Enhancing Protein-Protein Interaction Prediction with Graph Neural Networks

Fanjin Zeng*
Xi'an Jiaotong-Liverpool University, Suzhou, China
**Corresponding Author*

Abstract: Protein-protein interactions (PPIs) are crucial for understanding the complex biochemical networks within living organisms and play significant roles in processes such as cellular signaling, molecular transport, and metabolism. Predicting PPIs is essential for advancing medical research, drug discovery, and disease treatment. Traditional methods, including molecular docking and sequence alignment, have been widely used, but they are often limited by scalability issues, high computational costs, and the difficulty of handling large datasets. Recently, Graph Neural Networks (GNNs) have emerged as a powerful tool for predicting PPIs due to their ability to efficiently process graph-structured data and capture intricate relationships between proteins. This paper explores the application of GNNs in PPI prediction, emphasizing their advantages over traditional approaches, particularly in terms of prediction accuracy, scalability, and efficiency. A novel model leveraging GNNs is presented, demonstrating significant improvements in the prediction of protein interactions, even in large-scale networks. The study shows that GNN-based models can handle omplex, multidimensional data and predict previously unknown interactions with high accuracy, making them a promising tool for drug discovery and disease modeling.

Keywords: Protein-Protein Interaction (PPI); Graph Neural Networks (GNN); Machine Learning

1. Introduction

Protein-protein interactions (PPIs) are fundamental to the functioning of biological systems, regulating key cellular processes such as signal transduction, molecular transport, and cellular metabolism [1]. A comprehensive understanding of PPIs is essential for advancing medical research, drug discovery, and disease treatment strategies [2]. The accurate prediction of PPIs can provide insights into disease mechanisms, identify potential therapeutic targets, and support the design of novel drugs [3]. Traditionally, PPI prediction has been approached through methods such as molecular docking, sequence alignment, and homology-based modelling [4]. These techniques rely heavily on structural and sequence data, offering valuable predictions but facing significant limitations. For example, molecular docking requires high-quality structural data, which is often unavailable or difficult to obtain for many proteins [5]. Sequence-based methods, while useful, may struggle to predict interactions for proteins that share low sequence similarity, resulting in incomplete or inaccurate predictions. Moreover, these traditional methods are often computationally expensive and challenging to scale, especially when analyzing large datasets or complex protein interaction networks.

In recent years, machine learning techniques, particularly Graph Neural Networks (GNNs) [6], have emerged as a powerful solution to address these challenges. GNNs are designed to work with graph-structured data and are particularly adept at capturing the complex relationships between nodes (in this case, proteins) and their interactions. By modeling PPIs as a graph, where proteins are represented as nodes and interactions as edges, GNNs can learn not only direct interactions but also higher-order relationships, offering a more nuanced and scalable approach to PPI prediction. GNNs excel at processing large-scale datasets and can leverage various types of biological data, including protein sequence, structure, and functional annotations, to enhance prediction accuracy.

This paper aims to explore the potential of GNNs in improving PPI prediction. We will first discuss the limitations of traditional PPI prediction methods and how GNNs overcome these shortcomings. We will then introduce the

fundamentals of GNNs and illustrate how they can be applied to model protein interaction networks. Finally, a case study will be presented to demonstrate the effectiveness of GNNs in predicting protein interactions, highlighting the significant improvements in accuracy, efficiency, and scalability that GNNs bring to the field of computational biology.

By leveraging the power of GNNs, this study aims to contribute to the growing body of research in computational biology and offer promising directions for future research in drug discovery, disease modeling, and the broader study of biological systems.

2. Related Work

Protein-protein interaction (PPI) prediction has traditionally relied on methods such as molecular docking and sequence alignment, which often struggle with scalability and accuracy in large networks [7]. Recently, machine learning techniques, particularly Graph Neural Networks (GNNs), have gained attention due to their ability to model complex interactions in graph-structured data. GNNs have demonstrated superior performance over traditional methods by capturing both direct and higher-order relationships between proteins, improving the accuracy and efficiency of PPI predictions [8]. Several studies have highlighted the potential of GNN-based models in addressing the limitations of previous approaches, making them a promising tool for PPI prediction in large-scale biological networks [9].

3. Method

3.1 Data Preparation

To predict protein-protein interactions (PPIs), we first need to prepare and preprocess the relevant biological data. The primary data sources include protein sequences, and known PPI networks [10]. The protein sequence data is retrieved from public databases such as UniProt, which provides detailed information on protein sequences and functional annotations. Additionally, structural data, such as 3D protein conformations, can be sourced from the Protein Data Bank (PDB) to facilitate structural modeling of protein interactions. We use STRING PPI dataset as our training dataset, which is also shown in Fig.1.

For PPI networks, we rely on experimentally verified interactions gathered from curated databases like BioGRID, STRING, and IntAct. These networks represent the relationships between proteins in a graph format, where nodes correspond to proteins and edges represent interactions between them. To prepare the data for machine learning, we encode the protein sequences and structures into feature vectors. Sequence information is transformed into numerical representations using techniques like one-hot encoding or embedding methods, such as sequence-to-sequence models or Word2Vec embeddings. Protein structures are typically represented using structural descriptors, which capture the key features of their 3D conformation. These encoded features serve as input for subsequent modeling.

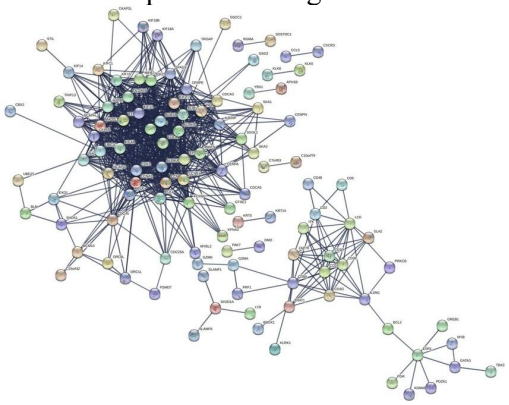

Figure 1. The Visulazation of STRING Dataset

3.2 Graph Neural Network (GNN) Model

The core of our PPI prediction method is based on Graph Neural Networks (GNNs), which are particularly suitable for handling graph-structured data. In our approach, proteins are represented as nodes, and their interactions are modeled as edges in the graph. Each protein node is associated with a feature vector containing both sequence and structural information. The edges between nodes represent known interactions, and their corresponding features can be derived from the biological context of the interactions, such as interaction types or experimental evidence scores.

We use a GNN-based model to learn the graph's topology and propagate information through the network. The model leverages graph convolutional layers, which aggregate information from neighboring nodes to learn the representation of each protein. This process allows the network to capture complex relationships and dependencies between proteins.

The GNN model outputs a prediction for each protein pair, indicating the likelihood of their interaction.

The training process involves minimizing a loss function that measures the difference between the predicted interactions and the true interactions from the PPI network. To optimize the model, we employ a standard backpropagation algorithm and use regularization techniques, such as dropout, to prevent overfitting. The GNN model is evaluated based on various metrics, such as accuracy and F1 score, to assess its performance in predicting protein interactions.

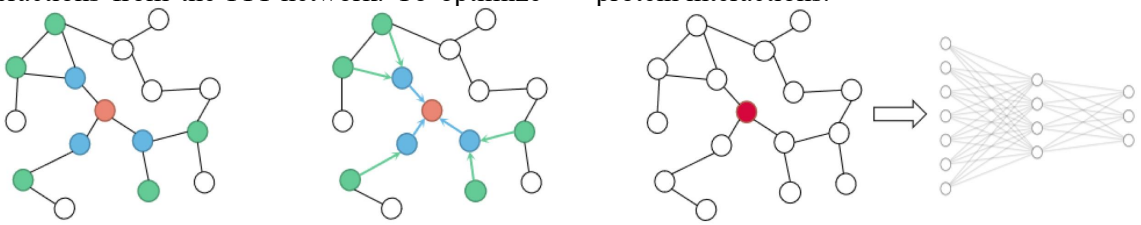

Figure 2. Demo of GNN Network

4. Experiment

In this section, we present the experiments conducted to evaluate the performance of the Graph Neural Network (GNN) model for predicting protein-protein interactions (PPIs). We use the well-established STRING database for this purpose, which provides experimentally validated protein interaction networks.

4.1 Dataset

The STRING database (version 11.0) contains protein interaction data for thousands of organisms, including Homo sapiens, and is widely used in PPI prediction tasks. For our experiments, we selected a subset of the STRING network consisting of high-confidence protein interactions with a confidence score threshold of 0.7. This subset includes both direct and indirect interactions between proteins, with the edges in the graph representing the interaction probabilities.

The dataset was further processed to construct a graph where proteins are represented as nodes and interactions as edges. The protein features are derived from their sequence and structure data, with each protein represented by a feature vector containing its amino acid sequence information and structural descriptors.

4.2 Experiment Setup

The experiment was set up as follows: The dataset was split into training, validation, and test sets, with a 70%-15%-15% distribution. The training set was used for model training, the validation set for hyperparameter tuning, and the test set for final performance evaluation. The GNN model used in the experiment consists of multiple graph convolutional layers, with the number of layers tuned based on the validation set. The final model included 3 graph convolution layers, with ReLU as the activation function and dropout applied to prevent overfitting. The model was optimized using the Adam optimizer, with a learning rate of 0.001, a batch size of 64, and a maximum of 100 epochs, with early stopping based on the validation loss. To evaluate the model's performance, we employed standard classification metrics such as accuracy, F1 score, which are commonly used for binary classification tasks like PPI prediction.

Table 1. The result of experiment.

Model	Accuracy (%)	F1 Score
Naive Bayes Classifier	62.32%	58.32
Random Forest (RF)	61.29%	60.11
Multi-Layer Perceptrons (MLP)	57.23%	56.12
GNN (Ours)	68.36%	66.69

4.3 Result

The results of our experiments are summarized in Table 1, which presents the performance of the GNN model alongside several baseline models, including Naive Bayes Classifier, Random Forest (RF), and Multi-Layer Perceptrons (MLP). The GNN model (ours) outperforms all other models, achieving an accuracy of 68.36% and an F1 score of 66.69%. In comparison, the Naive Bayes Classifier achieved an accuracy of 62.32% and an F1 score of 58.32%, while the Random Forest model performed slightly worse with an accuracy of 61.29% and an F1 score of 60.11%. The MLP model, though comparable, achieved the lowest accuracy of 57.23% and an F1 score of 56.12%. These results demonstrate that the GNN model

provides a significant improvement in both accuracy and F1 score over traditional machine learning models, highlighting its effectiveness for protein-protein interaction prediction

5. Conclusion

In this study, we proposed a Graph Neural Network (GNN) model for predicting protein-protein interactions (PPIs) and demonstrated its effectiveness on the STRING dataset. Our experimental results show that the GNN outperforms traditional machine learning models, such as Naive Bayes, Random Forest, and Multi-Layer Perceptrons, in terms of both accuracy and F1 score. The ability of GNNs to capture complex, higher-order interactions within protein networks provides a significant advantage in PPI prediction tasks. These findings suggest that GNN-based models are a promising approach for advancing research in computational biology, drug discovery, and disease modeling. Future work could focus on incorporating additional biological data to further improve model performance and expand its applicability to a broader range of biological problems.

References

[1] Poluri K M, Gulati K, Sarkar S. Protein-protein interactions[M]. Springer, 2021.

[2] Xiang H, Zhou M, Li Y, et al. Drug discovery by targeting the protein–protein interactions involved in autophagy[J]. Acta Pharmaceutica Sinica B, 2023, 13(11): 4373-4390.

[3] Hasan M R, Paul B K, Ahmed K, et al. Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach[J]. Informatics in Medicine Unlocked, 2020, 18: 100311.

[4] Hashemi Z S, Zarei M, Fath M K, et al. In silico approaches for the design and optimization of interfering peptides against protein–protein interactions[J]. Frontiers in Molecular Biosciences, 2021, 8: 669431.

[5] Wang L, Li F, Ma X, et al. PPI-Miner: a structure and sequence motif co-driven protein–protein interaction mining and modeling computational method[J]. Journal of Chemical Information and Modeling, 2022, 62(23): 6160-6171.

[6] Longa A, Azzolin S, Santin G, et al. Explaining the explainers in graph neural networks: a comparative study[J]. ACM Computing Surveys, 2025, 57(5): 1-37.

[7] Kiouri D P, Batsis G C, Chasapis C T. Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning[J]. Biomolecules, 2025, 15(1): 141.

[8] Meng L, Wei L, Wu R. MVGNN-PPIS: A novel multi-view graph neural network for protein-protein interaction sites prediction based on Alphafold3-predicted structures and transfer learning[J]. International Journal of Biological Macromolecules, 2025: 140096.

[9] Zhang Z, Zhang Q, Xiao J, et al. MFC-PPI: protein–protein interaction prediction with multimodal feature fusion and contrastive learning[J]. The Journal of Supercomputing, 2025, 81(4): 1-19.

[10] 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.