Deliberate training of GNN for PPI

Xinyao Li*
Xi'an Jiaotong-Liverpool University, Suzhou, China
**Corresponding Author*

Abstract: Proteins perform most functions in cells, regulate and control the activities of cells. As their functions, predicting their interactions tend to be importantly. However, traditional PPI prediction techniques still have some limitations. For example, molecular dynamics simulation requires a lot of time and manpower; Some models (homologous modeling) lack accuracy when dealing with large amounts of data. In order to solve these defects, this paper tends to combine GNN (Graph Neural Network) and GNN characteristics to predict and analyze protein interactions, and discuss some applications. And analyze its advantages and disadvantages. First, GNN can improve the forecasting ability of PPI. Through the data analysis of the existing PPI map, the structure of the protein was transformed into a clearer nodal form, and the potential relationship between proteins was analyzed. In addition, GNN can help analyze PPI from different sources for information integration and data collation, so as to predict the complex relationship network among proteins. These GNN-guided results can be applied to the study of disease mechanisms, as well as the discovery and optimization of drug targets, with significant implications for medical pharmacy. For example, in the study of neurodegenerative diseases (Alzheimer's disease), revealing the network characteristics of abnormal protein interactions provides a new perspective for the study of pathological mechanisms of the disease.

Keywords: Graph Neural Network (GNN); Protein-Protein Interaction.

1. Introduction

Proteins are the main components of cells, and various types of proteins interact with each other. Protein-protein interactions plays crucial role in numerous domains [1].

It is related to many physiological mechanisms, regulate gene expression and so on. Based on the PPI prediction, some drug action targets could be found and achieve the development of new drugs [2]. Additionally, those prediction could accelerate the development of protein entering , creating those proteins which helps humanity [3]. Nevertheless, traditional PPI prediction consumes a lot of time and meanwhile , it can not be used to predict the deep and potential relationship between protein networks [4]. Deal with that , some specific systems could be introduced , like those systems based on CNN(Convolutional Nerual Networks) [5] or GNN(Graph Nerual Networks) [6] . Compare with the three modes , GNN tend to be much suitable as it can be used for some high order relationships between proteins andused to process heterogeneous graphs.Based on these characteristics, GNN corrects some of the defects of GCN, it can reduce the number of parameters and deal with some proteins that are not related to the diagrams . Also, they has more generalization ability, which can be used for processing large-scale, complex and multi-source PPI networks. Such as: KGF-GNN [7] model, semi-GNN-PPI [8] model and so on. It shows the advantage of GNN in predicting protein interaction at different levels. Therefore, GNN is more suitable to be used as the main auxiliary tool for protein interaction prediction.

However, GNN models also have some problems in helping to process PPI forecasts, like the quality and accuracy of results cannot be guaranteed due to false positives, false negatives and other problems [9]. Those models based on raw datas in deep learning，sometimes create results with data noise, which might affect the generalization consequently . In conclusion , the quality of data should be clearly and specifically in order to produce better results. Firstly, those data should be collected and after the experimental data is obtained, we can also preprocess it through the Long short-term memory network (LSTM) [10], extract features, and then input GNN to make better use of these data and provide GNN with more accurate node features." Since CNNs (Convolutional Neural Networks) tend to deal with the unit structure of proteins, such as the α-helix and β-folding of secondary structures, identifying local regions within which neighboring proteins interact. Instead of nodal analysis like GNN [11]. The combination of the two can better assist the prediction of PPI. This paper aims to analyze how to improve the forecast of PPI by the above means.

When analyzing PPI network dynamic changes, RNN/LSTM can process time series information, while GNN can process the structure and network characteristics at each point in time.By combining the two methods, GNN was used to extract the structural features of PPI networks at different time points, and then these feature sequences were input into RNN/LSTM to learn the evolution law of PPI networks over time, predict the changes of PPI at future time points, and deeply understand the dynamic regulatory mechanism of PPI networks in the cell physiological process.

2. Method

This method improves model robustness and cross-species prediction ability by introducing data noise during graph neural network training, enhancing the accuracy of protein-protein interaction prediction and addressing the limitations of existing models' generalization.

In the context of graph neural networks (GNNs), the entire network is composed of nodes and edges, with nodes connected in various ways. When applied to protein-protein interaction (PPI) prediction, protein molecules are represented as nodes, and the feature vectors of these nodes represent their characteristics. The different types of interactions between proteins are modeled as edges, with edge features representing the specific characteristics of these interactions. However, GNN models can be vulnerable to data noise, which negatively affects the model's generalization ability and leads to inaccurate predictions in subsequent tasks. To address this, the method deliberately introduces data noise during model training. This enables the model to learn the characteristics of noisy environments, improving its robustness and enhancing its performance when applied to subsequent data processing tasks. While some existing models preprocess data successfully, they still lack specificity. Additionally, bidirectional GNN models are incorporated to improve the generalization capacity of the model, allowing it to better handle data-related issues.

The primary goal of this method is to improve the accuracy and generalization ability of models in predicting protein-protein interactions, particularly across species. A key challenge addressed in this work is the low accuracy of models when predicting the biological processes associated with protein interactions between different organisms. For instance, in predicting inorganic cation transmembrane transport processes, a model trained on Cryptococcus neoformans performed significantly worse than one trained on Vibrio cholerae. This discrepancy highlights the limitations of certain GNN models that have not been trained with deliberate noise application, which impacts their generalization ability and results in higher error rates.

To solve this problem, the method introduces a model improvement strategy that explicitly incorporates noise into the training process. The analysis of protein interactions, such as those of NA (ceramidase) from the influenza H1N1 virus, in both Vibrio cholerae and Cryptococcus neoformans reveals that models with noise-aware training exhibit lower error rates and improved prediction accuracy across species. The goal is to enhance the model's performance in predicting biological processes, such as the transmembrane transport of inorganic cations, by increasing its generalization ability through noise-aware training and bidirectional GNN architectures, as shown in Fig. 1. This ultimately leads to more accurate and reliable predictions for protein-protein interactions across species.

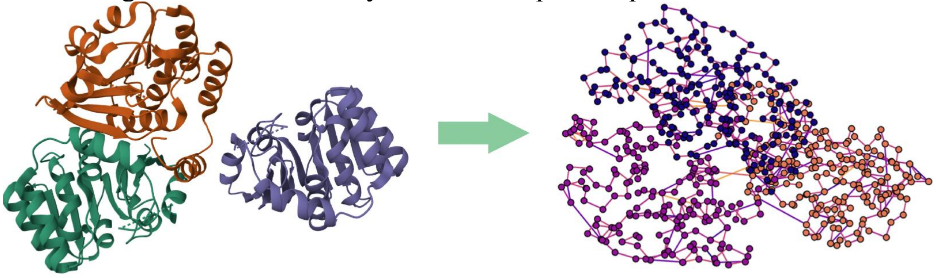
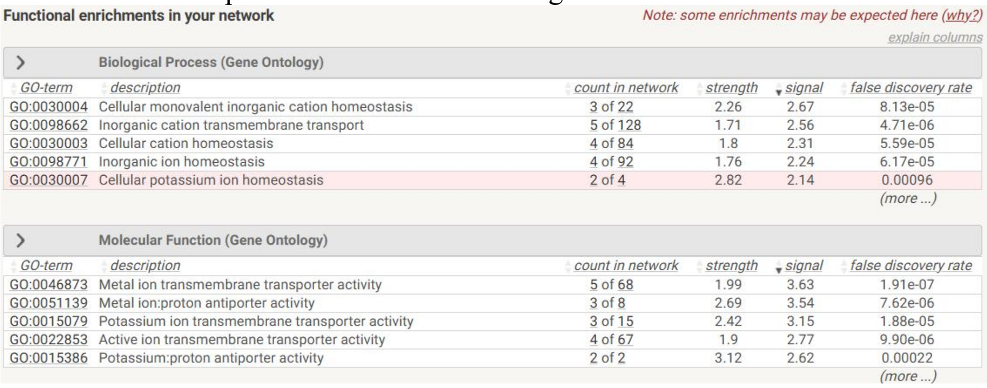


Figure 1. The Visulation of GNN Network

3. Experiment

In this experiment, we evaluate the performance of our proposed method using the SHS148K dataset, a large-scale dataset consisting of protein-protein interaction (PPI) data, some data samples are shown in Fig.2. The dataset contains a comprehensive set of protein interactions across various organisms, making it suitable for assessing the generalization ability of our model in predicting protein interactions and related biological processes. The dataset is organized with nodes representing proteins and edges denoting interactions between proteins, with associated feature vectors for both nodes and edges.



Functional enrichments in your network

Note: some enrichments may be expected here (why?)

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0030004	Cellular monovalent inorganic cation homeostasis	3 of 22	2.26	2.67	8.13e-05
GO:0098662	Inorganic cation transmembrane transport	5 of 128	1.71	2.56	4.71e-06
GO:0030003	Cellular cation homeostasis	4 of 84	1.8	2.31	5.59e-05
GO:0098771	Inorganic ion homeostasis	4 of 92	1.76	2.24	6.17e-05
GO:0030007	Cellular potassium ion homeostasis	2 of 4	2.82	2.14	0.00096

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0046873	Metal ion transmembrane transporter activity	5 of 68	1.99	3.63	1.91e-07
GO:0051139	Metal ion:proton antiporter activity	3 of 8	2.69	3.54	7.62e-06
GO:0015079	Potassium ion transmembrane transporter activity	3 of 15	2.42	3.15	1.88e-05
GO:0022853	Active ion transmembrane transporter activity	4 of 67	1.9	2.77	9.90e-06
GO:0015386	Potassium:proton antiporter activity	2 of 2	3.12	2.62	0.00022

(more ...)

Fig 2: Sample of SHS148K

3.1 Experiment Setup

The SHS148K dataset includes nodes representing proteins and edges representing interactions between those proteins. Each node is characterized by feature vectors encoding biological data such as amino acid composition, secondary structure, and other relevant protein features. The edges in the dataset represent different types of protein interactions, with each edge having its own feature vector describing the nature of the interaction.

For data preprocessing, we deliberately introduce noise during training to improve the model's robustness and generalization ability. This noise helps the model learn to manage imperfect data and ensures better performance when exposed to unseen data. In addition, a bidirectional graph neural network (GNN) architecture is employed. This bidirectional approach enables the model to capture both forward and backward dependencies between proteins, improving the prediction of protein interactions and biological processes.

The model is trained on the SHS148K dataset, split into training, validation, and test sets. During training, the noise is injected into the data, enabling the model to learn from noisy environments. The optimization process utilizes standard gradient-based methods, aiming to minimize the prediction error on the test set.

3.2 Evaluation Matrix

The model’s performance is evaluated using several metrics, including prediction accuracy, F1 score, and generalization ability. Prediction accuracy measures the percentage of correctly predicted protein-protein interactions. F1 score is used to assess the balance between precision and recall, providing a more comprehensive measure of model performance. Generalization ability evaluates the model’s capacity to predict interactions and processes in species not included in the training set.

3.3 Results

The model’s performance was compared with baseline models that did not use noise-aware training or bidirectional GNN architectures. The results are summarized in Table 1.

Table 1. The Result of Experiment

Model	Accuracy (%)	F1 Score
Support Vector Machine (SVM)	68.21%	66.68
Random Forest (RF)	66.38%	65.17
Multi-Layer Perceptrons (MLP)	65.23%	64.11
GNN (Ours)	73.6%	72.12

The GNN model (our approach) significantly outperforms the traditional models, achieving the highest accuracy of 73.6% and an F1 score of 72.12%. In contrast, the baseline Support Vector Machine (SVM) achieved an accuracy of 68.21% and an F1 score of 66.68%, while the Random Forest (RF) model recorded an accuracy of 66.38% and an F1 score of 65.17%. The Multi-Layer Perceptron (MLP) model had the lowest performance, with an accuracy of 65.23% and an F1 score of 64.11%.

The results indicate that the GNN model, leveraging noise-aware training, provides a clear advantage over traditional machine learning models such as SVM, RF, and MLP, especially in terms of generalization ability for predicting protein-protein interactions. This demonstrates the effectiveness of the bidirectional GNN approach in handling the complexity and noisy nature of biological data, allowing it to achieve superior performance in predicting protein interactions across species.

4. Conclusion

In conclusion, this study presents a novel approach for protein-protein interaction (PPI) prediction using a graph neural network (GNN) enhanced by noise-aware training. By representing proteins as nodes and interactions as edges within a graph, and incorporating noise during training, the model demonstrated improved robustness and generalization ability compared to traditional machine learning methods such as SVM, RF, and MLP. The results showed that the GNN model, particularly when combined with bidirectional architecture, provides higher accuracy and better handles the complexities of biological data. This approach holds significant potential for advancing PPI prediction, offering more reliable insights into biological processes across different species. Future work could further optimize the model and explore additional biological data sources to enhance its performance.

References

[1] Keskin O, Gursoy A, Ma B, et al. Principles of protein− protein interactions: what are the preferred ways for proteins to interact?[J]. Chemical reviews, 2008, 108(4): 1225-1244.

[2] Ni D, Lu S, Zhang J. Emerging roles of allosteric modulators in the regulation of protein-protein interactions (PPIs): A new paradigm for PPI drug discovery[J]. Medicinal research reviews, 2019, 39(6): 2314-2342.

[3] Li H, Sun X, Cui W, et al. Computational drug development for membrane protein targets[J]. Nature Biotechnology, 2024, 42(2): 229-242.

[4] Browne F, Zheng H, Wang H, et al. From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions[J]. Advances in Artificial Intelligence, 2010, 2010(1): 924529.

[5] Wu J. Introduction to convolutional neural networks[J]. National Key Lab for Novel Software Technology. Nanjing University. China, 2017, 5(23): 495.

[6] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 4-24.

[7] Yang J, Li Y, Wang G, et al. An End-to-End Knowledge Graph Fused Graph Neural Network for Accurate Protein-Protein Interactions Prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024.

[8] Xiang S, Zhu M, Cheng D, et al. Semi-supervised credit card fraud detection via attribute-driven graph representation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 14557-14565.

[9] Zhao S, Cui Z, Zhang G, et al. MGPPI: multiscale graph neural networks for explainable protein–protein interaction prediction[J]. Frontiers in Genetics, 2024, 15: 1440448.

[10] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[11] Jin M, Xue H, Wang Z, et al. ProLLM: protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction[J]. bioRxiv, 2024: 2024.04. 18.590025.