Protein-Protein Interaction Prediction Using Distributed Representation and Graph Convolutional Networks

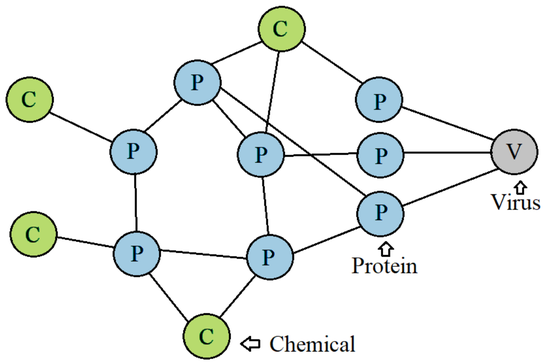
Hayden Gemeinhardt
Research Advisors: Adam Bess, Dr. Supratik Mukhopadhyay

Introduction

Protein-protein interaction (PPI) is vital in the discovery of new drugs, yet analyzing protein pairs in a lab is expensive. LSU's DeepDrug AI framework aims at alleviating the cost of drug discovery, so the question was raised: can PPIs be effectively predicted through the scope of network theory? Moreover, can using distributed representation of protein amino acid structure increase the accuracy of the model compared to other studies that address the same problem?

Background

The model is designed to be implemented into LSU's DeepDrug--an AI network that can discover drugs to target cellular diseases.



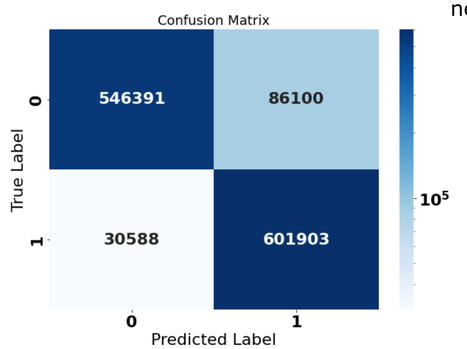
This study utilized two novel works from existing research:

- BioVec provided a pretrained model for distributed representation of protein sequences [1]
- GraphSAGE, a graph convolutional network (GCN) framework for representation learning on large graphs, was used to obtain node embeddings [2]

GCNs are a relatively new concept, so numerous applications have yet to be studied. Furthermore, little work has been done in predicting PPIs using GCNs except on smaller reference datasets.

Analysis

The following results were obtained through a sample of 632,000 positive and 632,000 negative test links.



Accuracy=90.7755% MCC=0.8187
Precision=0.8749 Recall=0.9516

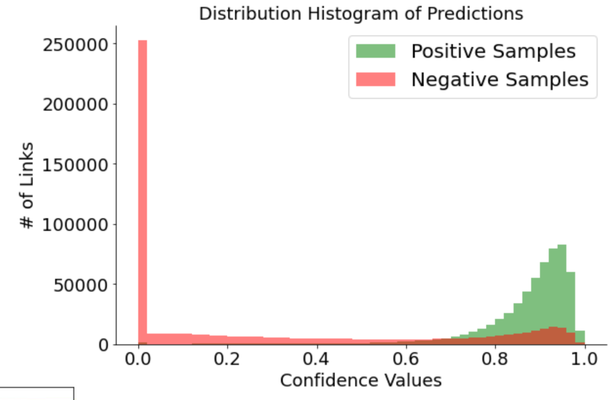
The model obtained an accuracy of over 90%, while our team expected an accuracy of around 80% for this type of large experimental dataset.

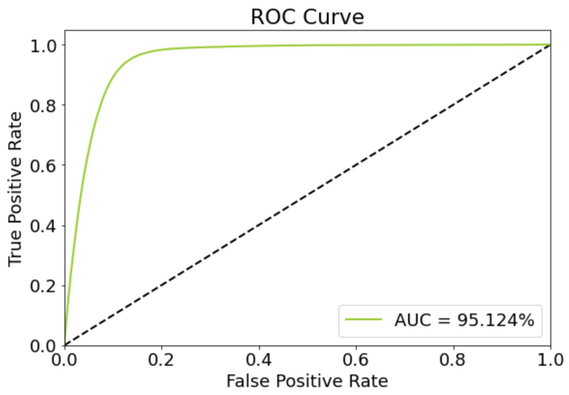
The model had a high false positive rate (13.6%) as expected, but a smaller false negative rate (4.8%).

Human verification found many high confidence false positives to be either nonsensical (and thus easily ignored by DeepDrug) or to contain two proteins with similar characteristics.

The histogram shows two clear peaks with a distinct positive threshold but some overlap.

The stretch of negative samples throughout the bottom of the graph, along with the slope of positive samples on the right, signifies that the model would benefit from prediction calibration in the future.





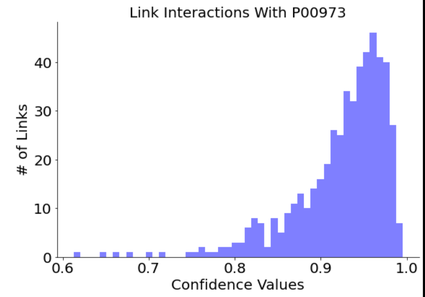
A high AUC percentage is especially important considering the unbalanced nature of our data. It is much easier to classify a false link than a true link.

A high AUC shows the strength of our model in identifying key signals and structure from a large, messy dataset.

COVID-19 Protein Test:

The model can predict select protein interactions against every other protein in the graph.
Protein P00973, or OAS1, is an important SARS-CoV-2 protein.
Its top predicted protein is Q00537, or CDK17.
OAS1 and CDK17 are strongly correlated for respiratory syndromes [3].
Moreover, DeepDrug found CDK proteins important for SARS-CoV-2.

Overall, 509 proteins interacted with P00973 in our dataset. The model found 76% of the protein interaction pairs with a confidence value above 0.9.



Conclusion

The study showed that PPIs can be predicted to an accuracy that allows the theoretical model, using both a GCN and distributed representation, to be a beneficial application on the DeepDrug framework.

Our model achieved high accuracy on a significantly large and messy dataset, which no other work using GCNs on PPI data has accomplished.

Although the system's hyperparameters are not yet optimally tuned, we believe we have an architecture that already outperforms similar predictive models.

Methodology

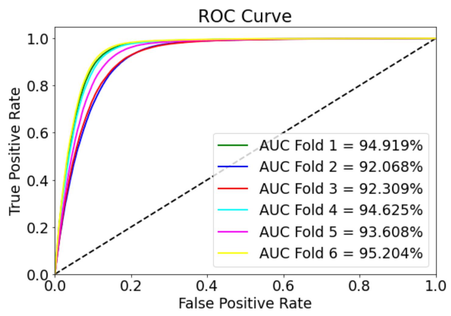
- PPI data was pulled from hu.MAP, IntAct, and STRING databases (95,000 proteins, 6,325,000 interactions weighted >0.01)
- Each protein's amino acid sequence was fed into the ProtVec model and assigned a vectorized representation
- A graph was constructed in which each vectorized node represented a protein and each weighted edge an interaction
- Using the graph, GraphSAGE was trained to generate node embeddings for link prediction

Validation

6-fold training was used to validate against overfitting. Due to the long training time of the model, only 10 epochs were used.

High accuracies were quickly achieved in each model, showing our model did not overfit.

	AUC	Accuracy
Fold 1	94.92%	90.36%
Fold 2	92.07%	86.41%
Fold 3	92.31%	86.54%
Fold 4	94.63%	90.02%
Fold 5	93.61%	88.58%
Fold 6	95.20%	90.83%



Future Work

- Biological systems representation clustering
- Further model training for higher accuracy
- Calibrate predictions for lower false positive rates
- Predict weights of the protein interactions

References

[1] Asgari, Ehsaneddin, and Mohammad R. Mofrad. "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics." *PLOS ONE*, vol. 10, no. 11, 2015.
[2] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. Neural Information Processing Systems (NIPS), 2017.
[3] Bai, Jianhui, et al. "A High-Throughput Screen for Genes Essential for PRRSV Infection Using a Piggybac-Based System." *Virology*, vol. 531, 2019, pp. 19–30.