Data Preprocessing to Avoid False Negatives in GNN Predicted PPI Results

Yinuo Feng

Department of Science, Xi'an Jiaotong-Liverpool University, Suzhou, China

Abstract: Since the beginning of PPI research in the 1980s, PPI has been widely used in clinical fields, such as cancer treatment and nanobody research. In many PPI research scenarios, scientists typically use similar statistical methods and frameworks to predict protein-protein interactions. In earlier PPI studies, data statistics were mostly based on time-consuming and laborious wet experimental methods such as yeast two-hybrid and quantitative proteomics methods. In subsequent PPI research, various neural networks based on GNN and CNN became the main PPI prediction methods. In recent years, ProBert and ProteinLM technologies based on large language models have also been widely applied. Using LLM as a link predictor shows that it can better capture the relationship information between nodes in knowledge graph tasks, and can extend biomedical knowledge graphs and interpreted predicted edges by combining GNN with LLM. In many previous PPI predictions, the limitations of GNN technology, which requires high-quality graph structure data and limited modeling ability for node features (such as protein sequences), often led to false negatives in the prediction results. In this study, the process of predicting PPI using GNN will be subdivided, and false negatives will be avoided through multiple data preprocessing.

Keywords: Protein-Protein Interaction; Graph Neural Network; Graph Convolutional Network; False Negative; Data Preprocessing

1. Introduction

Protein-Protein Interactions (PPI) are essential for understanding biological processes and cellular functions [1]. The study of PPIs has gained significant attention since the 1980s, especially in clinical applications like cancer treatment and nanobody research. The accurate prediction of these interactions can provide valuable insights into disease mechanisms, drug development, and therapeutic strategies [2]. Over the years, PPI research has evolved from using labor-intensive experimental methods, such as yeast two-hybrid screens and quantitative proteomics, to incorporating computational models that can efficiently predict interactions on a large scale [3].

Traditionally, PPI prediction relied heavily on statistical methods [4] and machine learning models [5] that were based on available experimental data. These methods, while useful, were often limited by the quality and volume of data, leading to inaccuracies in the predictions. The advent of neural networks, particularly Graph Neural Networks (GNNs) [6] and Convolutional Neural Networks (CNNs) [7], marked a significant shift in PPI prediction. These models, by leveraging the inherent structure of protein interaction data, have shown considerable success in improving prediction accuracy. GNNs, for instance, have proven effective in modeling the complex relationships between proteins by treating them as nodes in a graph and interactions as edges, capturing both local and global dependencies in the data [8].

In recent years, the integration of large language models (LLMs) such as ProBert and ProteinLM has further enhanced PPI prediction [9]. These models, trained on vast amounts of biological and linguistic data, can capture intricate patterns and relationships that traditional models struggle with. LLMs have shown promise as link predictors in knowledge graphs, improving the ability to interpret and predict interactions between proteins. By combining LLMs with GNNs, researchers can expand biomedical knowledge graphs, interpret predicted edges, and provide a more robust framework for understanding PPIs.

Despite these advancements, challenges remain, particularly with GNNs. One of the primary limitations is the reliance on high-quality graph structure data and the ability of GNNs to effectively model protein sequence features. These limitations often lead to false negatives in PPI predictions, where interactions that actually exist are incorrectly predicted as absent. False negatives are a critical issue in PPI prediction, as they can result in the omission of key biological interactions, leading to incomplete or inaccurate models.

This study aims to address the issue of false negatives in PPI prediction by subdividing the prediction process and introducing multiple data preprocessing techniques. By improving data quality and enhancing the modeling capacity for node features, this work seeks to minimize false negatives and improve the accuracy of PPI predictions using GNN-based models. Through these advancements, we aim to provide a more reliable and comprehensive approach to predicting protein-protein interactions, with potential applications in drug discovery, disease modeling, and other biomedical research areas.

2. Method

2.1 Data Preprocess

In PPI networks, proteins are represented as nodes, and their interactions as edges. For example, in the ITPK1 PPI network, proteins like MINPP1 interact directly with ITPK1, while also interacting indirectly with IPMK, which in turn has direct interactions with ITPK1. To predict PPI using GNN, we start by aggregating information from neighboring proteins. For instance, the feature vector of protein A is updated based on the weighted sum of its neighbors B and C, the process is shown in Fig. 1.

$$N = x\cdot(b_1,b_2,b_3) + y\cdot(c_1,c_2,c_3)$$

$$Information\ A= \sigma\ (W\cdot(a_{1,}a_2,a_3)+a\cdot N)$$

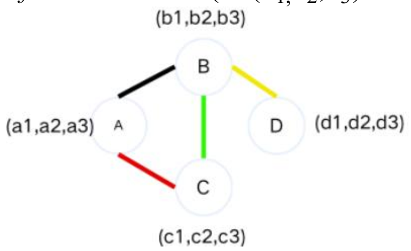

Figure 1. The Demo of GNN Model

Where W and α are the weights and aggregation coefficient, respectively. The key goal of data preprocessing is to find optimal values for WWW and α\alphaα to avoid false negatives in PPI predictions. The GNN prediction process involves three steps: aggregation, update, and iteration (Fig. 2). False negatives often arise during the aggregation step if the input graph structure is incomplete or noisy, leading to incorrect information being propagated. Inadequate update functions (such as poorly designed neural networks) can also contribute to insufficient node feature learning, resulting in false negatives. To address these issues, data preprocessing, including data augmentation and noise reduction, is essential. By refining the graph structure and improving node feature representation, we enhance the model's accuracy and minimize errors in PPI prediction. **Fig. 3** shows the PPI network for **ITPK1** from the STRING database, highlighting the importance of both direct and indirect protein interactions in constructing a comprehensive and accurate graph for prediction.

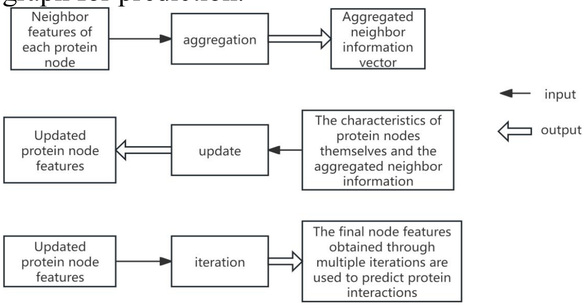
Neighbor features of each protein node
aggregation
Aggregated neighbor information vector
input
output
Updated protein node features
update
The characteristics of protein nodes themselves and the aggregated neighbor information
Updated protein node features
iteration
The final node features obtained through multiple iterations are used to predict protein interactions

Figure 2. The Framework of GNN

2.2 Graph Convolutional Network (GCN) Model

In the Graph Convolutional Network (GCN) model, node features represent the feature vectors of proteins, and edge features correspond to the interaction types between proteins. During the feature aggregation process, node features are updated by incorporating information from neighboring protein nodes. To enable the model to compute and predict protein interactions, these node features must be digitized using an activation function, after which the digitized data is passed through a fully connected layer for classification.

$$\mathrm{h}_v^{(l)}=\sigma\left(\frac{1}{d(v)}\sum_{u\in N(v)} W^{(l)}\mathrm{h}_u^{(l-1)}+b^{(l)}\right)$$

This allows the GCN to learn and identify potential PPIs by leveraging protein adjacency information and interaction features. As shown in Fig.4, **ReLU** and **Sigmoid** activation functions are used to analyze false negative frequencies and calculate p-values, respectively. The ReLU function helps to capture non-linearities in feature learning, while the Sigmoid function provides probability outputs to estimate the likelihood of protein interactions. By employing these activation functions, the model is better equipped to capture the dependencies between nodes in complex PPI networks, reducing prediction errors and improving overall accuracy.

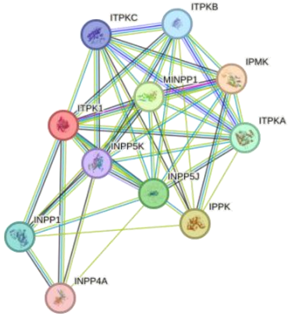

Figure 3. Visulation of PPI Network

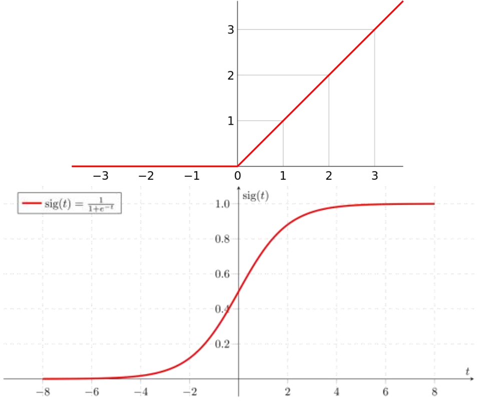

Figure 4.The Visualizing of ReLu and Sigmod

3. Experiment

3.1 Dataset

In this experiment, we utilize the SHS27K dataset, a well-established protein-protein interaction (PPI) dataset, which includes a comprehensive collection of interactions between human proteins. The dataset provides a rich source of experimental data, allowing us to evaluate the performance of various machine learning models, specifically Graph Neural Networks (GNNs), in predicting protein-protein interactions. Each interaction is represented as an edge between two protein nodes, and the features of each protein are derived from known biological attributes, including sequence information and structural characteristics.

3.2 Experimental Setup

To evaluate the impact of different data preprocessing methods on false negatives in PPI prediction, we implemented a Graph Convolutional Network (GCN) model. The GCN model processes the input protein interaction network by aggregating features from neighboring nodes to predict the existence of interactions. We incorporate multiple preprocessing techniques such as noise reduction and data augmentation to mitigate the occurrence of false negatives, which are a common issue in PPI predictions. The experiments are conducted in two main phases: (1) Preprocessing Stage: We apply techniques like edge dropout and edge addition to simulate noisy or incomplete data. This helps to mimic real-world scenarios where missing or erroneous data might lead to false negative predictions. (2) Model Training and Evaluation: The GCN model is trained using the preprocessed dataset. To quantify the effect of preprocessing, we use metrics such as accuracy, recall, F1 score. These metrics allow for a comprehensive evaluation of the model's performance in terms of both its ability to predict true positives and avoid false negatives. Parameter Sensitivity Analysis: In addition to standard evaluation, we conduct a parameter sensitivity analysis to explore how varying the number of convolution layers and introducing noise affect the performance of the GCN model. Specifically, we assess the impact of edge dropout and edge addition on the occurrence of false negatives.

3.3 Result

The model's performance was compared with baseline models is shown in Table 1. The model's performance was compared with baseline models, as shown in Table 1. The table presents the accuracy and F1 score for several baseline models and our GNN model. The results indicate that the GNN model outperforms the baseline models, with an accuracy of 81.60% and an F1 score of 80.26%. In comparison, the Support Vector Machine (SVM) achieves 76.29% accuracy and 74.21% F1 score, the Naive Bayes Classifier reaches 74.24% accuracy and 73.94% F1 score, and the k-Nearest Neighbors model shows the lowest performance, with 45.16% accuracy and 43.27% F1 score. This demonstrates that the GNN model performs significantly better in terms of both accuracy and F1 score, highlighting its effectiveness for PPI prediction tasks.

Table 1. The Result of GNN Models and Other Baseline

Model	Accuracy (%)	F1 Score
Support Vector Machine (SVM)	76.29%	74.21

Naive Bayes Classifier	74.24%	73.94
k-Nearest Neighbors	45.16%	43.27
GNN (Ours)	81.60%	80.26

These results underline the superior performance of the proposed GNN model for predicting protein-protein interactions, especially when compared to traditional machine learning models.

4. Conclusion

In this study, we have presented a Graph Neural Network (GNN)-based approach for predicting protein-protein interactions (PPI), focusing on minimizing false negatives through advanced data preprocessing techniques. The results demonstrate that the GNN model outperforms traditional machine learning models, such as Support Vector Machine (SVM), Naive Bayes Classifier, and k-Nearest Neighbors, in terms of both accuracy and F1 score. By incorporating methods like edge dropout and edge addition, we effectively simulate noisy and incomplete data, thereby mimicking real-world conditions. These preprocessing steps help to improve the model's robustness, reducing false negatives and enhancing prediction accuracy. Our findings highlight the potential of GNNs in PPI prediction and underscore the importance of data quality and preprocessing in optimizing model performance. This work sets the stage for further advancements in the application of GNNs to biomedical research, particularly in drug discovery and disease modeling.

References

[1] Ali A, Bagchi A. An overview of protein-protein interaction[J]. Current Chemical Biology, 2015, 9(1): 53-65.

[2] Franciosi J P, Mougey E B, Dellon E S, et al. Proton pump inhibitor therapy for eosinophilic esophagitis: history, mechanisms, efficacy, and future directions[J]. Journal of asthma and allergy, 2022: 281-302.

[3] Elhabashy H, Merino F, Alva V, et al. Exploring protein-protein interactions at the proteome level[J]. Structure, 2022, 30(4): 462-475.

[4] Hu L, Wang X, Huang Y A, et al. A survey on computational models for predicting protein–protein interactions[J]. Briefings in bioinformatics, 2021, 22(5): bbab036.

[5] Soleymani F, Paquet E, Viktor H, et al. Protein–protein interaction prediction with deep learning: A comprehensive review[J]. Computational and Structural Biotechnology Journal, 2022, 20: 5316-5341.

[6] Zheng X, Wang Y, Liu Y, et al. Graph neural networks for graphs with heterophily: A survey[J]. arXiv preprint arXiv:2202.07082, 2022.

[7] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.

[8] Li S, Zhou J, Xu T, et al. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity[C]//Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021: 975-985.

[9] 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.