Graph embeddings of enriched protein-protein interaction (PPI) networks for identification of disease nodes

Benjamin Angulo, Yusuf Roohani, Kendrick Shen
Department of Computer Science
CS 224W Final Project Writeup Fall 2019
{bangulo, yroohani, kshen6}@stanford.edu

Abstract: Understanding the genetic and molecular underpinnings of diseases is of crucial importance in biology and could provide valuable insights that help in the development of therapeutics against human diseases. However, predicting the genetic and molecular players in a given disease is a technically difficult task. Recent developments in graph and network analysis have provided computational tools for this task. Many previous attempts have leveraged existing molecular networks such as protein-protein interaction networks to predict disease nodes. We build on their successes by (a) augmenting molecular networks via incorporation of literature-derived data (b) enriching the graph with biologically-relevant molecular features, and (c) analyzing performance of recently developed graph algorithms such as GraphSage that utilize graph structure and node features to allow for the generation of high-quality embeddings useful in disease prediction. We show that while GNBR and PP-Decagon datasets had similar performance, their combined performance was lower than using independently, suggesting that more work and thought may be required in combining heterogeneous datasets. Further, our project data generated two key-insights: (a) incorporation of molecular node features in the PP-Decagon dataset nearly doubles predictive performance and (b) GraphSage performs comparably Graph Convolutional Networks. To the best of our knowledge, we are the first group to incorporate UniProt features in predicting disease labels on a PPI network. These results suggest promising new avenues of inquiry.

1 Introduction

Most cellular components within living organisms generally exert their functions through a network of interactions [1]. Protein-Protein interaction (PPI) networks are well known to be useful for identifying genes that are associated with a disease [2]. Several groups have explored relationships within protein-protein interaction networks, most of them aimed to discover communities via exploitation of the graph's structure and connectivity. However, disease modules–subcollections of nodes and their induced subgraphs whose collective abnormal function corresponds to disease–don't always correspond to topological or functional modules dependent on network connectivity. [1]

Barabási et al. [1] highlight that the two main blockers towards reliable network-based detection of disease-associated proteins are the incompleteness of available interactome maps and the limitations of the existing tools to explore the role of networks. Through this project, we aimed to address both these concerns through

a range of experiments that look at both integrating distinct protein interaction networks as well as interrogating the resulting graphs using embeddings. We focused on the following three objectives:

1. Integration of traditionally distinct data modalities (including literature based knowledge graphs) to enrich the information content of a PPI graph
2. Incorporation of molecular features into the PPI graph
3. Exploring new methods for embedding these enriched graphs within a shared space that allows for better detection of disease modules in the graph

2 Review of prior work

2.1 Learning intelligently from biological networks

In 2015, Ghiassian et al [3] found that disease-genes exist within the graph in what they describe as disease neighborhoods or disease modules. While these disease modules only capture approximately 10-30% of the disease-genes in the largest connected component, the agglomeration demonstrated by the disease genes is highly significant. Importantly, [3] demonstrated that connectivity significance rather than connection density was more important for predicting the observed interaction patterns. To study the importance of connection density, Ghiassian et al. used different methods, including the Louvain Method, Markov Cluster Algorithm, and a link-connection algorithm, all in order to detect communities of functionally related proteins. However, these approaches were not particularly useful in capturing the disease modules as only 15% of diseases had any significantly enriched community. By comparison, they found that connectivity significance was far more predictive, and disease modules had average connectivity p-values among the disease-gene sets of 10^{-241}.

While extending the general understanding of the relationship between diseases in their biological network, Agarwal et al. [4] raise a fundamental concern of applying classical topology- and connectivity- dependent analysis of disease pathways. Somewhat surprisingly, many of the proteins associated with a given disease phenotype do not interact within a PPI network, thus clustering, community-detection, and other proximity-based algorithms fail to extract valuable information for predicting disease subnetworks. As such, [4] shows that a study of the higher order PPI network structure, including features such as density of a given pathway, distance between its connected components, and network modularity, more adequately characterize disease pathways, and thus make the analysis methods more robust to loosely-connected or disconnected disease regions. In addition, [4] showed that diffusion (random walk) based methods, which construct a heuristic of the nodes' structural and functional neighborhoods, offer more predictive power over linkage based methods. Importantly, the authors in [4] conclude that in a conventional PPI network, disease proteins are "weakly embedded" rather than densely interconnected. As a consequence, it is critical that the design of any disease protein discovery algorithm should only rely on latent connectivity to a limited extent.

We are also interested in understanding the implications of the rapidly growing amounts of high throughput biological data on extracting more value from PPI networks. Schulte-Sasse et. al [10] looked at combining several heterogeneous -omics data types including into a single graph representation of that data. Their goal was to predict novel pan-cancer genes. This was done by attaching gene expression, methylation and other omics data to each node in a PPI before embedding the graph. This paper shows that using graph convolutional methods as a data integration

strategy is a feasible approach. However, there are several avenues for improving this approach. From a data perspective, the authors did not consider incorporating elements from the rich biomedical text mining literature. In terms of methodology, they only used prediction accuracy as a metric of performance and did not conduct any deeper analysis on the properties of the intermediate graph to validate it as a representation.

2.2 Node embeddings in biological networks

The application of graph embeddings to network biology was reviewed by Watson et al. [20], comparing random walk-based, convolutional neural network-based, and matrix decomposition-based methods for network alignment, community detection, and function prediction. This paper provides an informative starting point for reasoning about the trade-off between learning directly on networks by interacting with them and indirectly via their embeddings. In addition, it notes that embeddings that adequately describe network topology for the purposes of alignment (extracting analogous proteins across different tissue types, or species) may not accurately capture protein function. This is an important result to consider, and reinforces the idea that protein network analysis is intrinsically dependent on one's biological objectives. GraphSage is an algorithm developed at Stanford by Hamilton et al. that builds on the simple neighborhood aggregation method employed by graph convolutional networks (GCNs), adding in a skip connection by explicitly incorporating each node's features from the previous layer. The network may then learn to positively or negatively bias the node's own input features, depending on the specifics of the network.

3 Data collection

Our first goal was to gather all the relevant data and perform some conventional graph-based analyses on these sets. For this purpose, we downloaded two PPI networks: Decagon (obtained from the BIOSNAP repository) and ConsensusDB. This data is primarily drawn from Menche et al. For disease-gene relationships, we will use BIOSNAP's disease-gene association network (DG-AssocMiner), which draws from a variety of sources including the Online Mendelian Inheritance in Man (OMIM) database[13].

Our goal was to enrich these PPI graphs using external datasets including the Global Network of Biological Relationships (GNBR) which uses a natural language processing approach to extract biological relationships from scientific text as well as the multi-omics data from public databases TCGA and GTEX. We successfully incorporated TCGA/GTEX data together with Consensus DB into a dataset, but we decided against using these data as discussed in our milestones.

3.1 Initial findings and summary statistics

1. **PP-Decagon**: The PP-Decagon dataset is comprised of physical (direct) and functional (indirect) interactions between proteins. We loaded this data into an undirected graph, where the nodes are proteins, and the edges are the relationship between them. In this dataset, there are 19,081 nodes and 715,612 edges. The degree distribution of nodes is shown below. The distribution follows a similar pattern seen in other natural networks (Figure 1).
2. **GNBR**: In light of scientific estimates which suggest that roughly 80% [19] of biomedical data is locked away in unstructured states (e.g. scientific

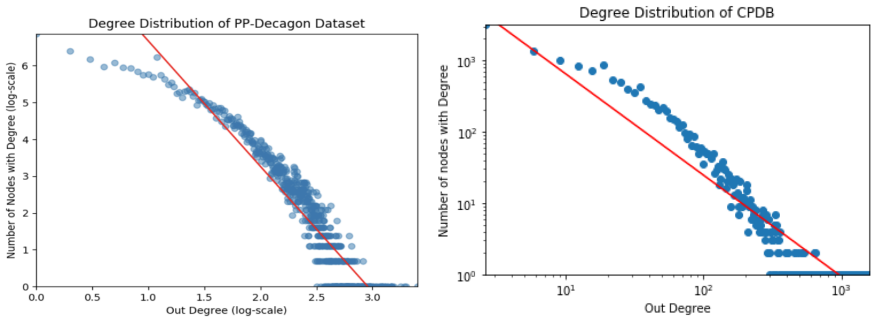

Figure 1: Degree Distribution for PP-Decagon and CPDB (Not Used)

publications, patents, clinical notes), we sought to augment our graphs by including data generated from natural language processing projects such as GNBR[8] which extract relationships between biological entities from single sentences of PubMed abstracts using their dependency paths. We focused on extracting the gene-gene interactions (filtered for human specific data) from the GNBR dataset. While we explored filtering this data to only the protein-protein subsection to conform better with our current PPI-oriented data, we ultimately were unsuccessful at disentangling the PPI subset from this source. In future experiments outside this project, we hope to experiment with utilizing the directionality data that GNBR has for many of these relationships. We did attempt to download the gene-disease data from GNBR, but have struggled to find an appropriate mapping of their disease codes (MeSH terms) to our DG-AssocMiner dataset.
Within the GNBR gene-gene dataset, there was 7,157 nodes and 194,633 edges. The degree distribution (below) is similar to the other datasets (Figure 3). We analyzed the overlap between our datasets, for instance the GNBR dataset and the PP-Decagon dataset. Of note, only 1,114 nodes overlapped between the two dataset, which seemed particularly low. This was not due to mapping as both datasets used NCBI's gene IDs. Further, the two datasets shared 48,020 edges, which appeared reasonable given that the two datasets contain slightly different data (gene-gene vs protein-protein). It is possible that GNBR's NLP did not capture a large portion of the data in the literature which could explain some of the reduced dataset overlaps. However, it should be noted that the 146,613 unique relationships added by GNBR represents a substantial augmentation of our data.

3. **UniProt Features**: UniProt contains relevant sequence and molecular features for virtually every protein in humans. Ben Angulo received permission from the company he works at (OccamzRazor) to use the features he generated for UniProt while working there. Data extraction from UniProt focused on extracting Post-translational modifications as well as sequenced-based features. In total, 71 features were extracted. However, some features were relatively sparse or considered to be redundant, and so that list was filtered down 10 total molecular features. We analyzed the features within the PP-Decagon network and generated visualizations. As an example, we include the distribution of phosphorylation features (magenta nodes) in the subgraph surrounding the Parkinson's disease gene LRRK2, which is itself an enzyme that phosphorylates other proteins (Figure 4).

4. **DG-AsocMiner**: The ultimate goal of our work is to predict proteins/genes that are associated with a given disease. Therefore, we need data that could

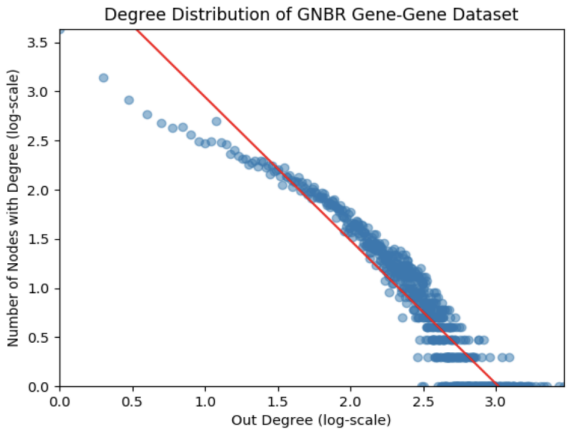

Figure 2: GNBR Degree Distribution

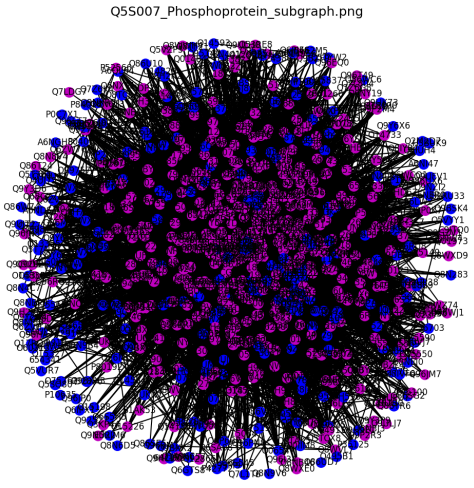

Figure 3: LRRK2 Subgraph with Phosphorylated Nodes in Magenta

serve as ground-truth labels with which we can learn a model that can make predictions on whether a gene is involved in a disease. We chose BIOSNAP's DG-AssocMiner dataset, which contains disease-gene relationships, to work with due to its wide-scope acrosss diseases. We loaded the data into a graph for analysis, where the nodes are diseases and genes (7,813 total, 519 for diseases, 7,294 for genes) and the edges are relationships between them (21,357). Once again, the degree distribution mapped similar to our other graphs (Figure 4). As opposed to the DIAMOnD paper, which only trained on 70 diseases, we train on 519 diseases. It's important to note that many of these diseases may not have a genetic component, or are molecularly heterogeneous, consequently leading to poor predictability.

Table 1: of Database Overlap

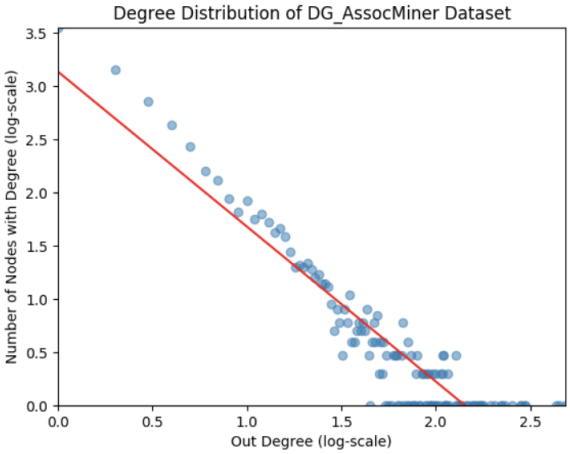

Figure 4: DG-AssocMiner Degree Distribution

Dataset	Entities	Edges	Overlap TCGA/GTex	Overlap GNBR (Entities)	Overlap GNBR (Edges)
PP-Decagon	19,081	715612	16,393	1114	48020
ConsensusDB	6262	67,250	6,155	Not calculated	Not calculated
GNBR	7157	194,633	Not calculated	NA	NA

4 Methodology

4.1 Mathematical formulation

We define a biological network $G = (V, E)$, where V is the set of all nodes and E is the set of all edges between them. The nodes in our model correspond to a set of genes. The edges represent interactions between them. We are defining the base graph G using information obtained from PPI-Decagon. We compared GNBR with both PPI-Decagon and CPDB and chose the one with greater overlap which was PPI-Decagon.

To enrich these graphs with additional edges (E'), we tapped into GNBR for connections derived from the literature, producing a new graph $G' = (V, E + E')$. Finally, we also used UniProt features that describe the properties of the proteins that the genes code for. This would be represented as a feature matrix $\mathbf{X}$ of size $(V \times d)$, where d is the dimensionality of the feature vector. Putting this together we would achieve an enriched graph $G'' = (V, E + E', X)$.

We have ground truth labels available for genes that are known to be disease causing ($V_d = v \in V : Y(v) = 1$) for a set of 519 diseases. For the rest, we assume that they are not disease causing ($V_{nd} = v \in V : Y(v) = 0$). Using disease gene prediction as the key task for assessing performance, we can create an embedding space Z of dimension d_z and map nodes V in this graph G'' into that embedding space. The goal of the embedding would be to create a representation of the data that best identifies the characteristics of disease causing nodes. This could be achieved through minimizing a simple binary cross-entropy loss $-y \log p + (1-y) \log(1-p)$. In summary, our data collection, aggregation, and augmentation scheme is as follows:

$$G = (V, E) \text{ [Original PPI graph]}$$
$$G' = \text{add\_links } (G) = (V, E + E')$$
$$\text{disease\_node\_predict } (G'' = (V, E + E', x))$$

4.2 Implementation

We built a pipeline to classify as disease causing or not using the labels we had available for 519 diseases. We used two models for this purpose: graph convolutional neural networks and GraphSage, with motivation explained in section 2.2. For both models, when training without node features, we used an input data dimensionality of 1, hidden layers dimensionality of 32, and output layer as 2 (d_z). The number of layers was set as 2. When training with features the input dimension as 11. We implemented this pipeline in PyTorch geometric with CUDA functionality for training. Training each model for all 519 diseases on an NVIDIA V100 GPU in Google Cloud instance took 3-4 hours.

4.3 Training and evaluation

For each scenario, we trained 519 different models, one for each disease. For each disease, we held out 10% of the set for testing. Within the training set, we then applied 5-fold cross validation for training. Both for train test split and for cross validation we used stratification to ensure that each fold would get the same number of positive and negative labels. At the end of training across all folds for a given disease, we reported test accuracy by selecting the model that had the highest validation score achieved in any fold. Many diseases showed very low scores, which could indicate poor model performance, small sample size of disease genes, or that the disease lacks a strong causative genetic component. Training and validation sets were implemented using node masks. So the full graph with all its edges was available during training and testing but only the unmasked nodes were used for calculating the loss and updating parameters.

We trained the model using the F1 score to measure performance. One of the main challenges with this dataset was the significant class imbalance between nodes that were linked with causing a disease (usually $< 10^2$) and those that weren't (usually around 10^4). The model could exploit this by trivially predicting all nodes as not disease-causing. To correct for this, we added weighting terms in the loss function to increase the penalty for false negatives. These were calculated based on the relative proportion of each class c among all classes $\mathbf{C}$ within that specific fold.

$$W_c = 1 - \frac{count(node_c)}{\sum_{d \in \mathbf{C}} count(node_d)}$$

We looked at all different combinations of graphs and features combinations. These resulted in a total of 12 scenarios (combined training time of over 48 GPU compute hours) are outlined in Table 1. We measured performance using a metric called 'Recall at 100'. This is the same metric used in Agrawal[4] and Ghiassian[3]. Here we take all the nodes in the test set, pass them through our network and use the output probability for being in the disease class to rank nodes. We then pick the top 100 nodes and measure what fraction of total disease causing nodes in the test set are accounted for in those 100 nodes. The advantage of this method was that we didn't need to rebalance our test set to contain an equal number of positive and negative classes.

5 Results

We used disease gene association pairs from DG-AssocMiner, and tested across 12 conditions. We report a summary of our results in Table 2. We note increased

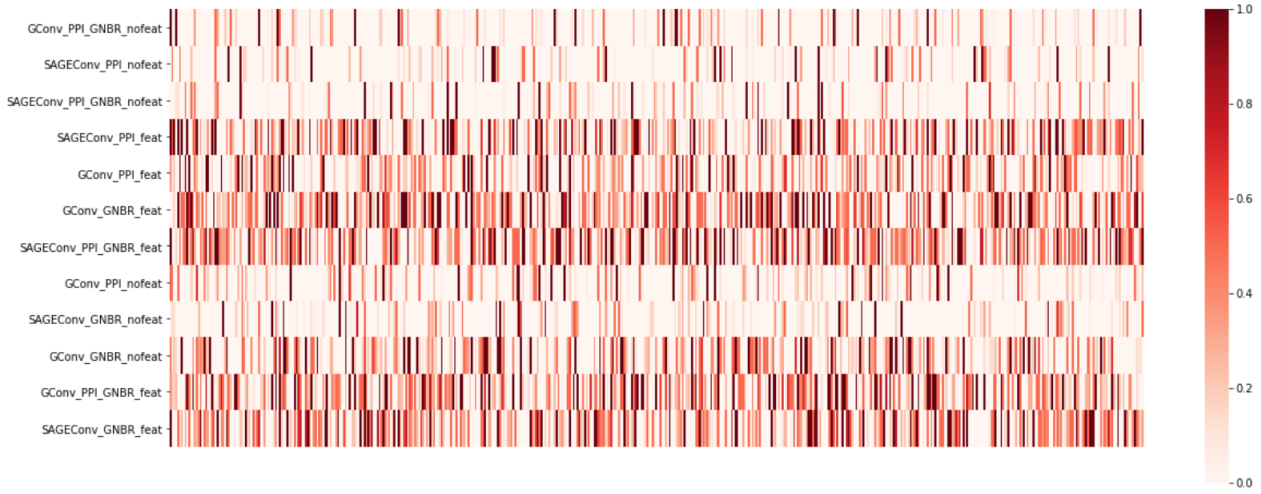

Figure 5: Heatmap across diseases for different Methods

performance across many conditions when using the UniProt-derived molecular features. To give an indication as to the relative improvement across diseases, we include plots of the difference in Recall at 100 between different conditions Figure 7. For space, we have included just a few examples.

Table 2: Model Performance

Model	Graph	Features	Mean recall	Weighted recall
GCN	PP-Decagon	None	0.101	0.000641
GCN	PP-Decagon	UniProt	0.204	0.00123
GCN	GNBR	None	0.231	0.00128
GCN	GNBR	UniProt	0.308	0.00198
GCN	PP-Decagon + GNBR	None	0.078	0.000453
GCN	PP-Decagon + GNBR	UniProt	0.303	0.00185
GraphSage	PP-Decagon	None	0.0767	0.000513
GraphSage	PP-Decagon	UniProt	0.273	0.00179
GraphSage	GNBR	None	0.0843	0.000603
GraphSage	GNBR	UniProt	0.303	0.00200
GraphSage	PP-Decagon + GNBR	None	0.069	0.000325
GraphSage	PP-Decagon + GNBR	UniProt	0.326	0.00204

6 Discussion

6.1 Augmentation of PPI using literature-derived Gene-Gene interactions

We sought to augment the PP-Decagon Protein-Protein Interaction network by ingesting data derived from the Global Network of Biomedical Relationships (GNBR)[8] which extracted biological relationships from 24 million scientific articles. We specifically incoporated GNBR's human Gene-Gene interaction relationships, which are a superset of Protein-Protein interactions. Augmentation of our PP-Decagon data, using GNBR failed to increase our predictive power as measured by the mean recall rate across diseases as seen in table 2. In fact, these two sets performed better independent of each other. We attribute this behavior to the fact that these datasets represent molecularly distinct interactions. We believe that combining them

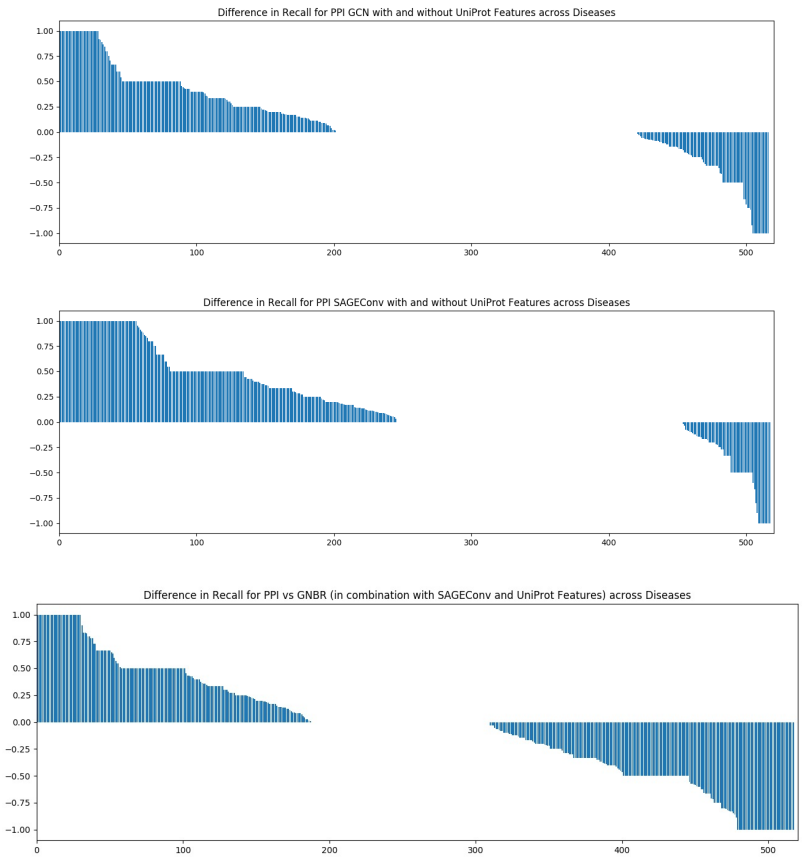

Figure 6: Plots of Difference between models in Recall Rates across diseases

uniformly reduced their independent predictive power Consequently, we believe these results raise potential hurdles in the concatenation of heterogeneous data.

6.2 Incorporation of molecular features increases predictive power

Previous efforts [1,4,10] have elucidated the power of molecular biological data, such as protein-protein interactions, in predicting disease nodes. While much of the past work has been done on network-generation and augmentation, we sought to complement these efforts by enriching our graphs with node-features derived from molecular databases. Conceptually, proteins can effect their changes through physical interactions. However, in addition, many proteins not only bind to their partners, but also chemically *modify* them by the covalent attachment of chemical groups, examples of these modifications can include phosphorylation, acetylation, ubiquitination, and glycosylation of their targets. These are known as post-translational modifications (PTMs) and can be thought of as protein-features. PTMs are important in cell function and are dysregulated in many disease pathologies. To ingest these molecular features, we extracted PTMs and additional protein features from the Universal Protein Resource (UniProt) which has the most extensive database of sequence and molecular features for nearly every known protein. We filtered our extracted UniProt feature set for a curated set of features that are widely distributed across proteins. Importantly, inclusion of UniProt molecular features led to significant improvement in the mean recall (MR) at 100 across all tested diseases, leading

to an approximate doubling of the MR at 100 from 0.204 from 0.101 (Table 2 and Figure 6).

6.3 GraphSAGE and GCN perform comparatively

Recent advances in algorithms, particularly GraphSage, have led to advancement in the ability to use node features in for learning neural embeddings of graphs. We found that with regards to our PP-Decagon set, GraphSage performed 30% better when the UniProt features were included than not. Interestingly, when using the GNBR dataset instead of the PP-decagon set, the GCN tended to performed better. We believe that this is in part due to the nature of GNBR, which is Gene-Gene relationships. The features we use from UniProt are protein-related features that should not necessarily improve performance for Gene-Gene relationships, which often have indirect relationships.

We have compiled our source code for our various tasks at `https://github.com/yhr91/CS224W_project/`.

7 Future work

Our work has suggested a critical role for molecular features in disease node classification. **Our results show roughly a doubling of performance.** Future work should further explore and expand the set of molecular features that have predictive power. Further, we would like to continue exploring the integration of literature-derived data in our molecular graphs by integrating these datasets in a manner that allows the model to differentiate between edges. One approach that implements this is Graph Attention Networks that weigh the importance of incoming messages differently based upon the task the model is training on. These two directions would be valuable to explore as follow-up to this work. Additionally, the task of classifying individually on 519 separate diseases was computationally-intensive and suffered from a large class imbalance (discussed previously, and dealt with via weighted loss). It would be valuable to reduce the number of classes for this multiclass classification task, as in the current setup, the total number of possible distinct labels is 2^{519} (powerset), as it is a binary vector $\in \mathbf{R}^{519}$. Employing a clustering algorithm such as hierarchical clustering, or even a simpler counting-based approach, may be useful in aggregating the information from the 519 diseases into a broader but more robust solution to the node classification task.

References

1. Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews genetics, 12(1), 56.
2. Navlakha, S., & Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. Bioinformatics, 26(8), 1057-1063.
3. Ghiassian, S. D., Menche, J., & Barabási, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS computational biology, 11(4), e1004120.
4. Agrawal, M., Zitnik, M., & Leskovec, J. (2018). Large-scale analysis of disease pathways in the human interactome. In PSB (pp. 111-122).
5. Zitnik, M., & Leskovec, J. (2018). Prioritizing network communities. Nature communications, 9(1), 2544.
6. Sosa, D. N., Derry, A., Guo, M., Wei, E., Brinton, C., & Altman, R. B. (2019). A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. bioRxiv, 727925.

7. Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864). ACM.
8. Percha, B., & Altman, R. B. (2018). A global network of biomedical relationships derived from text. Bioinformatics, 34(15), 2614-2624.
9. Zitnik, M., & Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. Bioinformatics, 33(14), i190-i198.
10. Schulte-Sasse R., Budach S., Hnisz D., Marsico A. (2019) Graph Convolutional Networks Improve the Prediction of Cancer Driver Genes. In: Tetko I., Kůrková V., Karpov P., Theis F. (eds) Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science, vol 11731. Springer, Cham
11. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
12. Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864). ACM.
13. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science [Internet]. 2015 Feb 20 [cited 2019 Apr24];347(6224):1257601. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25700523
14. https://www.omim.org/
15. https://cancer.sanger.ac.uk/cosmic
16. https://zenodo.org/record/1035500#.XafA2-dKiL4
17. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
18. http://lincsportal.ccs.miami.edu/dcic-portal/
19. HIT Consultant. Why unstructured data holds the key to intelligent healthcare systems [Internet] Atlanta (GA): HIT Consultant; 2015. [cited at 2019 Jan 15]. Available from: https://hitconsultant.net/2015/03/31/tapping-unstructured-data-healthcares-biggest-hurdlerealized/#.XFvZ1lwvOUk. [Google Scholar]
20. Nelson, Walter, et al. To Embed or Not: Network Embedding as a Paradigm in Computational Biology. Front. Genet., 01 May 2019 | https://doi.org/10.3389/fgene.2019.00381
21. Singh, Rohit, et al. Global alignment of multiple protein interaction networks with application to functional orthology detection PNAS September 2, 2008 105 (35) 12763-12768; https://doi.org/10.1073/pnas.0806627105
22. Gao, Zheng. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2914-2