding and representation from ESM-1b, E_c represents the adjacency matrix, defined by the residue pair feature matrix after MLP. In the branch of protein graph, the mainly operations

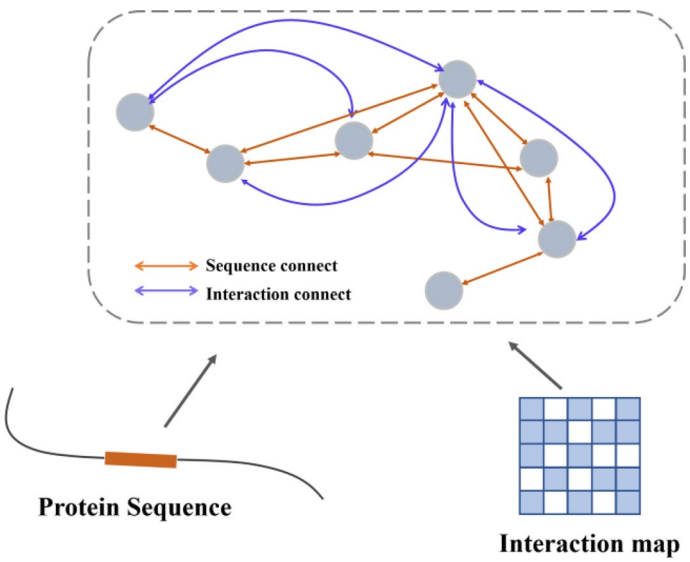

Fig. 2 Flow diagram of protein edge connects

are graph neural network and full connection layer. The most important thing for a graph neural network is the aggregation function, that is, for graph data, it is proved how to update the node features after aggregating all the adjacent node features and aggregating all features. Among them, GraphSage [29], as a widely used advanced graph neural network in recent years, is a general inductive generation framework, and its most obvious feature is that it can efficiently generate node embedding features for data that have never appeared before through existing node features. In the GCN (Graph Convolutional Network) model before GraphSage, the whole graph training method was adopted, that is to say, all nodes had to be updated in each iteration. If faced with a large graph, this training method would require a lot of resources and time, and the effect was very poor for the big graph. Therefore, the neighboring nodes in the GraphSage algorithm aggregate a certain number of neighboring node features by random sampling, and not only aggregate the neighboring node features of the same level but also aggregate the node features of multi-level levels to generate the embedded value of the target node. Finally, the embedded value of the target node is input into the fully connected network to get the predicted value of the target node. The GraphSage paper provides various aggregation functions to handle the features of aggregation nodes. These include average aggregation, pooled aggregation, and LSTM aggregation functions. While the difference between average and pooled aggregation lies in the order of averaging and nonlinear transformation, and the effect is not much different. As for the LSTM aggregation function, although it does not satisfy the permutation invariance and the calculation amount is one order of magnitude higher than the first two, it will be considered an aggregation function because it has more parameters. Given that some amino acid sequences in this study exceed 1000, so the first average aggregation function is adopted to deal with the node features after polymerization. Sampling the node feature randomly from GraphSage and calculating the feature of K-order neighbor nodes of the current node instead of considering the global nodes of the graph, which also makes GraphSage the ability of inductive learning.

The nodes feature the representation from ESM-1b and one-hot encoding as the initial input of the MLP. In GraphSage, the adjacent node features are aggregated to obtain multi-node feature mapping, which passes through a fully connected layer after the average operation and splicing with the node's features. The specific formula is as formula (2) and formula (3):

$$h\_mean_i^l = Mean(W_j^l * h_j^l), \forall j \in N(i) \tag{2}$$

$$h_i^{l+1} = ReLU(W_i * (h_i^l || h\_mean_i^l)) + h_i^l \tag{3}$$

where i stands for subscript, Mean stands for average operation, l stands for the number of layers in the graph neural network, N(i) stands for all nodes connected with i-th node after random sampling, Formula (2) shows that the embedded value of i-th node is obtained after sampling aggregation average, and in Formula (3), W represents learnable weight matrix, || represents concatenation operation, and ReLU is a activation function.

After feature processing, the embedded value of the i-th node is added to the original node features, Similar to the residual neural network, the next part of

gradient feedback is reserved in each deep learning training to prevent over-fitting and gradient disappearance.There are three layers of graph convolution layers in the branch of the protein graph, in which the fully connected layers are connected before and after the GraphSage of each layer, and activation functions and regularization operations are used to improve the feature expression. The overall node update module is shown in the following Fig. 3.

In the branches of the complete graph, the graph can describe $G(V_p, E_c)$, which the V_p represents updated amino acid feature from the protein graph, and take the attention map from Esm-1B as the initial edges feature. The edge feature updating module mainly acts on the branches of the complete graph, to update the feature matrix of the remaining pairs. In the fully connected graph, each node represents a residue in the amino acid sequence, the edge connected with other nodes represents the interaction between residues, and the edge features represent the characteristic information of the interaction between residues. In this module, after each node updates the module, the fully connected graph will aggregate the node features and the remaining pair of feature matrices. For multi-dimensional features in the fully connected graph, to explore the interaction between features, it is necessary to compress the node features and node interaction features in the aggregation function. The specific formula is formula (4)

$$e\_aggregation_{i,j}^{l} = W_{i,j}^{l}\left(\sigma\left(W_i * v_i^{l}\right)\left\|e_{i,j}^{l}\right\|\sigma\left(W_j, v_j^{l}\right)\right) \tag{4}$$

where $e\_aggregation_{i,j}^{l}$ represents the edge features of node pairs with subscripts i and j in the l-th layer, w_i, w_j and $W_{i,j}^{l}$ represent trainable parameters, and the full connection transformation is performed on node i and node j respectively to extract feature maps, and σ represents the activation function. To further mine the feature information of residue pairs, the feature matrix after each update will be spliced with the feature matrix of the upper layer to expand the dimension of features, and formula is formula (5). After two edge feature updates, a 276-dimensional residue pair feature matrix is formed. The specific edge feature update module is shown in the following Fig. 4.

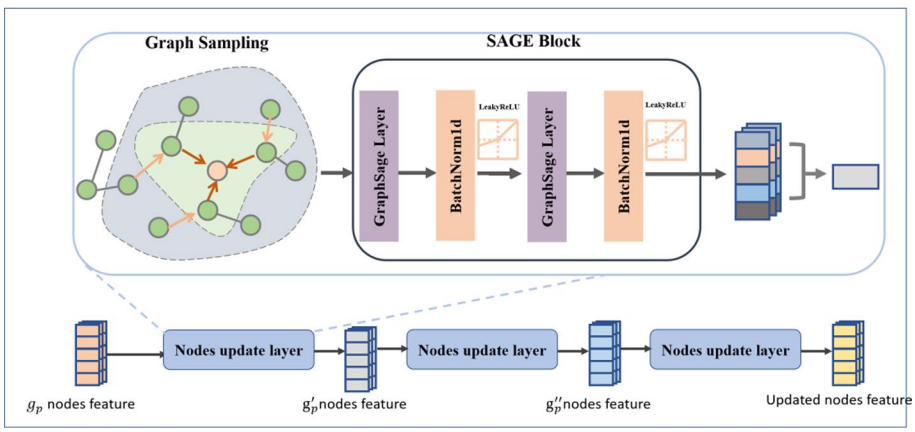

Fig. 3 Nodes update Module

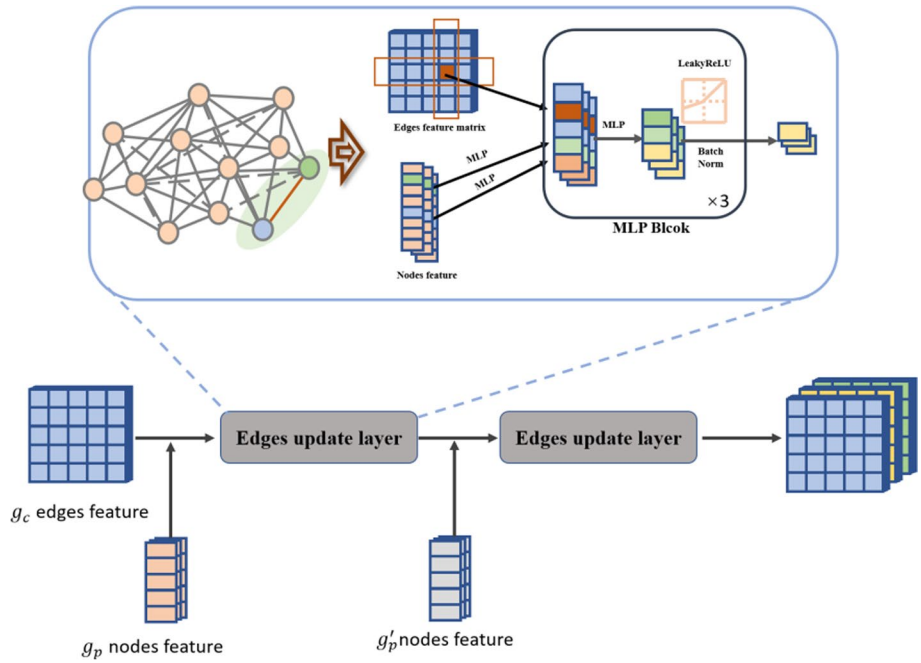

Fig. 4 Edges update Module

$$e_{i,j}^{l+1} = \left(W_{i,j} * \left(e_{i,j}^{l} \middle\| e\_aggregation_{i,j}^{l}\right)\right) \middle\| e_{i,j}^{l} \tag{5}$$

The final prediction module comprises two convolution filters and a convolution module that work together to generate accurate predictions. The input to this module consists of the updated residues sequence feature and the residue pairs feature matrix, which are obtained through a series of update steps. When the input is received, the first convolution filters come into play, and their primary function is to fuse the features of two different modes. This is achieved by smoothing the features through the convolution kernel, which helps to eliminate noise and improve the accuracy of the predictions. The resulting features are then sent to the convolution module, which further refines them and prepares them for the final stage of the prediction process. The last convolution filters are responsible for predicting the residues' contacts, which is a critical step since accurate predictions are essential for understanding the structure and function of proteins. These filters use the refined features generated by the convolution module to generate predictions, and the resulting output is a contact map that provides valuable insights into the protein's structure and function. In the latter part, we continue to predict the protein distance. The main body model is essentially consistent with the contact prediction model, except that different output functions are added to the final convolution filter to generate different outputs.The detail of the prediction module in our method is shown in Fig. 5.

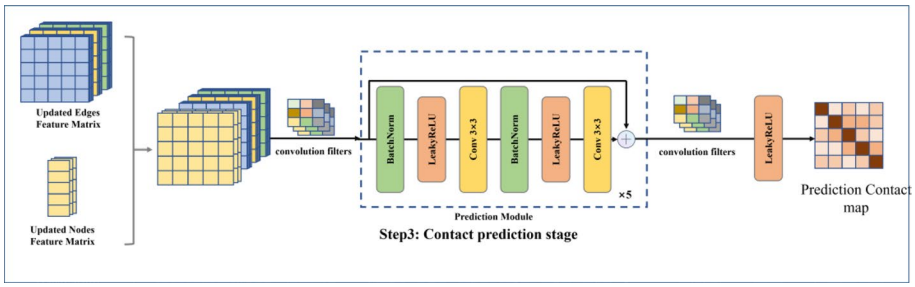

Fig. 5 Prediction Module

Results

Datasets

Because our method takes protein sequences as input, it is reasonable to train and compare all methods on the protein database without homologous information. In the SPOT-Contact-LM work, they selected the ProteinNet [30] dataset for trainnig, using a sequence identity cutoff of 95% to minimize redundancies and obtain sequences with maximum diversity. To effectively reduce the possible over-fitting, they separated 100 proteins from the ProteinNet set and compared them with other HMMS in protein [31]. When the E-value cut-off value is less than 0.1, the length is more than 500, and the final training and validation sets have 34 691 and 88 proteins, respectively. To make a better prediction on the long protein sequence, we remove the proteins with sequence lengths less than 100 on Validation Set and then predict this Validation-100 Set. Furthermore, for the test set, they also put forward the SPOT-2018 strict test set which has 669 proteins.In addition, we used other independent test sets like Casp14-FM and Casp15 target proteins.

Although homologous-dimer proteins were not used as training in the proposed method, two independent test sets (DeepHomo test set and CASP-CAPRI test set) were also used to predict the performance of PCP-GC-LM to test the contact between complex proteins in the proposed method. DeepHomo test set only includes C2-symmetric homodimeric complex structures that have less than 30% sequence identity. And CASP-CAPRI dataset includes 28 homodimers from the target from the recent CASP-CAPRI competition.

Performance evaluation

Our research aims to predict which residue pairs in a protein are in contact. So in the Critical Assessment of Structure Prediction (CASP) [32] definition, residues are contacted when there is an inter-distance of 8 Å. The contact between two residues can be divided into three types: short-range (7–11 residues apart), medium-range (12–23 residues apart), and long-range (at least 24 residues apart). For each contact type, we also calculate the top L/k highest-ranked predictions of precision in the model which the L is the length of the protein sequence, and the k is usually to be 1, 2, 5, 10.

Method comparison

In our study, we will be comparing our protein contact prediction model with other existing models that are based on single-sequence protein data. One of the models that we will be comparing is Esm-1b, which is an excellent language model that has been trained on a large-scale protein dataset. This model is capable of providing sequence embedding for many downstream tasks, including protein contact prediction. Considering that the number of existing protein contact prediction models based on a single sequence is insufficient, we also compare with Esm-1b in some datasets, although it may not be fair. Another model that compares is SSCpred, which is a single-sequence-based contact predictor that performs prediction through the deep fully convolutional network (Deep FCN) without additional homology information. We have chosen the SSCpred code in an offline version, which can be downloaded from https://github.com/chenmc1996/SSCPred. In addition to these models, we will also be comparing our network with SPOT-Contact-LM and its different sub-models. SPOT-Contact-LM is a combined network that obtains different prediction results through different inputs and training strategies and averages the different prediction results. We will be comparing our network with the different sub-models under the SPOT-Contact-LM, especially when choosing a training strategy that is direct inter-residue contact prediction. Based on the size of the input feature and training strategy in the sub-model of SPOT-Contact-LM, it is named points SPOT-Contact-1 to SPOT-Contact-6, respectively. To provide a comprehensive comparison of the performance of various methods and the parameters of different models, we will be comparing our model with other models in the validation-100 set and other test datasets. Furthermore, we will also be comparing our model with DeepHomo and PGT in the DeepHomo test set and CASP-CAPRI test set. By comparing our model with these existing models, we aim to demonstrate the effectiveness and accuracy of our protein contact prediction model.

Performance comparison

Feature and dual-graph importance

To understand the influence of different input features and the branches of a dual graph, we also trained different models and compared their performance on the test datasets. Table 1 outlines the different models that we utilized in our study. Our findings, as depicted in Fig. 6, indicate that the models trained using one-hot code had a prediction accuracy of 13.8% and 12.2% for the top L/2 type in medium- and long-range contact.

Table 1 Input feature vector composition from ablation experiments

Model	Input	Graph
Model 1	ESM-1b attention map (last layer only) + one-hot encoding	Protein graph + Complete graph
Model 2	ESM-1b attention map (last layer only) + one-hot encoding + ESM-1B representation	Only protein graph
Model 3	ESM-1b attention map (last layer only) + one-hot encoding + ESM-1B representation	Only complete graph
PCP-GC-LM	ESM-1b attention map (last layer only) + one-hot encoding + ESM-1B representation	Protein graph + Complete graph

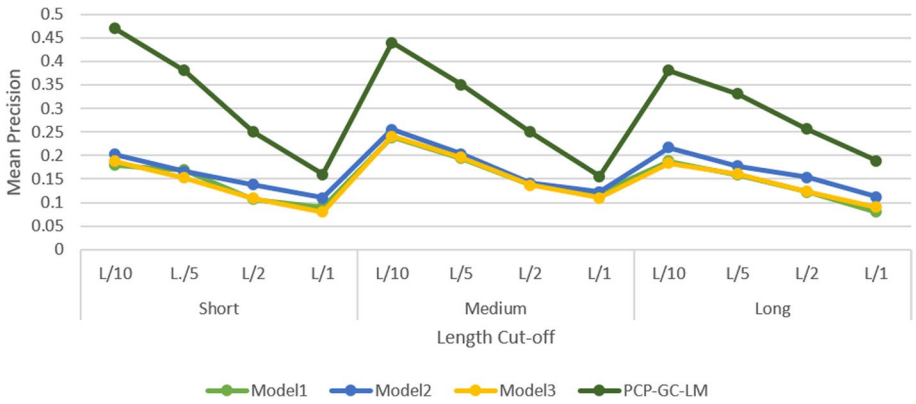

Fig. 6 The mean prediction accuracy in SPOT-2018 set for Short, Medium and Long range contact

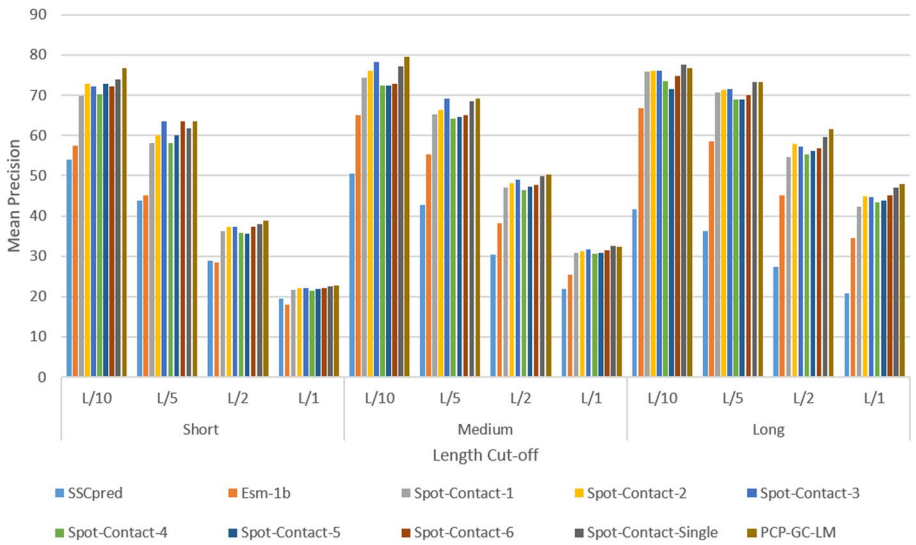

Fig. 7 Comparison of our method and other methods on Validation-100 set for short-,medium- and long-range contacts

However, when we trained the model by adding representation to the one-hot code, we observed a significant improvement of 58.5% and 74.5% respectively. We also examined the effects of individual graphs in the graph encoder module. The only protein graph (PG) refers to the edges matrix generated solely by MLP from the edges feature matrix, without going through the edges update layers. On the other hand, the only complete graph (CG) does not generate the edges matrix. Our results, as shown in Figure 6, demonstrate that GP performs better than PG in most contact types. As a result, we have chosen one- hot encoding and representation from Esm-1b as inputs and adopted the dual-graph updating module as our best model.

Result comparisons

First, we evaluated the performance of our model on a validation set and compared it with other existing methods. Our model outperformed ESM-1b and spot-contact-LM and their sub-models in terms of three contact types and top/k (k = 1, 2, 5, 10). Figure 7

Table 2 Comparison of our method, sub-model of SPOT-Contact-LM, and ESM-1b on SPOT-2018 set for medium-, long- contacts

Model	Medium-range				Long-range			
	L/10	L/5	L/2	L/1	L/10	L/5	L/2	L/1
SPOT-Contact-1	39.17	31.84	21.75	14.84	35.14	30.34	22.75	17.02
SPOT-Contact-2	40.03	32.83	22.49	15.26	36.03	30.92	23.75	18.13
SPOT-Contact-3	42.03	34.38	23.32	15.65	38.75	33.23	25.22	18.94
SPOT-Contact-4	38.52	31.53	21.86	14.85	35.32	30.19	22.80	17.20
SPOT-Contact-5	40.16	32.85	22.13	14.93	37.34	31.82	23.87	17.77
SPOT-Contact-6	40.52	33.37	22.59	15.33	37.43	32.22	24.31	18.44
SPOT-Contact-LM	42.43	34.41	**23.63**	**15.88**	**39.60**	**34.35**	**25.94**	**19.62**
PCP-GC-LM	**44.44**	**35.05**	25.08	15.68	38.09	33.04	25.63	18.98

Bold values denote the most outstanding performing values for that category within the table

shows a detailed comparison of the validation set. We also analyzed the effect of different sub-modules of SPOT-Contact-LM and found that the performance improved with the increase of input features. For the long-range and top L/1 contact type, the mean prediction accuracy of spot-contact-1 to spot-contact-6 were 0.423, 0.448, 0.446, 0.433, 0.438, and 0.45, respectively, while the mean prediction accuracy of our method was 0.478. We also tested our model on the SPOT-2018 test set and compared it with other methods. Table 2 shows that our method performs much better than ESM-1b and the sub-model from spot-contact-LM with long-range type contact precision for length cut-offs of L/1, L/2, L/5, and L/10. In the L/1 of Long-range contact type, PCP-GC-LM was 11.51%,4.68%,0.21%,10.34%,6.8%, and 2.92% higher than Spot-Contact-1 to 6, respectively. Our method also improved the performance measures for other contact types compared to these sub-models of SPOT-Contact-LM. However, there is still room for improvement when compared with the final result of SPOT-Contact-LM. We believe that further research can help us to improve the accuracy of our model and enhance its performance in predicting protein contact maps.

During the 14th edition of the Critical Assessment of Protein Structure Prediction (CASP14), several free modeling targets were released, and various methods were employed to predict inter-residue full-length contacts. For these methods, the simple sub-model from SPOT-Contact-LM achieved a contact precision of 0.1194 in Top L/2, while the SPOT-Contact-LM achieved 0.154. However, our method outperformed these methods by achieving a mean precision of 0.16. In the latest protein targets of CASP15, the average prediction accuracy of spot-contact-1 to spot-contact-6 in the long-range l/10 type was found to be 55.49%, 62.1%, 62.47%, 61.05%, 63.35%, and 63.04%, respectively. The final model of spot-contact-LM achieved a precision of 63.81%. our method outperformed all these methods by achieving a precision of 64.74%. These results demonstrate the effectiveness of our method in predicting inter-residue full-length contacts accurately. In summary, our method has proven to be a reliable and effective approach for predicting inter-residue full-length contacts in protein structure prediction. The details of the precision of different methods are shown in Figs. 8 and 9.

The SPOT-Contact-LM sub-model has been designed with two strategies: direct inter-residue contact prediction and inter-residue distance bin prediction. In order

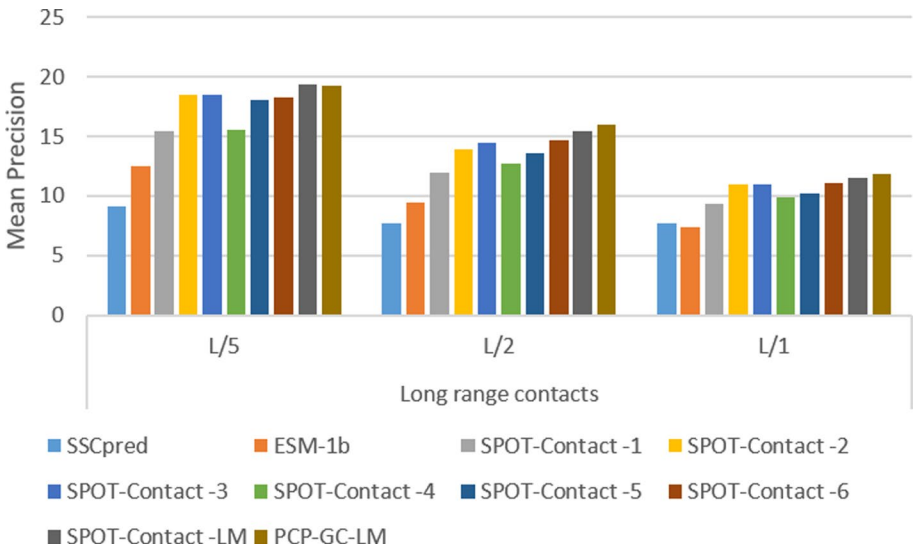

Fig. 8 Precision-based comparison of SPOT-Contact-LM and our method on CASP14-FM set for long range contact

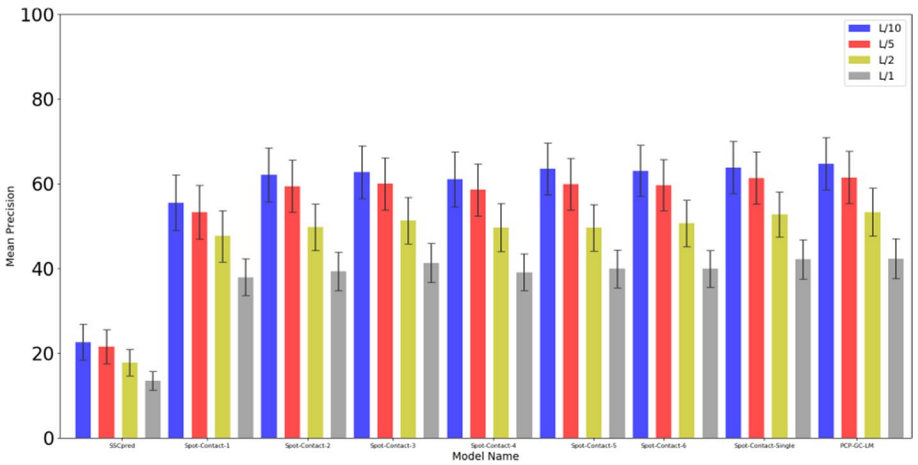

Fig. 9 Precision-based comparison of SPOT-Contact-LM and our method on CASP15 target set for long range contact

to evaluate the effectiveness of these strategies, we conducted a comparison with our own method using validation databases. Our findings suggest that the direct inter-residue contact prediction strategy performs marginally better than the interresidue distance bin prediction strategy. However, when it comes to long-range predictions, our method outperformed the direct contact prediction strategy. The difference in performance between the two strategies was found to be smaller for medium-range predictions. To illustrate this, we have included precision comparisons of the two training strategies and our method for medium- and long-range predictions in Fig. 10. Overall, our results suggest that our method is more effective than the direct contact prediction strategy, particularly for long-range predictions.

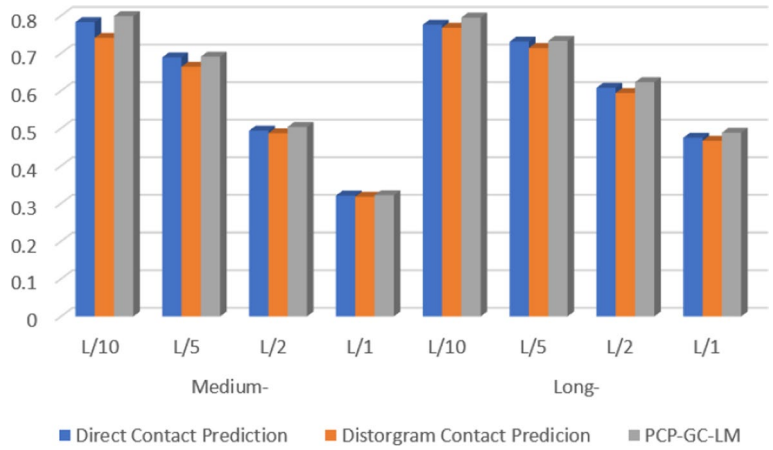

Fig. 10 precision comparison of two training strategies for medium-range, and long-range on the Validation-100 set

Table 3 Comparison of the parameters and running time between our model and the other two model

Model	Number of parameters (M)	Time (s)
SPOT-Contact-LM	14.13	74
SSCpred	5	128
PCP-GC-LM	4.79	66

Table 4 Precision on the DeepHomo test DataSet (300)

	Methods		
threshold	DeepHomo	PGT	PCP-GC-LM
L/10	52.1	67.33	**78.78**
L/5	47.8	64.71	**75.25**
L/2	39.4	60.29	**65.33**

Bold values denote the most outstanding performing values for that category within the table

Furthermore, the comparison of the parameters and running time between our model and the other two models, SSCpred and Spot-Contact-LM, is presented in Table 3. It is evident that our model has fewer parameters and requires less running time compared to the other two models. This indicates that our model is more efficient and computationally less expensive. This is a significant advantage, especially when dealing with large-scale protein structure prediction tasks. Our experiment employs the Ubuntu operating system and utilizes the Pytorch deep learning development framework. The central processing unit (CPU) used is the 11th Gen Intel(R) Core(TM) i9-11900K, while the graphics processing unit (GPU) employed is the Nvidia GeForce 3090Ti.

Performance in complex protein

Proteins are essential biomolecules that play a crucial role in various biological processes. They carry out their biological functions by interacting with other biomacromolecules, such as DNA, RNA, and other proteins. In particular, protein-protein interactions are critical for the formation and function of protein complexes, which are involved in many cellular processes, including signal transduction, gene regulation, and metabolic pathways. However, predicting the three-dimensional structure of protein-protein complexes remains a major challenge in structural biology. This is because the structure of a protein complex is determined by the interactions between individual subunits, which can be highly complex and dynamic. Therefore, some work has focused on interchains contact prediction to help predict protein complex structure due to the importance of residues to residues interactions between individual subunits of protein complexes.

The results of our study, as presented in Table 4, demonstrate that our model is highly effective in predicting contacts in complex proteins. Specifically, in the

Table 5 Precision on the CASP-CAPRI datasets

	Methods			
threshold	DeepHomo	Glinter	PGT	PCP-GC-LM
L/10	46.98	50.54	54.56	**57.51**
L/5	44.11	48.09	49.48	**50.59**
L/2	36.95	41.9	42.55	**45.33**

Bold values denote the most outstanding performing values for that category within the table

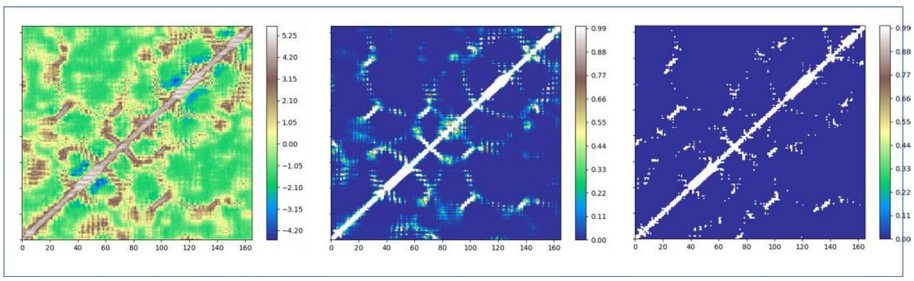

Fig. 11 Comparison of the outputs for 2XZ4 protein by PCP-GC-LM as labeled. In these three columns, the left side is the edge matrix, the middle is the final output from PCP-GC-LM, and the right is the true contact from example protein 2XZ4

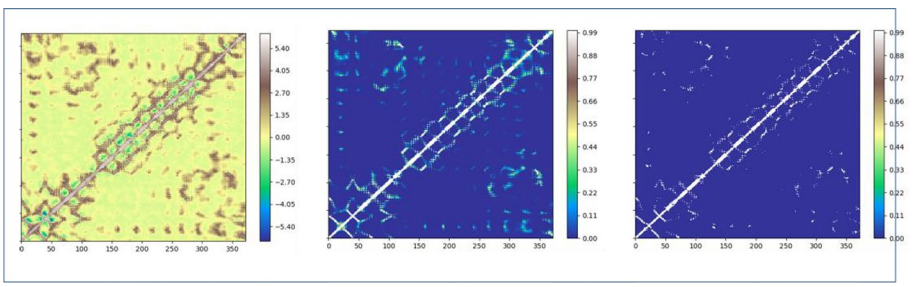

Fig. 12 Comparison of the outputs for 2GGE protein by PCP-GC-LM as labeled. In these three columns, the left side is the edge matrix, the middle is the final output from PCP-GC-LM, and the right is the true contact from example protein 2GGE

Deephomo test dataset, our model achieved a contact prediction accuracy of 78.78 in the top/10, which is significantly higher than PGT's accuracy of 67.33. This indicates that our model outperforms PGT in terms of predicting contacts in complex proteins. Furthermore, our model also outperformed other methods in the CASP-CAPRI dataset, as shown in Table 5.

Performance analysis of edges matrix

To compare the effects of interaction matrices generated by intermediate components and construct heatmaps for observation, we can analyze the heatmaps of real contacts on two example proteins, namely 2GGE and 2XZ4, and compare them with the edge matrices generated by PCP-GC-LM. This allows us to visualize the real contacts in these proteins through the interaction matrix heatmap. The results for these example proteins are shown in Figs. 11 and 12, respectively, and each set of images consists of three different pictures.

The left image represents the connection matrix obtained through our model, and the heatmap quantifies the interaction relationship of residue pairs at different positions. The darker the color means the higher the vector value, indicating a higher likelihood of interaction of different positions in the protein sequence. And the middle image represents the final protein contact prediction matrix from our model, while the image on the right represents the real protein contact matrix generated by our model. In both heat images, the closer the color is to white, the higher the probability of contact, whereas closer to blue indicates a lower probability of contact.

Performance in protein distance prediction

Based on this model, we have enhanced the architecture to predict the distance between proteins. Unlike contact prediction, which tackles the probability of contact between residues and can be approached as a regression or binary classification problem, distance prediction involves categorizing distances into multiple bins, making it a multi-classification task. In our protein distance prediction model, the overall structure is

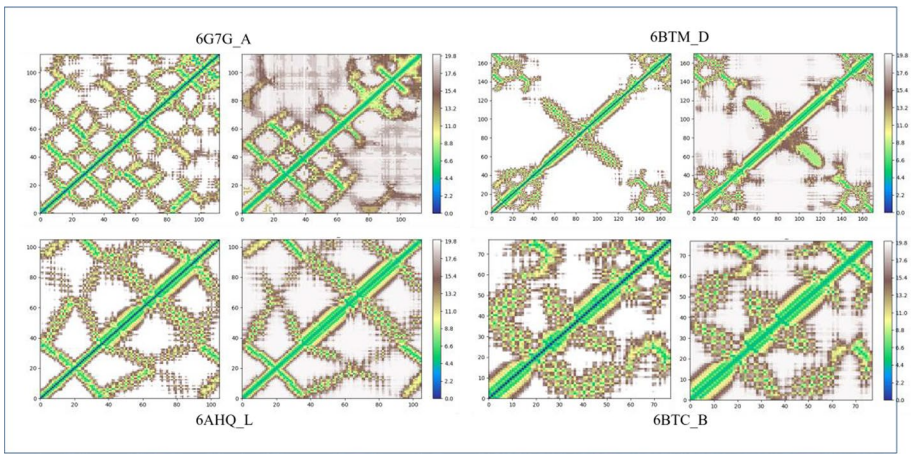

Fig. 13 Comparison of the outputs for these proteins by PCP-GC-LM as labeled. In different combinations, the left side is the real protein distance distribution and the right side is the predicted distance distribution by our distance model

similar to the contact prediction model, with the only difference being the utilization of the SoftMax function instead of the LeakReLU function, as distance prediction involves multiple classes and improves the depth of our model. We have selected four target protein domains (such as 6G7G_A, 6BTM_D, 6AHQ_L, 6BTC_B) for predicting protein distances. Figure 13 illustrates a specific comparison diagram, where each combination show cases the actual distance on the left side and the predicted distance on the right side. The distance range is limited to [0, 20]. The distance map reveals that the distance prediction matrix offers more spatial information than the contact matrix. The whiter the color distribution in Fig. 13, the farther the distance between residues, while a shift towards green and blue indicates a closer distance between residues.

Discussion

In this research, we have introduced a novel approach for predicting protein contacts using a single-sequence model that combines graph neural networks (GNN) and convolutional neural networks (CNN). The proposed method has been shown to significantly outperform other recent single-sequence-based contact prediction models. The study also examines the impact of dual-level graphs and input features on the performance of the proposed model. The authors found that the lack of more expressive sequence features in the input has the greatest influence on the model's performance. Therefore, incorporating more effective and expressive sequence features can potentially improve the accuracy of the model's predictions. Additionally, the study shows that the dual-graph model is effective, as only the complete graph has the worst effect on the model's performance. The proposed model's architecture involves a preprocessing graph representation stage, updated graph representation stage, and prediction stage that preprocessing graph representation stage the protein sequence information into a graph representation and update graph representation by different branches of updated modules. This graph representation is then fed into the prediction module, which extracts the relevant features for predicting the protein contacts. We have evaluated the performance of their proposed model on several benchmark datasets and compared it to other state-of-the-art methods. The results show that the proposed model achieves higher accuracy and outperforms other single-sequence-based contact prediction methods. Overall, this study provides a new approach for predicting protein contacts using a single-sequence model that combines GNN and CNN. The proposed method has been shown to significantly improve the accuracy of protein contact prediction and has the potential to be useful in various applications, such as protein structure prediction and drug discovery.

Conclusion

Our experiment shows that protein input and the connection of the graph are all essential for improving contact prediction. In our method, we use a graph neural network to get potential information from the output of the protein pre-training language model, which effectively improves the accuracy of prediction, but there is still a lot of room for improvement. In the following-up work, we have the following improvement ideas:

(1) Explore other possible features. In the protein sequence feature, SPOT-Contact-LM added more structure probability feature to its input feature like three-state-secondary-structure (SS3) [33] and eight-state-secondary structure (SS8) [34], solvent accessible surface area (ASA) [35] and protein backbone torsion angle and so on from SPOT-1D-Single [36]. We can also consider using more features such as more expression residues as input to improve prediction accuracy. In the residues-residues feature, mostly attention map from Esm-1b. In our early thoughts, we also want to use more attention map layers, but due to the shortage of our own hardware, we only use all the features of the last layer of attention heads. So, after this, we can try these ideas.

(2) Graph neural network selection. In the process of transmission and aggregation of a graph neural network, all neighboring nodes of a node are treated equally. However, it should also be considered that not all neighboring nodes are equal, so different neighboring nodes can be given different weights. Graph attention network (GAT) can learn the structure representation of nodes by considering the weight and relation embedding of neighboring nodes. So we can use GAT instead of GraphSage in the graph encoder.

(3) Using a more powerful multi-scale network Res2Net [37]. ResNet, as a prediction module, is very important for generating high-precision contact maps, so the use of a more novel and efficient convolution model is also the focus of the improvement.

Acknowledgements
No applicable.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by J. OuYang, Y. Yang, and Y. Gao. All authors reviewed the manuscript.

Funding
This research has been supported by Key Projects of the Ministry of Science and Technology of the People Republic of China (2020YFC0832401).

Availability of data and materials
The CASP14 and CASP15 datasets used in this study and the source are availabel at https://predictioncenter.org/download_area/CASP14/targets/ and https://predictioncenter.org/download_area/CASP15/targets/. The DeepHomo test set and the CASP-CAPRI test set are available at http://huanglab.phys.hust.edu.cn/DeepHomo/ and can downloaded by http://huanglab.phys.hust.edu.cn/DeepHomo/download/DeepHomo_testset.tgz. The train-set list and validation-set list can be downloaded from https://github.com/619yong/PCP-GC-LM/tree/master/dataset.

Declarations

Ethics approval and consent to participate
No applicable.

Consent for publication
The authors declare that they consent to publication.

Competing interests
The authors declare no competing interests.

Received: 4 May 2023 Accepted: 22 August 2024
Published online: 02 September 2024

References

1. Zheng W, Li Y, Zhang C, Pearce R, Mortuza S, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins Struct Funct Bioinform. 2019;87(12):1149–64.
2. AlQuraishi M. Machine learning in protein structure prediction. Curr Opin Chem Biol. 2021;65:1–8.

3. Joshi RS, Jagdale SS, Bansode SB, Shankar SS, Tellis MB, Pandya VK, Chugh A, Giri AP, Kulkarni MJ. Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. J Biomol Struct Dyn. 2021;39(9):3099–114.
4. Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, Wei Z. Drug–target affinity prediction using graph neural network and contact maps. RSC Adv. 2020;10(35):20701–12.
5. Lai B, Xu J. Accurate protein function prediction via graph attention networks with predicted structure information. Brief Bioinform. 2022;23(1):502.
6. Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics. 2022;38(1):125–32.
7. Song B, Luo X, Luo X, Liu Y, Niu Z, Zeng X. Learning spatial structures of proteins improves protein–protein interaction prediction. Brief Bioinform. 2022;23(2):558.
8. Buchan DW, Jones DT. EigenTHREADER: analogous protein fold recognition by efficient contact map threading. Bioinformatics. 2017;33(17):2684–90.
9. Fukuda H, Tomii K. DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. BMC Bioinform. 2020;21(1):1–15.
10. Fukuda H, Tomii K. Deep neural network for protein contact prediction by weighting sequences in a multiple sequence alignment. bioRxiv, 2018. p. 331926.
11. Valdez R, Roig K, Pinto-Roa DP, Colbes J. Analysis of protein contact prediction by deep learning algorithms in CASP13.
12. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol. 2017;13(1):1005324.
13. Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in casp12. Proteins Struct Funct Bioinform. 2018;86:67–77.
14. Li Y, Zhang C, Bell EW, Zheng W, Zhou X, Yu D-J, Zhang Y. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS Comput Biol. 2021;17(3):1008865.
15. Chen M-C, Li Y, Zhu Y-H, Ge F, Yu D-J. SSCpred: single-sequence-based protein contact prediction using deep fully convolutional network. J Chem Inf Model. 2020;60(6):3295–303.
16. Singh J, Litfin T, Singh J, Paliwal K, Zhou Y. SPOT-contact-Lm: improving single-sequence-based prediction of protein contact map using a transformer language model. Bioinformatics. 2022;38(7):1888–94.
17. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. Biorxiv. 2020. 2020–12.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. Pp. 770–778.
19. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
20. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556
22. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016. arXiv preprint arXiv:1609.02907
23. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. 2017. arXiv preprint arXiv:1710.10903
24. Berg R, Kipf TN, Welling M. Graph convolutional matrix completion. 2017. arXiv preprint arXiv:1706.02263
25. Wang F, Lei X, Liao B, Wu F-X. Predicting drug–drug interactions by graph convolutional network with multi-kernel. Brief Bioinform. 2022;23(1):511.
26. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug–target interactions based on graph convolutional network and deep neural network. Brief Bioinform. 2021;22(2):2141–50.
27. Wang H, Zhou G, Liu S, Jiang J-Y, Wang W. Drug-target interaction prediction with graph attention networks. 2021. arXiv preprint arXiv:2107.06099
28. Budak C, Mençik V, Gider V. Determining similarities of COVID-19–lung cancer drugs and affinity binding mode analysis by graph neural network-based GEFA method. J Biomol Struct Dyn. 2023;41(2):659–71.
29. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Advances in neural information processing systems, vol 30. 2017.
30. AlQuraishi M. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinform. 2019;20(1):1–10.
31. Söding J. Protein homology detection by hmm–hmm comparison. Bioinformatics. 2005;21(7):951–60.
32. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005;15(3):285–9.
33. Ho C-T, Huang Y-W, Chen T-R, Lo C-H, Lo W-C. Discovering the ultimate limits of protein secondary structure prediction. Biomolecules. 2021;11(11):1627.
34. Cong P, Li D, Wang Z, Tang S, Li T. Spssm8: an accurate approach for predicting eight-state secondary structures of proteins. Biochimie. 2013;95(12):2460–4.
35. Dong T, Gong T, Li W. Accurate estimation of solvent accessible surface area for coarse-grained biomolecular structures with deep learning. J Phys Chem B. 2021;125(33):9490–8.
36. Singh J, Litfin T, Paliwal K, Singh J, Hanumanthappa AK, Zhou Y. Spot-1d-single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. Bioinformatics. 2021;37(20):3464–72.
37. Gao S-H, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr P. Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell. 2019;43(2):652–62.

J. Ouyang a professor at Xiangtan University. His research interest is bioinformatic

Y. Gao is a graduate student at Xiangtan University. His research interest is bioinformatic

Y. Yang is a graduate student at Xiangtan University. His research interest is bioinformatic