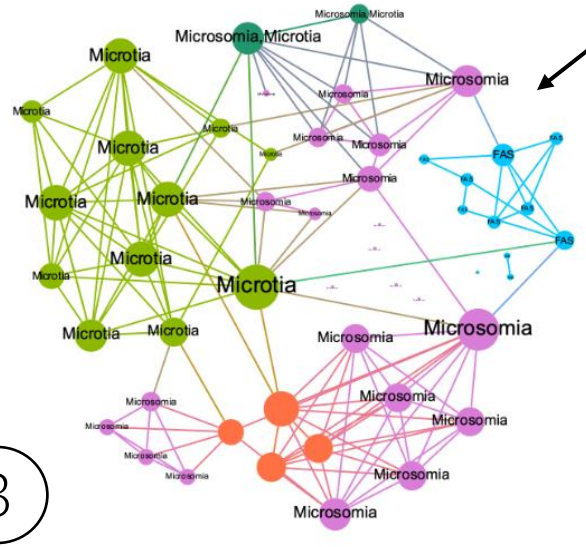
Graphical Abstract

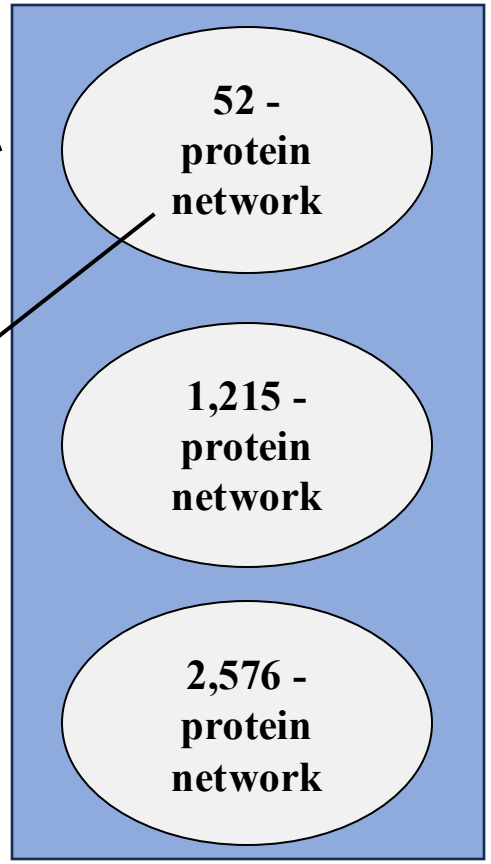
Machine Learning Approaches for Predicting Craniofacial Anomalies with Graph Neural Networks

Methodology

1. Protein networks created from StringDB
2. Disease-Related proteins gathered from FAS, Microsomia, and Microtia



3. Relevant protein features obtained from external databases



Results

Model	52-Protein Network	1,215-Protein Network	2,576-Protein Network
RF	53.5%	72.4%	62.8%
XGBoost	63.5%	72.2%	63.2%
AdaBoost	57.0%	66.0%	63.4%
Stacked	55.0%	67.5%	62.1%
GCN	66.7%	81.3%	63.3%
GAT	69.2%	80.0%	79.7%

Conclusion

The results highlight the powerful potential of Graph Neural Networks (GNNs) for predicting craniofacial anomalies by leveraging protein-protein interaction (PPI) networks. Both GCN and GAT models consistently outperformed traditional machine learning approaches, particularly on larger networks, demonstrating their capacity to effectively capture complex relationships within biological data. The superior performance of GNNs, especially with larger protein networks, suggests that these models are well-suited for the intricacies of disease-related protein interactions.

Machine Learning Approaches for Predicting Craniofacial Anomalies with Graph Neural Networks

Colten Alme[a], Harun Pirim[b], Yusuf Akbulut[b]

[a]*Mechanical Engineering North Dakota State University*
[b]*Industrial and Manufacturing Engineering North Dakota State University corresponding e-mail: harun.pirim@ndsu.edu*

Abstract

This study explores the use of machine learning algorithms, including traditional approaches and graph neural networks (GNNs), to predict certain diseases by analyzing protein-protein interactions. Protein-protein interactions (PPIs) are complex, multifaceted, and sometimes ever-changing. Therefore, analyzing PPIs and making predictions based on them present significant challenges to traditional computational techniques. However, machine learning, particularly GNNs, with their powerful ability to identify complex patterns within large, convoluted datasets, emerge as compelling and revolutionary tools for unraveling these intricate biological networks. We apply machine learning, aided by SHAP explainability and GNNs, on three networks of distinct sizes, ranging from small to large. While the ML results highlight the higher importance of network features in prediction, GNNs exhibit superior accuracy.

Keywords: craniofacial anomalies, machine learning, SHAP, graph neural networks

1. Introduction

Recent advancements in computational biology, data analysis, and machine learning have created vast opportunities for improving healthcare, particularly in the area of disease prediction at the molecular level [19]. One critical area involves studying protein-protein interactions (PPIs), which play a central role in virtually all biological processes. By analyzing PPIs, researchers can detect patterns and abnormalities, which are often linked to

the development of diseases. This ability to observe and interpret these interactions is crucial for advancing early diagnosis, prevention, and treatment of various conditions.

Craniofacial anomalies, constituting approximately one-third of all congenital birth defects, present a diverse range of challenges. These anomalies vary in severity, from mild conditions like shortened palpebral fissures or reduced pinna length, to more severe forms such as craniosynostosis, microphthalmia, coloboma, orofacial clefting, and microtia with aural atresia [8]. Addressing these complex conditions requires a multidisciplinary medical team, including specialists such as neurosurgeons, ophthalmologists, audiologists, oral surgeons, and genetic counselors. Despite the availability of advanced treatment approaches, a significant gap remains in understanding the genetic causes of these defects. A deeper exploration into the genetic basis of craniofacial anomalies is necessary to enhance prevention strategies, facilitate earlier diagnoses, and create personalized treatment plans that improve outcomes for those affected [5].

This research focuses on three craniofacial conditions: Fetal Alcohol Spectrum Disorder (FASD), Microtia, and Microsomia. By applying machine learning algorithms, particularly graph neural networks, the study aims to analyze protein-protein interactions related to these conditions, uncovering shared and distinct genetic mechanisms that may drive these defects. Through this approach, the study seeks to advance knowledge in disease prediction and shed light on the molecular processes underlying craniofacial anomalies.

Fetal Alcohol Spectrum Disorder involve a number of conditions that arise from being exposed to alcohol before birth. Conditions range from physical conditions to behavioral disorders. FASD happens when alcohol in the mother's bloodstream passes through the umbilical cord to the baby. This alcohol interferes with the development of a child's brain, other critical organs, and physiological conditions [15]. This can occur with any amount of alcohol at any stage of pregnancy. Therefore, it is typically advised to stop the consumption of alcohol if a women is pregnant or might become pregnant, due to the fact that people can be pregnant and not know for up to six weeks. Also, if alcohol is consumed, stopping consumption will improve the baby's health as development occurs throughout the entirety of pregnancy. Prevention of FASDs simply occurs by not consuming alcohol during pregnancy.

Like previously stated, there are many different symptoms that can arise

due to Fetal Alcohol Spectrum Disorder. Since alcohol affects a baby's brain development, operations from simple motor control (walking, standing, etc.) to complex executive functions (self-control, organization, etc.) can all be affected [15]. On the same page, individuals with FASD may have smaller heads and a shorter stature due to development being prematurely affected. On top of that, many organ functions can be affected as well, like hearing, vision, heart, and breathing issues. There can be a ranging severity in FASD related to the amount of alcohol consumed, and there are different diagnoses for different levels of FASD.

The next disease of interest is Microtia. Microtia refers to birth defects relating to a baby's ear. The child's ear is abnormal and does not develop properly [22]. Microtia ranges in severity from Type 1 to Type 4. Type 1 Microtia is the most moderate level of Microtia, where the ear retains its normal form, but is smaller than usual. Type 4 Microtia is the most severe, also called Anotia, where all external ear structures are missing. Microtia can affect one or both of a child's ear. Typically, babies born with Microtia do not have damage in their internal ear structures, but, in some cases, babies affected can have a narrow or even missing ear canal [22].

The exact cause of Microtia remains a scientific obscurity. The CDC states that, like many diseases, Microtia occurs due to a genetic change. In some cases, a baby develops Microtia from a single gene alteration. Another potential cause of Microtia stems from the consumption of isotretinoin (Accutane®) during pregnancy [4]. This medication can lead to a number of birth defects, including Microtia. However, the cause of Microtia remains a scientific challenge, and much research is ongoing to attempt to solve this challenge.

The final disease analyzed is Microsomia. In children with Microsomia, part of the face is underdeveloped. This disease is commonly called Hemifacial or Craniofacial Microsomia. Microsomia literally means "smallness" [3]. The majority of Microsomia cases only affect one side of the face, hence the term "hemifacial". The level of deformity can greatly vary from mild to severe from child-to-child. However, Microsomia always involves underdevelopment of the lower jaw. Other underdevelopments, such as issues relating to the eye, ear, cheek, teeth, and other facial features, may occur due to Microsomia as well [3].

Similar to Microtia, Microsomia's cause is widely a mystery. Research leads to the conclusion that Microsomia occurs within the first six weeks of pregnancy, and is due to a disrupted process in the baby's fetal development.

External factors may also be involved in the development of Microsomia.

In this study, proteins associated with the three discussed diseases will be used to construct three distinct PPI networks of varying sizes. By supplementing these networks with information gathered from external databases, multiple ML models will be employed for classification purposes. These models aim to correctly classify each individual protein to its corresponding disease. Additionally, Graph Neural Network (GNN) structures will be utilized in an attempt to enhance accuracy scores.

In this approach, the study introduces a unique integration of protein-protein interaction (PPI) networks and external biological data to improve disease classification. While traditional machine learning models can handle protein features independently, this study leverages the interconnected nature of proteins by using graph-based representations. By embedding proteins within their respective networks, the relationships and interactions between proteins can be fully explored, providing richer data for classification. This approach not only captures individual protein attributes but also incorporates the broader context of their interactions, allowing for a more holistic view of the underlying biological processes. Combining both traditional ML models and GNNs introduces a new way to enhance predictive power, especially as GNNs are designed to analyze these complex networked relationships. This dual approach sets the study apart by combining the strengths of classical ML and advanced deep learning architectures tailored to graph-structured data.

A review of similar studies will be conducted, and their methodologies will be examined. These studies offer valuable insights into effectively conducting this investigation. Following this review, a detailed outline of the study's methodology will be presented. Finally, the results and accuracy scores will be reported and discussed, with considerations for future adjustments.

2. Literature Review

Predicting disease associations and understanding the underlying molecular mechanisms are pivotal in biomedical research. Advances in machine learning and graph neural networks (GNNs) have shown significant promise in tackling these challenges by leveraging the intricate relationships encoded in biological networks such as protein-protein interactions (PPIs).

Numerous studies have explored the potential of machine learning in disease prediction. Grampurohit and Sagarnal [6] analyzed 41 diseases by focus-

ing on patient symptoms closely associated with specific conditions. Using machine learning techniques such as random forests and gradient boosting, their models achieved approximately 95% accuracy in predicting diseases based on symptom input. This study underscored the diverse applications of machine learning in healthcare and its potential benefits. Similarly, Uddin et al. [21] provided a comprehensive review of various machine learning methods for disease risk prediction. They compared models such as regression, decision trees, and artificial neural networks, with a particular focus on Support Vector Machines (SVMs). Their review found that SVMs were especially effective for binary data classification in predicting diseases such as heart disease, Parkinson's, diabetes, and breast cancer.

Machine learning techniques have also been applied to predict PPIs, which are crucial for understanding protein functions and disease mechanisms. Traditional methods such as SVMs, random forests, and artificial neural networks have been used for PPI prediction with significant success [28] [16]. However, these methods often fail to fully capture the complex, multifaceted nature of protein interactions. Recent advancements in GNNs have enabled the integration of PPI network information into predictive models, offering a significant leap in performance, especially in tasks like drug response prediction. GNNs excel in capturing the complex relationships within PPI networks, improving the prediction of interactions and their functional implications [25].

For instance, Jia et al. [9] introduced ModulePred, a deep learning framework that incorporates functional modules and graph augmentation techniques to predict disease-gene associations. By building a heterogeneous module network that integrates disease-gene associations, protein complexes, and augmented protein interactions, ModulePred demonstrates how GNNs can capture both topological and biological features. This addresses limitations in traditional models, such as overlooking the cumulative effects of functional modules or incomplete data. Further developments include MultiPPIs, developed by Zou et al. [29] which predicts PPIs by integrating various biological associations. MultiPPIs achieved high prediction accuracy using DeepWalk, a graph-based feature extraction technique, demonstrating the effectiveness of leveraging multi-source biological data. Similarly, Zeng et al. [27] introduced GNNGL-PPI, which improves PPI prediction accuracy by combining global and local subgraph features via Graph Isomorphism Networks (GINs). This approach addresses category imbalance issues often encountered in PPI datasets, providing a robust solution for multi-category

prediction.

In addition to these methods, Zhao et al. [9] proposed the MGPPI model, which incorporates multiscale graph convolutional neural networks to capture both local and global protein structure information. This model introduces a novel visual explanation method, Grad-WAM, to highlight key binding residue sites, improving interpretability in disease prediction. Zhang et al. [28] also contributed with DSSGNN-PPI, a model that uses double structure and sequence graph neural networks to predict complex PPIs. By combining structural and sequence data, this approach offers a more comprehensive understanding of protein interactions, enhancing predictive accuracy.

Smeriglio et al. [18] provided a GNN-based method for phenotype prediction, incorporating PPI networks and successfully outperforming existing models in predicting Alzheimer's disease. Their work showcases the utility of PPI networks in understanding complex diseases, especially those with multifactorial causes. Lastly, Wang et al. [23] addressed the challenges of constructing accurate graph structures by proposing a hybrid multimodal fusion approach. By integrating raw data with latent embeddings and applying graph pruning techniques, their model significantly improved performance in diagnostic prediction tasks.

The increasing use of GNNs in recent years demonstrates their potential for advancing research in disease prediction and biological network analysis. GNNs provide a unique capability to model complex biological relationships, as seen in these advancements. Despite this progress, however, there remains a notable gap in applying GNNs and PPI analysis to childhood diseases. Research in this area is crucial for understanding early-onset disorders and developing targeted interventions, but it remains underexplored. In this study, we aim to address this gap by applying machine learning, deep learning, and GNNs, to predict specific diseases through the analysis of protein-protein interactions in pediatric populations. PPIs are complex, multifaceted, and dynamic, posing significant challenges for traditional computational methods. By employing GNNs, we can leverage their ability to discern intricate patterns within large and complex datasets. Additionally, we use SHAP (SHapley Additive exPlanations), a powerful tool that explains the output of machine learning models by assigning each feature a contribution value for a specific prediction. This is crucial for interpreting model predictions, particularly in biomedical applications where understanding the role of specific protein interactions can inform clinical decision-making. Our approach includes the application of GNNs to three distinct networks of varying sizes,

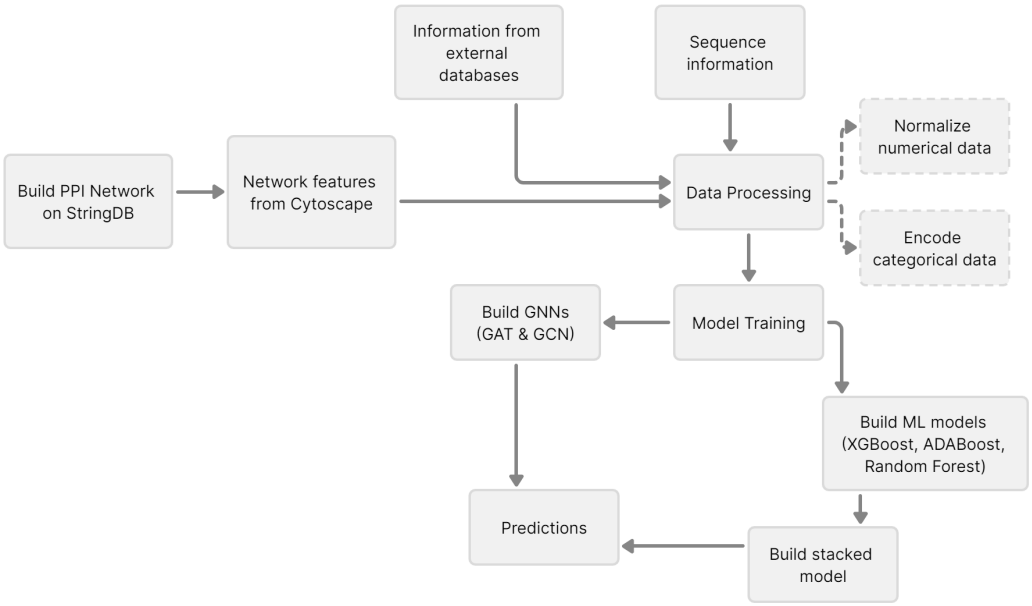


Figure 1: Process Flow Chart

from small to large.

3. Methodology

The flow chart that illustrates the methodology is seen in Figure 1.

3.1. Data Collection

To initiate the data collection process, we aimed to visualize protein networks for the three diseases of interest. Several databases are available for this purpose, but StringDB [20] was selected for its superior visualization capabilities for these specific diseases. StringDB provides excellent information on both known and predicted protein-protein interactions. Using this database, we identified a protein network for each disease individually.

However, the objective was to create an interconnected protein network for all three diseases to enhance the ability to differentiate among them and improve predictive accuracy. By combining the networks, we aimed to capture shared and unique protein interactions that could provide clearer distinctions between the diseases. Instead of looking at each disease's protein network individually, integrating them into one larger network allows us to

observe overlapping proteins, while also highlighting the unique interactions specific to each disease. This approach enhances the model's ability to detect subtle differences between the diseases, ultimately leading to more accurate predictions and insights into disease mechanisms.

To achieve this, we first identified all the proteins involved in each disease's individual protein network using STRING's "proteins by disease" tool. This tool allowed us to compile a comprehensive set of proteins associated with each specific disease by leveraging STRING's extensive database of known and predicted protein interactions.

When selecting proteins to include in each individual disease network, we focused on those that STRING identified as having a significant association with the disease. This was determined by a set confidence score threshold of 70%, ensuring that only high-confidence interactions were considered. This threshold was chosen to balance the need for comprehensive network coverage with the requirement for interaction reliability, thereby excluding lower-confidence interactions that might introduce noise into the analysis.

STRING assigns interactions from two primary sources: curated databases and experimentally determined evidence, which provide a strong, research-backed foundation for protein interactions. Additionally, STRING predicts interactions based on computational methods, including gene neighborhood (genes located near each other on the genome), gene fusions (genes encoding separate proteins fused in certain organisms), and gene co-occurrence (proteins consistently found together across different species) [20]. By applying these criteria, we constructed robust and reliable protein networks for each individual disease.

After constructing the individual disease networks, we proceeded to integrate these proteins into a combined network. We manually entered the identification keys of the proteins from each disease into STRING's "multiple protein" search tool. This tool generates a comprehensive network by linking all entered proteins (nodes) and visualizing the interactions (edges) across the different diseases.

Initially, this approach produced a relatively disconnected network due to the diversity of proteins from multiple diseases. To enhance connectivity, we employed STRING's network expansion tool, which adds additional proteins to fill in gaps and improve network cohesion. We applied only one expansion, carefully balancing the need for connectivity with the importance of maintaining network specificity. This resulted in a well-connected network with numerous protein-protein interactions, ultimately creating a network of

52 proteins (see Figure 2. It's important to note that, during this process, four proteins (highlighted in orange) were identified that were not related to any of the three diseases; their IDs were recorded for further analysis.

In applying this method, we maintained the same stringent criteria as with the individual disease networks, ensuring that only high-confidence interactions were included in the expanded network. This approach allowed us to create a robust and biologically meaningful combined network, effectively capturing the complexity of protein interactions across the diseases studied.

Once this comprehensive protein network was created, we exported it from StringDB's visualizer to Cytoscape [11]. Cytoscape is an excellent tool for visualizing protein-protein interactions and providing important network data.

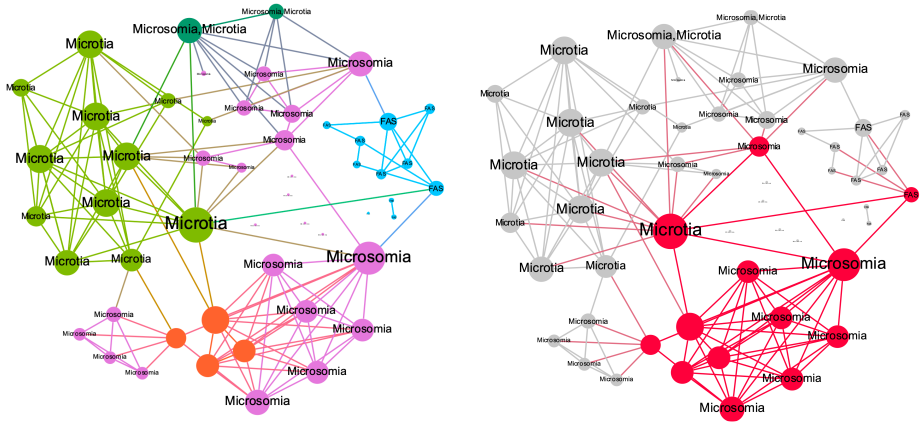


Figure 2: Visualization of Small PPI Network

To facilitate data import into other sources, a comprehensive Excel file was created to consolidate all information regarding the proteins in the network. The first two columns of this file included the protein's ID and the corresponding disease. Subsequent columns contained various data points about the proteins. The initial data imported was sourced from Cytoscape, which provides two types of data: network features and localization predictions. Network features include key values such as connectivity, degree, stress, and eccentricity, which indicate how integrated a protein is within the network. Localization data offers insights into the typical locations of these

proteins within the body (e.g., blood, bone, kidney) and within individual cells (e.g., nucleus, mitochondria, cytosol). Higher values in these localization data points suggest a greater likelihood of the protein being found in those locations, offering valuable information for disease prediction.

After utilizing Cytoscape's tools, additional data was obtained from external sources, beginning with two of the largest protein databases: NCBI and UniProt. These databases provide comprehensive information on proteins across various species. The primary data collected included each protein's sequence, which is essential for employing tools that generate inputs beneficial to the machine learning model. Additionally, protein family information was extracted, aiding in the identification of relationships between proteins and their potential impact on disease presence. Although data from multiple sources were ultimately utilized, UniProt was favored over NCBI for its slightly more detailed information and more user-friendly interface.

Next, the sequence information was harnessed by converting it into FASTA format using Biopython [2], a bioinformatics tool in Python. FASTA format allows for easy manipulation and analysis of protein sequences, representing individual amino acids with single-letter codes. Using Biopython, molecular weight and amino acid counts were calculated, providing insights that could enhance the accuracy of the machine learning model.

Moving beyond NCBI and UniProt, data was gathered from DisGeNET, a discovery platform that offers insights into gene and protein variants and their effects on human disease. DisGeNET, one of the largest gene-disease association repositories, was used to strengthen the Excel dataset. The process involved manually searching for individual proteins and selecting the "Summary of Variant-Disease Associations" section. Variants were sorted by their association score, and the top variant for each protein was recorded, including its ID, type of mutation, specific alleles involved, associated disease, disease class, and association score. This information can reveal connections between proteins, improving the predictive model.

Following this, JensenLabs was utilized to gather further information. Similar to DisGeNET, JensenLabs studies gene-disease relationships. For each protein, the top disease association and knowledge information, along with the corresponding z-score (indicating confidence in the association), were extracted. This data complements the information from DisGeNET and may reveal new connections.

Finally, The Human Protein Atlas was employed. This comprehensive database provides extensive information on all human proteins. While it in-

cludes many visual features, the focus for this data collection was on biological processes and molecular functions, two components of Gene Ontology [1]. Gene Ontology offers a standardized way to describe and represent genes, facilitating a streamlined data collection process and better understanding of genes. For this study, Gene Ontology information provides insights into the functions of individual proteins, potentially revealing previously unknown connections and improving predictive models.

In summary, the additional databases used for this study include UniProt, DisGeNET, JensenLabs, and The Human Protein Atlas. These resources were carefully selected for their comprehensive coverage of protein sequences, variant-disease associations, gene-disease relationships, and functional annotations. The data from these databases were carefully selected and processed to ensure accuracy and relevance, contributing to a dataset that aids the construction of the protein networks and strengthens the overall predictive model.

3.2. Expanding the Network

Initially, a smaller network of 52 proteins was constructed to serve as a proof of concept for the pipeline. The goal of this network was not to produce conclusive results, as the dataset was too limited in size to support robust model performance. Instead, it allowed us to develop, test, and refine the workflow for data processing and model application. By using this smaller network, we ensured that the pipeline was functional and prepared for larger-scale analysis.

After successfully constructing the initial pipeline on the 52-protein network, we sought to extend the analysis to a larger dataset. Returning to the StringDB database, the PPI networks for each individual disease were significantly expanded. STRING database enhances PPI networks by identifying proteins with the highest confidence scores for interactions with existing proteins and adding them to the network. Each disease's network was increased to approximately 400 proteins, and when combined, this resulted in a final network comprising over 1,200 proteins. After accounting for duplicates, the overall network included 1,215 unique proteins (see Figure 3). The primary issue with analyzing the increased size of the PPI network is the data collection process. With the smaller, 52-protein network, data was manually added to the dataset from each database of interest. To handle the larger network, it was necessary to leverage each database's API (Application Programming Interface). The API allows communication between

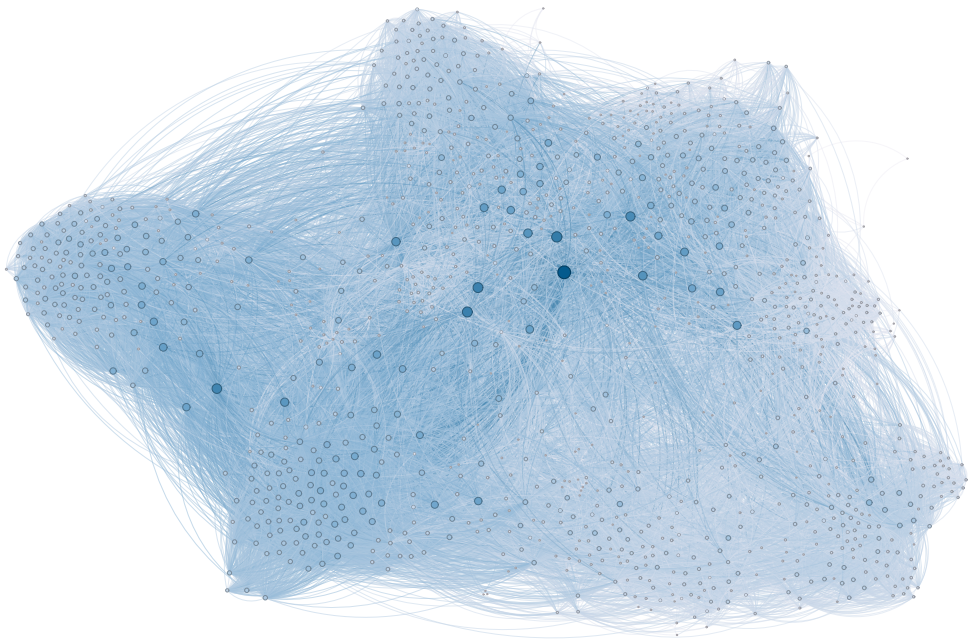


Figure 3: Visualization of Large PPI Network

software programs, enabling the automation of the data collection process. Once all necessary feature information from each database was collected, the same data cleaning and normalization steps were applied as with the smaller set.

While the use of APIs streamlined the data collection process, several issues had to be addressed. For many data components, there were numerous empty results. The presence of empty values in some categories introduces unnecessary noise that could skew accuracy scores. Consequently, some categories of information collected through the API had to be removed from the dataset. However, analysis of the small network indicated that the removed values had a minimal impact on the model's output.

After creating and modeling the complex protein network, a final, larger iteration was completed. This network was generated by creating individual PPI networks for the three target diseases using Cytoscape, rather than the StringDB website. Therefore, this final network was made independently of the first two networks; it was not another expansion of the 1,215-protein net-

work. This network was accomplished through the "Import Network from Public Databases" tool, which is part of the StringDB add-on in the Cytoscape app. This tool allows for the direct import of PPI networks related to the three target diseases into Cytoscape, bypassing the need to export from StringDB. Once the three individual networks were created, the network merger tool in Cytoscape was used to combine them into a final, merged network, which was larger than the previous two.

Since this method of building a network did not directly use the StringDB site, approximately 20 proteins were incomplete, lacking network or localization characteristics as well as sequence information. These incomplete proteins would hinder the models' success, so they were removed, resulting in a 2,576-node network for analysis.

There were a few errors encountered during the data collection process for this final, merged network. One of the primary sources for external information, DisGeNet, was experiencing errors that rendered their API unavailable. As a result, the collection of specific protein attributes could not be achieved. Similar issues were encountered when trying to collect information from other databases. Therefore, for this final iteration, only network, localization, and sequence information was fed into the ML models. Due to this limitation, a drop in prediction accuracy is expected.

To conclude, a summarized flow of the process of creating PPI networks and collecting relevant data, conducted over three iterations for this study is provided. Initially, a 52-protein network was constructed using StringDB. This network was exported to Cytoscape, where localization and network features were harvested. Subsequently, additional databases (NCBI, Uniprot, DisGeNet, Jensen Labs, and The Human Protein Atlas) were utilized to gather supplementary information on these proteins. In the first iteration, data from these databases was manually obtained.

Following the development of the 52-protein network, a more extensive 1,215-protein network was created, marking the second iteration in this three-step process. The protein networks related to each of the three diseases under investigation were expanded using StringDB's network expansion tool, which identifies proteins with high-confidence interaction scores with the existing network proteins. Once each protein network was sufficiently expanded, they were combined in StringDB and exported to Cytoscape, mirroring the procedure of the first iteration. Subsequently, the same databases were used to gather additional information on the proteins. Due to the increased number of proteins, API was employed to enhance the data collection process.

The final iteration involved creating a 2,576-protein network. Unlike the second network, this was not an expansion but rather constructed through the Cytoscape app. After the network's creation, the same localization and network features were collected, and the same databases' APIs were utilized. The primary difference in this iteration was the exclusion of data from DisGeNet due to its API being offline. With the completion of this final step, the data collected from all three iterations is now ready to be processed for its use in ML applications.

3.3. Data Processing

With sufficient and informative datasets created from the three protein sets, data modeling can begin. The first step in creating an effective ML model is data cleaning, which involves removing outliers, addressing missing or infinite values, and correcting poorly formatted data. For this dataset, there were a few missing values, but these were easily corrected as they were attributed to zeros. There were no true outliers or infinite values.

The final aspect of data cleaning is normalizing numerical data. Many machine learning models are sensitive to the scale of data, so it is necessary to ensure all values are properly and similarly scaled to prevent any one column from having more impact than another. This process also keeps the data consistent and manageable for the model. The method used for normalization was Z-score standardization, which adjusts the mean to zero and the standard deviation to one. After standardization, data cleaning was complete.

Next, data transformation is required to successfully run machine learning models. This involves converting all data to Boolean or numerical formats, which are compatible with machine learning models. Numerical data, such as localization information, degree, and closeness, were predominant in this dataset. Boolean data, which has two possible values (0 for false and 1 for true), allows machine learning to incorporate binary decisions into its algorithms.

String data, which includes textual information like disease and disease class, is more challenging for models to handle due to its inefficiency and the variation involved. Therefore, it was necessary to convert string data into a compatible format.

Two main methods were used for transforming string data: label encoding and one-hot encoding. Label encoding assigns a unique integer to each category. For example, a dataset of [“Blue”, “Yellow”, “Red”, “Blue”] would be

transformed to [1, 2, 3, 1]. However, issues can arise if the model assumes an order to the data. Therefore, the only data columns utilizing label encoding in this application were the class, target family, and alleles categories.

One-hot encoding, the predominant method used here, creates separate binary columns for each category. The same dataset would produce columns [1, 0, 0, 1], [0, 1, 0, 0], and [0, 0, 1, 0]. This method is more suitable for machine learning models but can become inefficient with many unique values, leading to potential overfitting. For instance, columns where each protein has a unique value were removed to save efficiency and prevent overfitting. Similarly, columns with unique values, like "Variant ID," were removed as they are not helpful in this context.

With all necessary preprocessing steps completed, the dataset was now in an ideal state for model development. The careful attention to data cleaning, normalization, and transformation ensured that the data was both consistent and suitable for machine learning algorithms. This foundation is crucial, as it maximizes the accuracy and reliability of the models by eliminating noise and standardizing input variables. With the data fully prepared, the next step involved leveraging this dataset to construct and evaluate various machine learning models.

3.4. Machine Learning Application

With the data processed, construction of machine learning models could begin. Three models were created: Random Forest, XGBoost, and AdaBoost. These models were selected for their strong performance in handling complex datasets, widespread popularity, and ease of interpretation, making them among the most effective machine learning models available [14]. Additionally, a stacked model combining all three methods was developed.

Each ML model has inherent advantages and trade-offs. Random Forest (RF) is versatile and handles large datasets well, making it suitable for large PPI networks [17]. RF models also have built-in feature importance tools, enhancing transparency and interpretability. However, they can be slow and overly complex.

XGBoost is often considered a high-performance model, typically outperforming others. It also includes packages to describe feature importance, aiding transparency [17]. However, XGBoost models can be time-consuming and computationally expensive, and they require careful tuning to avoid overfitting, where the model performs well on training data but poorly on new data.

AdaBoost iteratively improves model performance by focusing on misclassifications [7]. It is less prone to overfitting but is sensitive to noisy data, which can distort results. This sensitivity can be a drawback when handling inherently noisy PPI data. Additionally, AdaBoost lacks explainability tools, making it less interpretable than RF or XGBoost models.

The stacked model combines the strengths of Random Forest, XGBoost, and AdaBoost, offering flexibility and strong performance [12]. However, it is extremely complex, time-consuming, and computationally expensive. Its complexity also reduces interpretability.

When determining the best fit for these models, it is crucial to understand their application. Given the presence of noise in PPI data and the importance of feature importance in this project, it is hypothesized that Random Forest and XGBoost will be strong options.

A final note on the construction of the machine learning models relates to data splitting. Splitting the dataset is crucial, involving separating data randomly into "test" and "train" sets. The training set allows the model to learn general patterns, while the test set evaluates the model. For individual models, the data was split in an 80:20 ratio (Train: Test). For the stacked model, the data is in a 70:15:15 split (Train: Validation: Test).

The differing proportions between individual and stacked models arise from the necessity of a validation dataset in the stacked model. Individual models, being simpler, require only a single data split, and an 80:20 ratio provides sufficient data for the model to effectively learn patterns while still allowing for the evaluation of its performance on unseen data. In contrast, stacked models involve greater complexity and necessitate an additional validation step. This validation is crucial for fine-tuning the model, similar to the testing phase, as it provides essential feedback on model performance. Therefore, a 70:15:15 split is optimal for stacked models, ensuring a balanced approach that maximizes accuracy while minimizing the risk of overfitting.

With the data cleaned, transformed, and split, the construction of machine learning models was swiftly executed.

3.5. Graph Neural Networks

Another class of models for disease prediction is Graph Neural Networks (GNNs). Unlike traditional machine learning (ML) methods, which operate on feature vectors, GNNs consider both the features of proteins and the graphical structure of the protein interaction network. Designed to handle

nodes (proteins) and edges (relationships between proteins), GNNs are particularly well-suited for this type of biological data.

In this study, two types of GNNs were employed: Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). Both models are effective but approach the problem differently. GCNs aggregate feature information from a node's neighbors, treating all neighbors equally. This makes GCNs suitable for networks where relationships between nodes are relatively uniform, such as protein-protein interaction networks. GATs, by contrast, use an attention mechanism to assign different importance to each neighboring node, making them more adept at capturing complex, non-uniform relationships between proteins.

We leveraged both GCN and GAT architectures due to their complementary strengths. GCNs were effective in cases where uniform relationships were present, while GATs offered a more refined approach by identifying critical interactions in networks with complex topologies. This dual strategy allowed for a comprehensive analysis of protein networks, ensuring that the study benefited from the unique advantages of both models.

To construct these GNNs, we imported protein interaction networks from Cytoscape as GraphML files, which contained both node and edge data. Protein features were integrated from our dataframe, and the models were trained using cross-validation to fine-tune hyperparameters and assess performance. The GCN model employed two graph convolution layers, a hidden layer of 64 units, and an output layer tailored to the classification task. Meanwhile, the GAT model incorporated an 8-head attention mechanism in the first layer and a single attention head in the output layer, allowing it to better capture nuanced relationships in the network. Both models used Adam optimization and were trained over 100 epochs with a standard 70/15/15 split for training, validation, and testing.

While GNNs are highly effective for graph-structured data, traditional descriptor-based ML models, such as Support Vector Machine (SVM), Random Forest (RF), and XGBoost, often excel in molecular property modeling. A study by Jiang et al. [10] compared these two model categories, showing that descriptor-based models frequently outperform GNNs in terms of accuracy and computational efficiency across various datasets. For example, SVM performed strongly in regression tasks, while XGBoost and RF were reliable for classification [10]. The interpretability of descriptor-based models, especially when paired with tools like Shapley Additive Explanations (SHAP), adds to their practicality, especially in scenarios where model transparency

is critical.

However, in domains like biological data, where the relationships between entities are inherently graph-structured, GNNs still offer significant advantages. While descriptor-based models may be simpler and more interpretable, GNNs are better equipped to model complex relationships within protein networks, aligning with the goals of this research.

Another study by Wu et al. [24] compared the performance of GCN and GAT against advanced GNN models like GraphSAGE, GIN, MoNet, and GatedGCN using benchmark datasets like Cora, PubMed, and ENZYMES. GAT achieved the highest accuracy on the Cora dataset due to its attention mechanism, while GCN outperformed other models on PubMed and DD in terms of training speed and computational efficiency [24]. While GAT's fine-grained attention is advantageous in some contexts, the increased computational cost can be a drawback, whereas GCN provides a balanced trade-off between performance and resource efficiency.

In summary, while traditional ML models and GNNs each have their strengths, the choice of model depends on the specific application. For graph-structured data, particularly in biological networks, GNNs like GCN and GAT are invaluable. Even within the GNN category, selecting between GCN and GAT hinges on task-specific needs, balancing accuracy, computational efficiency, and the complexity of relationships within the data.

4. Results and Discussion

4.1. Small Network - Classical Models

Once all machine learning models were constructed, they were evaluated using accuracy scores. K-fold cross-validation was applied to each ML model to ensure proper training and prevent bias. In k-fold cross-validation, the dataset is divided into "k" subgroups, and the model is trained and tested "k" times, with a different subgroup used as the testing set each time. This approach ensures that every part of the dataset is both trained and tested, making it particularly useful with limited datasets, where accuracy numbers might otherwise be skewed. Overall, cross-validation provides a more comprehensive assessment of ML models.

After applying k-fold cross-validation to the four machine learning models, the results were more consistent and accurate. Although the accuracy scores were generally lower, they were more reliable, indicating proper model evaluation. The Random Forest model achieved a cross-validation accuracy

of 53.5%. XGBoost and AdaBoost produced cross-validation scores of 63.5% and 57%, respectively, while the stacked ML model returned a score of 55%. These results are summarized in Table 1.

While these scores are lower, they align with expectations for initial machine learning models analyzing 52 proteins. It is unrealistic to expect over 70% accuracy with such a small dataset. Scores in the 50s and 60s still demonstrate significant potential for applying ML models to disease prediction. However, improvements such as better handling of proteins belonging to multiple disease classes, eliminating proteins not associated with any disease classes, and increasing the dataset size are necessary.

Another beneficial process with the model results is identifying feature importances. Analyzing which input features are most important in determining the output provides several advantages. It helps in understanding the model better by revealing which features the model relies on for predictions, offering insights into underlying data relationships. Additionally, feature importance can guide future feature selection, identifying and removing irrelevant features to create a simpler, more efficient model.

For these models, network feature characteristics such as degree, eccentricity (maximum distance to another node/protein), and stress (number of paths passing through this node) were the most influential (refer to Figure 4). The features with the least impact on the model were disease association and biological process information. This is likely due to the wide variety of results among the 52 proteins, with around 35 unique results in both columns, making it challenging for any ML model to identify patterns.

An additional method for assessing feature importance was applied to the most accurate model, XGBoost, using SHAP (SHapley Additive exPlanations). SHAP values explain the prediction of an instance by computing the contribution of each feature. A SHAP summary plot combines feature importance with feature effects, showing how much each feature contributed to model predictions, whether positively or negatively. A positive SHAP value indicates that the feature increases the model output, while a negative SHAP value indicates a decrease. The magnitude of the SHAP value indicates the strength of the feature's effect, with higher impact features appearing higher on the SHAP graph.

From the SHAP explainer, shown in Figure 5, the features that impacted the model the most were neighborhood connectivity, nucleus, and degree, followed by other localization and sequence features. It is clear to see that for the simple ML models (RF, XGBoost, and AdaBoost), network features,

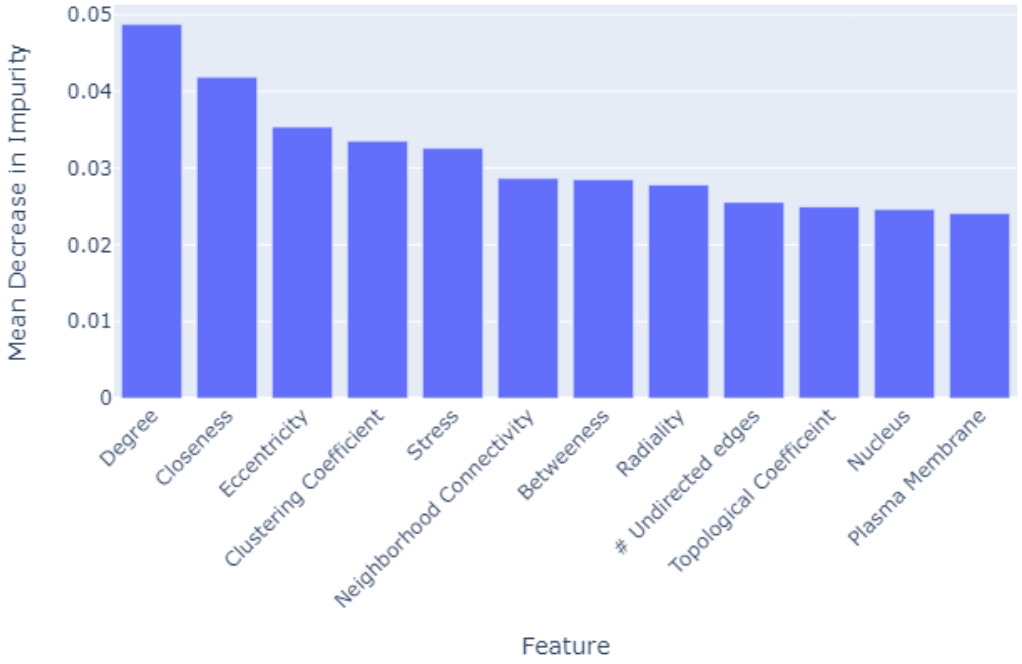


Figure 4: Feature Importances - Small Network

sequence data, and cellular localization information were the most impactful.

Despite network features being the most significant factors influencing disease classification in the small network, some biological insights can still be drawn. For FAS, proteins localized in the nucleus tend to be classified less frequently as FAS. Beyond nuclear localization, no other biological features had a substantial impact on FAS classification within this small network.

In the case of Microsomia, two biological features—bone and pancreas—had the greatest influence on classification. Proteins with high bone localization scores were more frequently classified as Microsomia, while those with high pancreas scores were classified as Microsomia less often. Conversely, proteins with high bone localization scores were classified as Microtia less frequently. This inverse relationship between Microsomia and Microtia is also observed in the enzyme target family feature, where proteins belonging to the enzyme target family are classified more often as Microtia and less often as Microsomia.

Additionally, individual decision trees were examined within the Random Forest ML model using the plot tree feature. This visualizes the first two

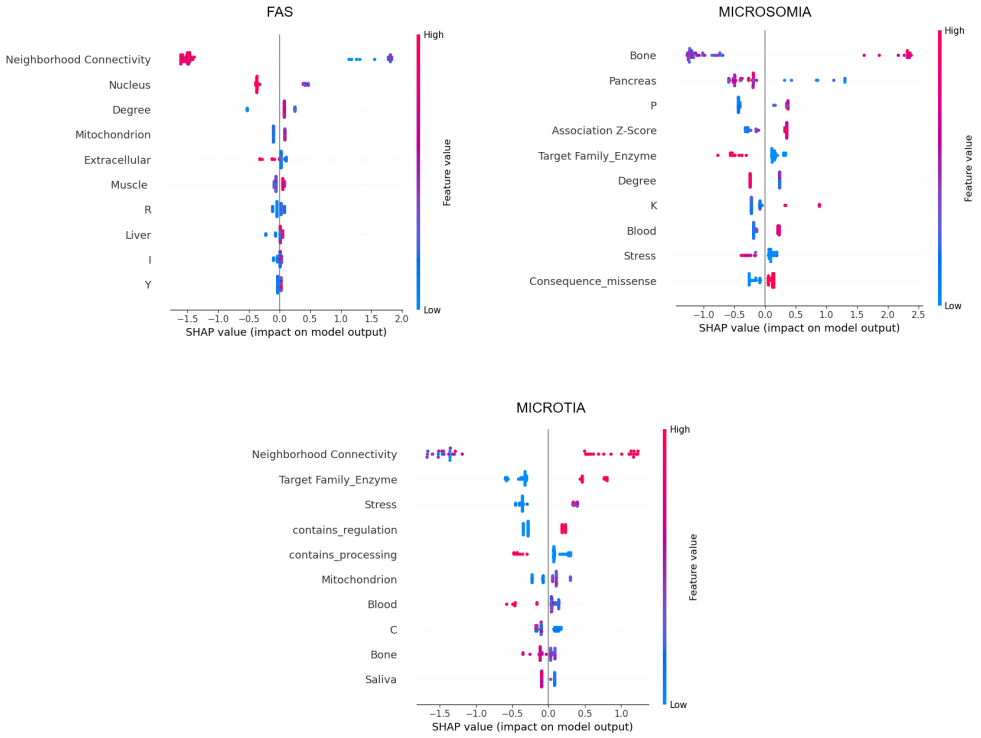


Figure 5: SHAP Values for Small Network

levels of a single decision tree, showing how the root node splits into branches based on certain feature thresholds. This visualization (see Figure 6) provides insight into the decision-making process of one tree in the forest, although it represents only a small fraction of the entire Random Forest model.

For example, one decision tree in the Random Forest model categorized a protein into disease classes based on the presence of the amino acid S in the protein's sequence. The presence of S corresponded to a higher chance of predicting Microsomia. The next feature considered was whether the protein's target family is an enzyme, which split Microtia from Microsomia. Proteins labeled as Microsomia and belonging to the enzyme target family were correctly classified, while those labeled as Microtia were correctly classified due to the absence of an enzyme as their target family. This feature did not affect the classification of FAS disorder. These individual decision trees provide valuable insights into the important features for classifying proteins.

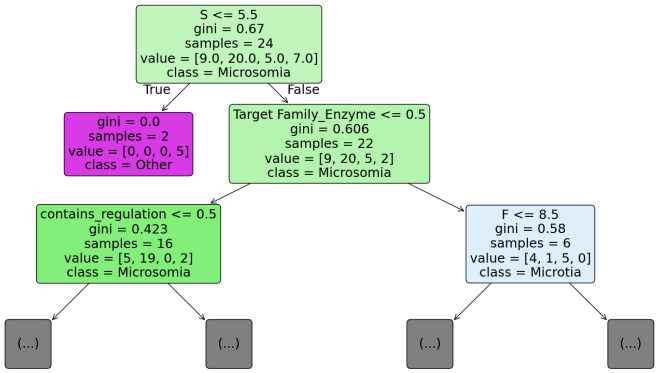


Figure 6: Single Decision Tree within RF Model for Simple Network

These early ML models offer valuable insights into protein features and their corresponding networks. Although the current accuracy scores are too low to identify these models as reliable disease predictors, expanding this knowledge and using more advanced models like graph neural networks, which specialize in handling graphical data such as protein networks, may lead to effective disease prediction models.

4.2. Small Network - GNNs

After gathering results from the four classical ML models, Graph Neural Networks (GNNs) were developed and tested. There was an internal expectation that GNNs would yield slightly higher accuracy numbers due to their suitability for modeling graph-like data such as PPI networks.

This hypothesis was confirmed after creating and testing the accuracy of the GNNs. The Graph Convolutional Network (GCN) achieved an accuracy of 66.7%, which was 3 percentage points higher than the most accurate classical model, XGBoost, which had an accuracy of 63.5%. The Graph Attention Network (GAT) further improved accuracy, achieving 69.2%, the highest score of any model for the simple network. All these results are summarized in Table 1.

4.3. Large Network - Classical Models

The same models were created to analyze the larger, 1215-protein network. The same methods of confirming accuracy scores for the classical ML models, such as k-fold validation, were used on this large network. Due to the larger size of the network, initial accuracy scores were quite similar to the k-fold validation scores. For the Random Forest model, the initial accuracy was 74.8%, and after k-fold validation, the accuracy score finalized at 72.4%.

Similar consistency was observed for the other models as well. The final accuracy for the XGBoost model was 72.2%. The worst-performing model was AdaBoost, with a score of 66%. Interestingly, this 66% accuracy would still be the highest-performing model for the simple network, highlighting the impact of increased network size on ML model accuracy. The stacked model produced an accuracy score of 67.5%. While all four classical ML models showed similar performance, the Random Forest and XGBoost models were the top performers.

The same feature importance methods were applied to the larger network as were used on the smaller network. For the Random Forest model, the features heavily relied upon were consistent across both network sizes.

In both network models, network information (degree, stress, connectivity) remained the most important features for training the model. Cell localization features (nucleus, plasma membrane, mitochondrion) and body localization features (blood, bone, heart) were the next most impactful features. The importance of the top features in the Random Forest model for the larger network is illustrated in Figure 7.

For the XGBoost model, the same SHAP explainer tool was utilized for the large network. Similar features were identified as important for categorizing each protein into individual disease classes, consistent with the feature importances identified by the Random Forest model.

For the classification of Fetal Alcohol Spectrum Disorder (FAS), the most impactful features were nucleus, connectivity, mitochondrion, bone, and clustering coefficient (Figure 8). For Microsomia, the most prominent features were connectivity, enzyme target family, bone, degree, and the presence of amino acids S and I (Figure 8). Lastly, the classification of Microtia was most heavily influenced by the localization features of extracellular matrix, nucleus, and mitochondrion, as well as the network features of degree and closeness (Figure 8).

From the SHAP figures, Figure 8, distinct biological patterns emerge. In FAS, proteins with higher localization in the nucleus appear to reduce the

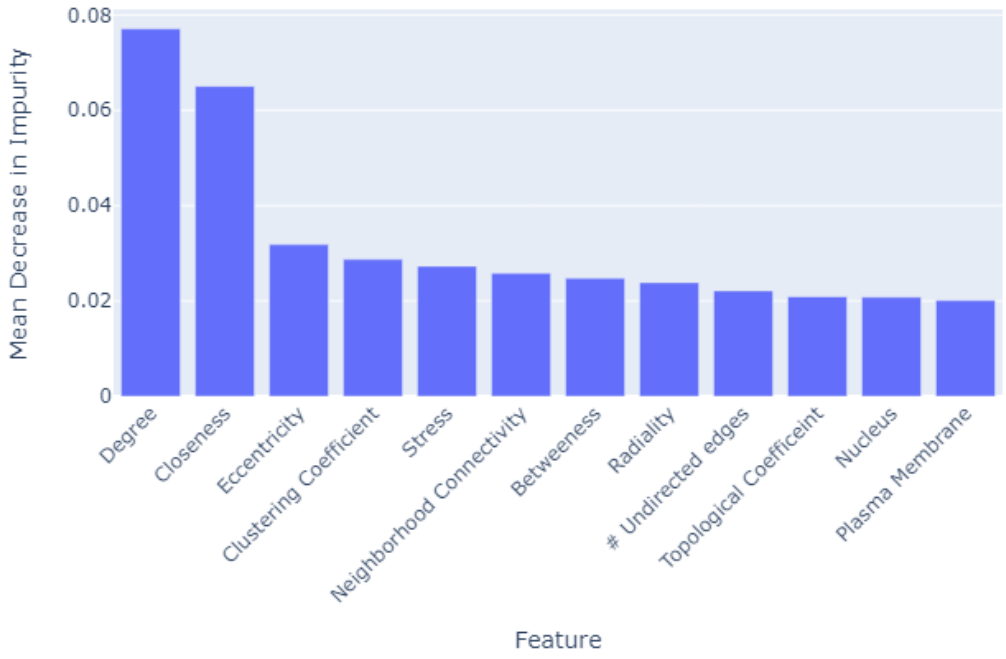


Figure 7: Feature Importances - Large Network

likelihood of the disease being classified. Conversely, in Microtia, proteins with elevated nuclear concentrations increase its classification. A similar inverse relationship is observed with the mitochondrion: proteins with high mitochondrial localization lead to a higher prediction rate for FAS, whereas they result in a lower prediction rate for Microtia. Furthermore, Microtia tends to be associated with proteins present in saliva.

Biologically, Microsomia is frequently classified in proteins localized within the bone, bone marrow, and muscle. Interestingly, proteins belonging to the enzyme target family are less likely to be classified as Microsomia. This stands in contrast to FAS, where proteins within the enzyme target family are more frequently classified as contributing to the disease.

SHAP dependence plots are a powerful tool to visualize how the actual feature values impact the SHAP values (i.e., the contribution of each feature to the model's prediction). These plots allow us to understand the main effects of a single feature and the interaction effects with other features, represented by the color gradient. The y-axis shows the SHAP values for the feature, while the x-axis displays the actual values of that feature. The color

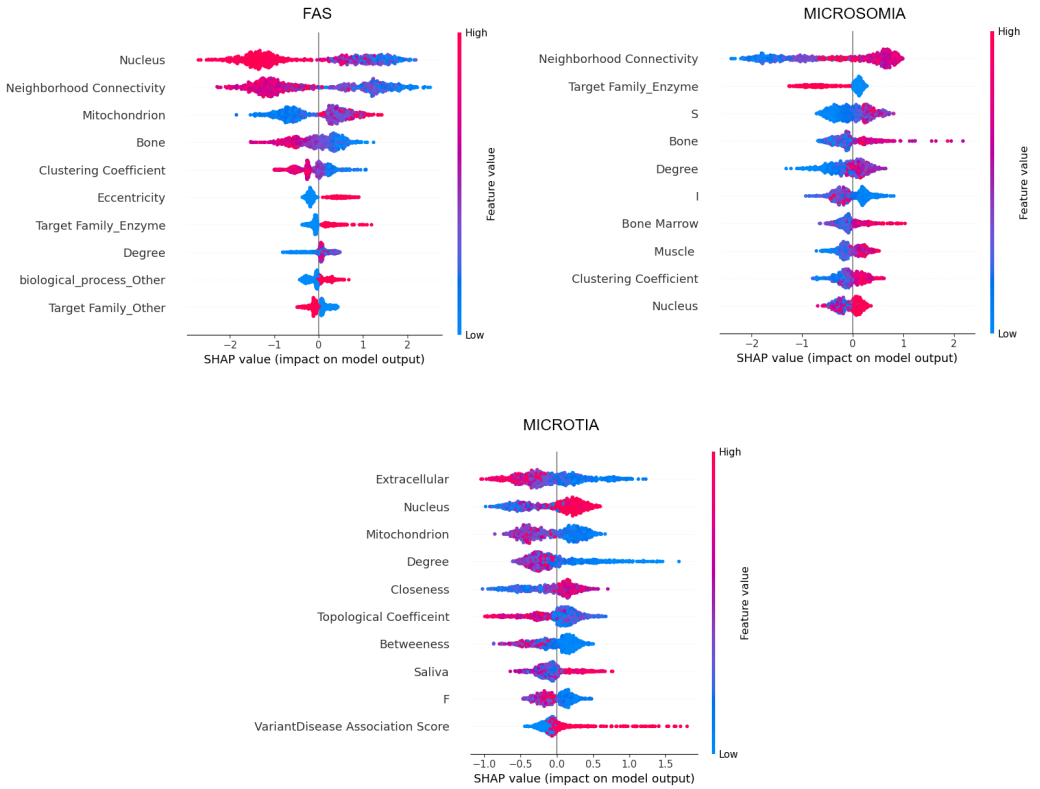


Figure 8: SHAP Values for Large Network

of the dots represents another feature that interacts with the primary feature, highlighting how two variables together influence the model's predictions.

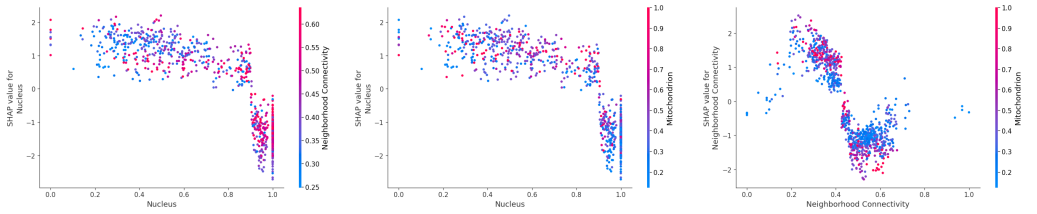


Figure 9: Dependence Plot for Top 3 SHAP Values for FAS Disease - Large Network

In Figure 9 that contains first SHAP dependence plots about FAS Disease, we can see powerful interactions between key features in the model's predictions. The "Nucleus" feature has a strong impact on predictions, especially

at higher values, and its effect is amplified when "Neighborhood Connectivity" is high and "Mitochondrion" is low, indicating strong relationships. As "Nucleus" increases, its contribution grows, particularly when "Neighborhood Connectivity" is elevated, suggesting a cumulative influence. On the other hand, "Neighborhood Connectivity" exhibits a non-linear relationship with predictions, contributing positively at intermediate values, which is further boosted when combined with high "Mitochondrion." Together, these interactions highlight that "Nucleus" plays a central predictive role, significantly influenced by "Neighborhood Connectivity" and "Mitochondrion" in different regions of their value ranges.

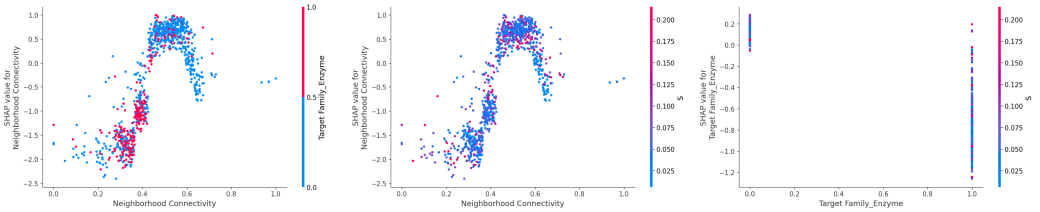


Figure 10: Dependence Plot for Top 3 SHAP Values for Microsomia Disease - Large Network

The second SHAP dependence plots in Figure 10 reveal that Neighborhood Connectivity plays a important role in the model's predictions, showing a non-linear relationship where lower values contribute negatively and higher values (above 0.4) contribute positively. This effect is amplified by its interaction with the S feature, where higher values of S further increase the positive impact of Neighborhood Connectivity. In contrast, the Target Family Enzyme feature, which behaves like a binary variable, has an influence on the predictions, with interaction with Neighborhood Connectivity. It seems that while the existence of the Target Family Enzyme interacts with Neighborhood Connectivity, resulting in a positive contribution, the reverse is also true.

The last SHAP dependence plots in Figure 11 show that Extracellular has a non-linear effect on the model's predictions, where lower values contribute positively, and higher values contribute negatively. This effect is amplified by Nucleus, which increases the negative impact of high Extracellular values. Similarly, Mitochondrion interacts with both features, enhancing the positive influence of lower Extracellular values and further boosting the positive contribution of high Nucleus values. Overall, Nucleus and Mitochondrion form

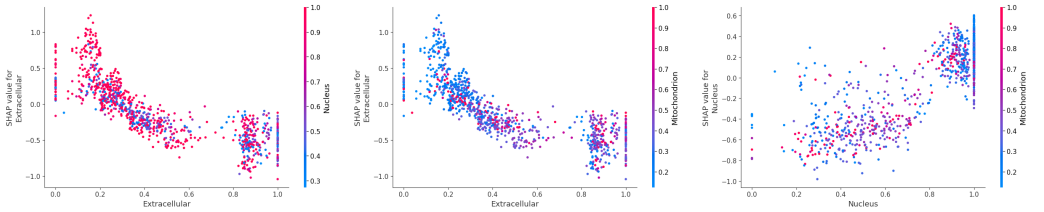


Figure 11: Dependence Plot for Top 3 SHAP Values for Microtia Disease - Large Network

a strong synergistic pair that significantly drives the model's predictions in a positive direction, particularly when Nucleus high and Mitochondrion low values.

4.4. Large Network - GNNs

Once the creation and analysis of ML models associated with the large, 1215-protein PPI network were finished, GNNs were developed. Similar to the smaller network, the two main models created were the GCN and GAT. The only significant difference in constructing the GNNs for the larger network was the increased number of hidden layers to learn more intricate relationships within the data. As with the smaller network, the GNNs achieved the highest accuracy scores for the larger network.

For the GCN of the large network, the accuracy score was 81.3%, the highest accuracy score of any model for any network. The GAT model also demonstrated strong accuracy, with a score of 80%, the second-highest rating. It is evident that in both the small and large networks, GNNs were better equipped to handle these PPI networks and classify diseases accurately.

4.5. Final Network - Classical Models

Once again, the same classical ML models were created and implemented on the final, 2576-node network. As stated previously, this network was made independently of the large, 1215-node network. As with the previous applications, the same models were used with the same k-fold cross-validation processes. The Random Forest model resulted in a predictive accuracy of 62.8%.

This low 60s accuracy score was consistent across the other three classical models. The XGBoost model achieved an accuracy score of 63.2%, AdaBoost had a performance score of 63.4%, and the stacked model scored 62.1%. While consistent, all four models exhibited much lower performance

compared to the large, 1215-protein network. This decrease in performance is likely due to the lack of data collected from external databases, as DisGeNet's API was down during the study of the final network.

When observing feature importances for this final network, many similarities to the previous two iterations were noted. For the top 12 features influencing the Random Forest model, the same network characteristics (degree, stress, eccentricity, etc.) played the largest role. However, unlike the large network importances shown in Figure 7, all of the top 12 features had a relatively equal impact on the model. This can be visualized in Figure 12. After determining the overall feature importances for the models, individ-

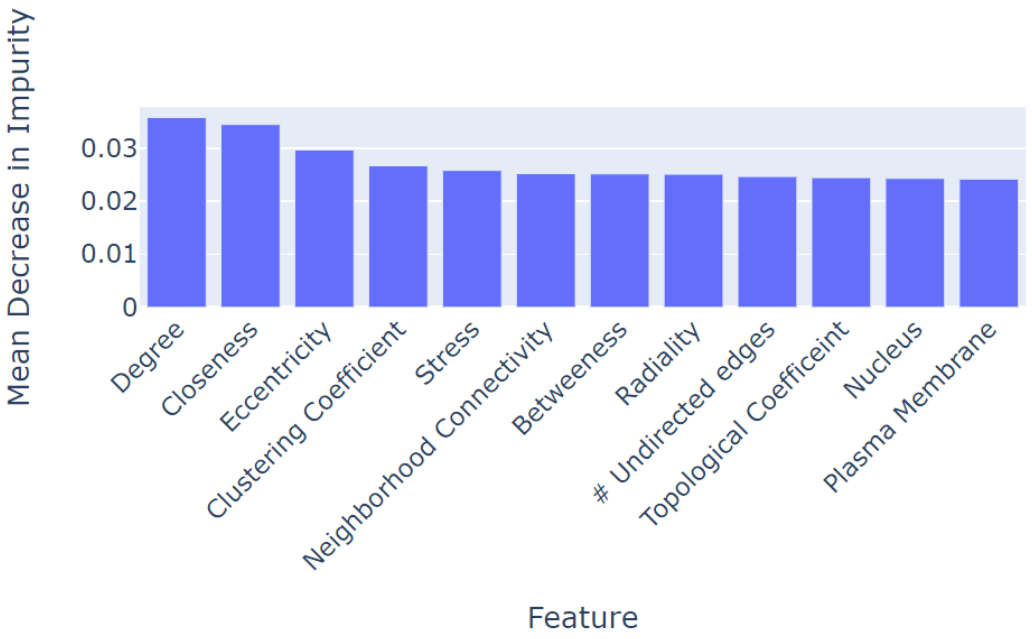


Figure 12: Feature Importances - Final Network

ual disease classification importances were identified using the SHAP tool. These results showed some differences compared to the previous iteration. Bone emerged as the most impactful feature in the classification of FAS and Microsomia. In the large network SHAP charts, bone was still significant but not the top feature. These deviations likely stem from the addition of new, unique proteins to the network or the lack of external information from outside databases. All of this information is visualized in Figure 13.

Similar biological interpretations can be performed on this final network

using visualizations in Figure 13. For both FAS and Microsomia, proteins with high bone localization scores influence their classification, though in opposite directions. In FAS, higher bone localization scores result in fewer classifications, whereas in Microsomia, these scores lead to increased classification. Although bone localization also has a positive effect on Microtia classification, its impact is notably smaller compared to FAS or Microsomia.

Microsomia is significantly affected by proteins localized within the nervous system, with high nervous system localization scores leading to fewer classifications of Microsomia. In contrast, FAS demonstrates a strong positive correlation, where higher nervous system localization scores increase the likelihood of FAS classification.

Across the three different networks, distinct patterns emerge in the role of biological features in disease classification. While network features dominate the predictive power in the small network, biological features still reveal meaningful insights, particularly in relation to protein localization. In both the small and medium networks, FAS is negatively influenced by proteins localized in the nucleus, whereas Microtia shows a contrasting pattern, with higher nuclear concentrations increasing its classification. This inverse relationship is consistent across multiple features. For instance, proteins localized in the mitochondrion lead to higher FAS classification in the medium network, while reducing Microtia classification. Similar opposing tendencies are found between Microsomia and Microtia in relation to the enzyme target family, highlighting how specific biological features can produce divergent disease outcomes.

As network size increases, the trends grow more nuanced, particularly for Microsomia. While bone localization drives Microsomia classification in both the small and large networks, in the large network, the nervous system emerges as a significant factor, reducing Microsomia classifications when proteins have high localization scores within this system. In contrast, FAS, which did not show nervous system involvement in the smaller networks, displays a strong positive correlation with nervous system localization in the large network. This suggests that as the network size expands, the influence of biological features such as tissue localization becomes more pronounced, and different diseases are affected by similar features in diverging ways.

Taken together, these results suggest that certain biological features—such as nuclear, mitochondrial, and bone localization—are consistently influential across network sizes, but their effects can vary significantly depending on the disease. The larger the network, the more pronounced the biological pat-

terns become, especially in complex systems like the nervous system. These findings highlight the importance of integrating both network topology and biological context in disease prediction models. While small networks are predominantly shaped by topological features, biological insights become increasingly relevant in medium and large networks, offering a more detailed understanding of how protein interactions relate to specific diseases.

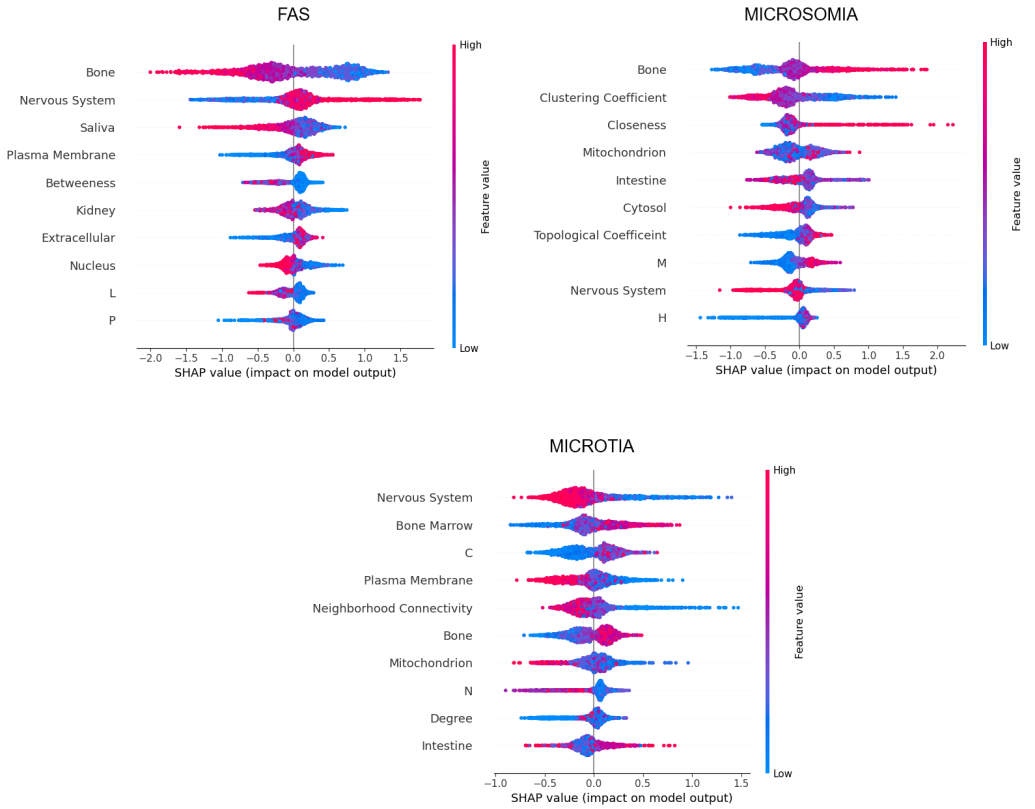


Figure 13: SHAP Values for Final Network

Now, we are performing dependency analysis for the final network, focusing on how key features interact and influence the model's predictions. The first SHAP dependence plots in Figure 14 show that Bone has a non-linear effect on predictions, contributing positively at lower values and negatively at higher values, with minimal interaction from the Nervous System feature. However, Saliva enhances the positive impact of low Bone values, as higher Saliva levels cluster with positive SHAP values for Bone. The Nervous Sys-

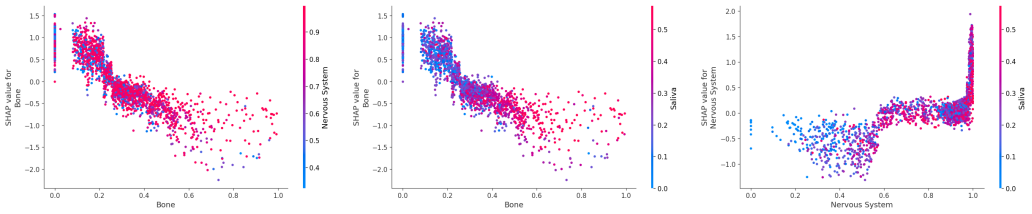


Figure 14: Dependence Plot for Top 3 SHAP Values for FAS Disease - Final Network

tem feature also shows a non-linear pattern, contributing negatively at low values and increasingly positively as its values rise, especially above 0.6. This effect is further amplified by high Saliva values, suggesting a strong positive interaction between Nervous System and Saliva in driving the model's predictions.

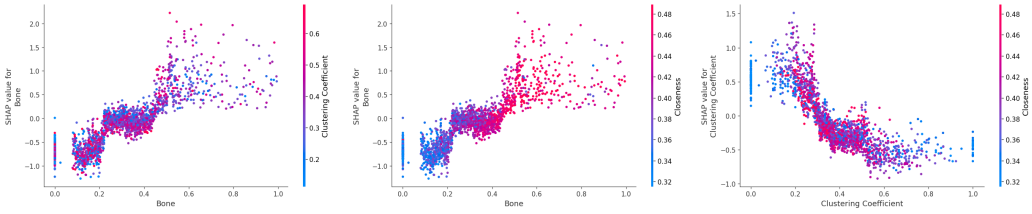


Figure 15: Dependence Plot for Top 3 SHAP Values for Microsomia Disease - Final Network

Second SHAP dependence plots in Figure 15 reveal that Bone has a strong positive linear effect on the model's predictions, with higher values contributing more positively, and this effect is weakly amplified by interactions with Clustering Coefficient and Closeness at mid-to-high ranges. Clustering Coefficient shows a non-linear relationship, where lower values contribute positively to predictions, but higher values contribute negatively. Closeness interacts positively with both Bone and Clustering Coefficient, particularly enhancing the positive contribution of lower Clustering Coefficient values, indicating that Closeness and Clustering Coefficient work synergistically in the lower range to drive the model's predictions.

In the last SHAP dependence plots of Figure 16, the Nervous System feature exhibits a non-linear relationship with model predictions, where lower values contribute positively and higher values (above 0.4) negatively, with minimal interaction from Neighborhood Connectivity or Bone. Similarly, Neighborhood Connectivity shows a non-linear effect, contributing positively

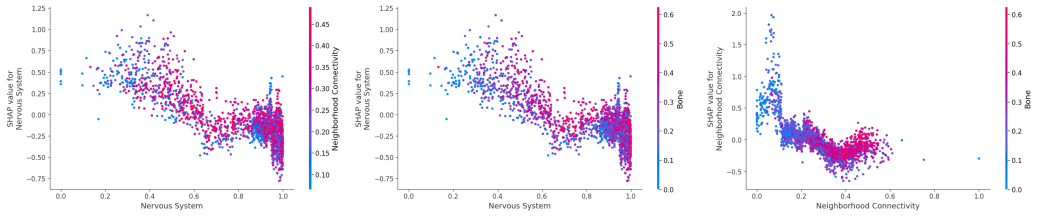


Figure 16: Dependence Plot for Top 3 SHAP Values for for Microtia Disease - Final Network

at lower values but negatively beyond 0.2, where higher Bone values amplify this negative impact.

4.6. Final Network - GNNs

Finally, the same two GNNs, GCN and GAT, were applied to this final 2576-node network. As with the large network, hidden layers were adjusted to uncover more intricate relationships in the data. Consistent with the other two networks, the GNNs were the highest-performing models. The GCN model achieved an accuracy score of 63.3%, while the GAT model attained an accuracy score of 79.7%.

In this iteration, the GCN performed significantly worse than the GAT. Given that the GAT is a more computationally expensive GNN, it is expected to perform better. However, the GAT model's performance being over 16 percentage points higher than the GCN is surprising. This discrepancy could again be attributed to the lack of external database information for this final iteration.

Network	Random Forest	XGBoost	AdaBoost	Stacked	GCN	GAT
Small	53.5%	63.5%	57%	55%	66.7%	**69.2%**
Large	72.4%	72.2%	66%	67.5%	**81.3%**	80%
Final	62.8%	63.2%	63.4%	62.1%	63.3%	**79.7%**

Table 1: ML Model Scores

4.7. Ablation Experiments

To gain a better understanding of the models, ablation experiments were conducted on the top three performing models: XGBoost, GCN, and GAT, using the two largest datasets. Feature ablation was specifically applied

to the XGBoost model. Five distinct sets of features were systematically removed to assess the model's performance: the top five features as identified by SHAP charts, the top ten features based on SHAP charts, all network features, all API features, and a random 10% of the features. This approach provides a comprehensive evaluation of how various feature groups impact the model's performance, helping to isolate the contributions of key features as well as less significant ones. Results of the feature ablation are presented in Table 2.

For the large network, the top five features according to SHAP were degree, nucleus, neighborhood connectivity, bone, and mitochondrion, while the top ten features added clustering coefficient, betweenness, extracellular, saliva, and nervous system. In the final network, the top five features included bone, nervous system, plasma membrane, extracellular, and mitochondrion, with the top ten adding betweenness, nucleus, degree, bone marrow, and clustering coefficient.

	Large	Final
Original Accuracy	72.2%	63.2%
Top 5 SHAP removed	69.7%	61.9%
Top 10 SHAP removed	70.0%	62.0%
All network features removed	70.5%	61.2%
All API features removed	71.9%	62.1%
Random 10% of features removed	71.5%	62.6%

Table 2: Ablation Experiment Results for XGBoost Model of Large and Final Networks

In the large network, removing the top five features resulted in a drop in accuracy to 69.7%, a noticeable decline of 2.5 percentage points. Removing the top ten features led to a similar decline, with the model achieving 70.0%. These results indicate that the top features, as identified by SHAP, play a crucial role in the model's predictions. The removal of all network features caused a slightly smaller drop, reducing accuracy to 70.5%, while the removal of all API features had a more modest effect, bringing the accuracy down to 71.9%. Interestingly, removing a random 10% of features also produced a smaller drop, resulting in an accuracy of 71.5%.

For the final network, the top five feature removal led to a reduction in accuracy to 61.9%, while removing the top ten features resulted in a slightly better performance of 62.0%. The removal of all network features had a more

pronounced impact on this network, bringing accuracy down to 61.2%. In contrast, removing all API features resulted in a more moderate decline, with the model achieving 62.1%. As in the large network, randomly removing 10% of features had a small impact, with accuracy dropping slightly to 62.6%.

In addition to the XGBoost experiments, ablation studies were also conducted on the GNN models (GCN and GAT) for both the large and final networks. Since both models utilized two-layer architectures, three iterations of ablation were performed. The first experiment used the full two-layer architecture. The second iteration involved reducing the model to a single layer, while the third ablation removed the GNN structure entirely, running the model with only the node features. These results are displayed in Table 3.

	Two Layer (Original)	One Layer
GCN - Large	81.3%	65.4%
GAT - Large	80.0%	74.5%
GCN -Final	63.3%	62.2%
GAT - Final	79.7%	73.9%

Table 3: Ablation Experiment Results for GNNs of Large and Final Networks

From the ablation studies, it is clear that the two-layer GNNs performed significantly better across both datasets, with the GCN achieving 81.3% and 63.3% on the large and final networks, respectively. Reducing the models to a single layer caused a noticeable drop in performance, particularly in the large network where GCN dropped to 55.4%. This demonstrates the importance of the second layer in capturing complex relationships within the graph. The final ablation, which removed the GNN architecture entirely, resulted in near-zero performance, indicating the critical role that the graph structure plays in these models' ability to learn meaningful patterns.

Overall, the XGBoost ablation experiments reveal the model's reliance on key features, as shown by the performance decline when removing the top 5-10 SHAP-ranked features. In the large network, accuracy dropped from 72.2% to 69.7% after removing the top five features, and a similar drop occurred when removing the top ten features. While a 3% decrease in accuracy might seem minor, it suggests that XGBoost is fairly robust, handling the removal of important features without a significant loss in predictive power. This resiliency is beneficial in applications where the model may encounter missing or noisy data.

In the final network, the overall performance was lower, but the pattern persisted. Removing all network features or API features also caused modest drops, further illustrating that XGBoost does not overly rely on any single subset of features. While the model remains stable, the fact that removing a random 10% of features had the smallest impact highlights its robustness in less controlled scenarios. This indicates that while key features are important, XGBoost is flexible in handling feature variability.

The GNN ablation experiments demonstrate how model architecture, particularly depth, is crucial for capturing graph patterns. Both GCN and GAT performed best with two layers, with the large network reaching 81.3% and 80% accuracy, respectively. Reducing the models to one layer resulted in significant performance declines, particularly in GCN, where accuracy dropped to 55.4%, emphasizing the need for depth in larger datasets. GAT showed a smaller drop, likely due to its attention mechanism, which compensates for reduced layers.

In the final network, the accuracy decrease was less severe with one-layer models, but still notable. The complete removal of layers (zero-layer models) caused the models to perform poorly, with near-zero accuracy. This reinforces the idea that the GNNs rely heavily on graph-based learning, and merely using node features without leveraging the graph structure is insufficient for predictive tasks.

The results from both XGBoost and GNN ablation experiments highlight the balance between feature importance and model architecture. XGBoost's moderate performance drop when removing top features suggests that while certain features drive predictions, the model retains flexibility. In contrast, the GNNs' sharp performance decline when reducing layers underscores the importance of model depth for graph-based learning. Overall, these findings guide future efforts to fine-tune feature selection in XGBoost and optimize GNN architectures for varying datasets.

4.8. Discussion

The results of our study highlight several important insights regarding the performance of machine learning models when applied to protein-protein interaction (PPI) networks. Initially, the classical models like Random Forest, XGBoost, and AdaBoost yielded moderate accuracy scores across the different network sizes. For the smaller, 52-protein network, XGBoost emerged as the most accurate classical model with a cross-validation accuracy of 63.5%, while the Random Forest and AdaBoost models performed slightly worse.

These results indicate that even with relatively small datasets, these models can provide some predictive capability, but their performance is limited due to the inherent complexity of PPI networks.

As we expanded the dataset to the larger 1215-protein network, the performance of these classical models improved significantly. Both the Random Forest and XGBoost models achieved accuracy scores above 70%, with the Random Forest model slightly outperforming XGBoost. This suggests that the increase in network size provided more comprehensive information for the models to learn from, allowing for more reliable disease predictions. However, even with this improvement, the classical models were ultimately outperformed by the GNNs, which were better suited for the network-based structure of our data.

The most striking results came from the application of Graph Neural Networks (GNNs), particularly the Graph Convolutional Network (GCN) and Graph Attention Network (GAT). The GNNs consistently achieved the highest accuracy scores in both the small and large networks, with the GAT model reaching an impressive 69.2% accuracy in the small network and 80% in the large network. This marked improvement can be attributed to the GNNs' ability to capture the relationships between proteins more effectively, as these models are specifically designed to handle graph-like data. The GAT model, which leverages attention mechanisms to prioritize certain protein interactions over others, proved especially powerful in this context.

One notable observation from our study is the difference in GNN performance across network sizes, with GNNs showing significantly better accuracy in larger networks compared to smaller ones. This disparity is likely driven by two factors: the richness of the feature space and data sparsity. Larger networks provide more comprehensive information about protein interactions, allowing the models to better learn underlying patterns and relationships, whereas smaller networks tend to have less data, which may limit the models' capacity to generalize effectively. Additionally, in smaller networks, the number of connections between nodes is lower, which reduces the GNNs' ability to capture meaningful structures in the graph. To mitigate this, we adapted the GNN models to suit the network sizes by adjusting hyperparameters such as learning rates, hidden layer dimensions, and dropout rates. For smaller networks, we reduced model complexity, decreasing the number of hidden units and increasing dropout rates to avoid overfitting due to the limited data. For larger networks, we allowed for deeper architectures and lower dropout, enabling the models to fully capture the complexity of the pro-

tein interactions. These adaptations helped the models maintain competitive performance across varying dataset sizes, though the inherent differences in data availability and network structure still affected the results.

The performance of different models varied widely across network sizes, with GNNs performing significantly better in large networks than in small ones. This performance difference may stem from insufficient feature space or data sparsity in smaller networks, where fewer connections between proteins lead to less informative graph structures. Additionally, larger networks provide more comprehensive interaction data, which allows GNNs to better capture complex patterns. To address these challenges, we slightly adapted the GNN models, adjusting hyperparameters like the number of hidden units and layers to accommodate the varying complexities of different network sizes. However, further improvements could be achieved by exploring more advanced architectures and regularization techniques specifically designed for smaller graphs. Techniques like graph data augmentation or transfer learning from larger networks could also help improve model performance on smaller datasets, reducing data sparsity issues while maintaining the robustness of the predictions.

When we applied the models to the final, 2576-protein network, we observed a noticeable decline in accuracy for the classical models, which all returned scores in the low 60s. This drop in performance is likely due to the incomplete dataset, as the DisGeNet API was unavailable during this phase of the study. Despite this, the GNNs remained the highest-performing models, with the GAT model once again showing superior accuracy at 79.7%. The discrepancy between the GCN and GAT models' performance in this final network iteration is notable, with the GAT model outperforming the GCN by over 16 percentage points. This difference underscores the importance of using more advanced GNN architectures, such as GAT, when handling highly complex networks with missing data.

These results have significant implications for the application of machine learning to PPI networks and, more broadly, for disease prediction tasks. While classical ML models can offer reasonable performance with sufficient data, their limitations become clear when dealing with more intricate, graph-structured data. The superiority of GNNs, particularly models like GAT that leverage attention mechanisms, highlights the need for models capable of understanding the underlying network structure and the relationships between proteins.

However, several limitations should be acknowledged. One key limitation

of this study is the incomplete dataset used for the largest network, which affected the performance of both classical and GNN models. Access to a more comprehensive dataset would likely lead to better performance and a more accurate comparison between the models. Additionally, while GNNs showed impressive performance, they tend to lack interpretability compared to traditional machine learning models, particularly in terms of feature importance. Unlike models such as Random Forests or XGBoost, where feature importance can be easily derived and interpreted, GNNs are more opaque, making it difficult to identify which specific protein features contribute most to the predictions. This "black box" nature of GNNs poses challenges in clinical settings, where understanding the decision-making process is critical. Another limitation is the computational intensity required to train GNNs, particularly on large networks, which can lead to resource constraints and longer processing times. Furthermore, the scalability of GNNs to handle even larger and more complex networks remains a challenge that requires optimization.

A further limitation is the size of the networks themselves. While efforts were made to expand the dataset using STRING and Cytoscape, the initial 52-protein network was insufficient for conclusive results, serving primarily to create a pipeline for analysis. The expanded network sizes, though more informative, still fell short of the complexity observed in real biological systems. StringDB, while helpful for expanding protein interactions, is inherently limited by its available data, which may not cover all relevant interactions or include emerging data on less-studied proteins. Additionally, the static nature of these networks does not capture the dynamic changes in protein interactions over time or under different conditions. Future work should aim to further extend the size and complexity of these networks, either through additional data sources or through data augmentation techniques, in order to better represent real-world PPI networks and improve model robustness.

As noted, exploring more advanced GNN architectures, such as Fuzzy-Based Deep Attributed Graph Clustering (FDAGC), offers a promising direction for overcoming these challenges. FDAGC excels at clustering tasks by simultaneously considering network structures and node attributes through graph convolution, reducing information loss and improving cluster cohesion via self-monitoring training [26]. Its ability to tightly integrate node embeddings with clustering optimization through fuzzy memberships makes it particularly suitable for protein interaction networks. The combination of

structural and attribute data positions FDAGC as a strong candidate for enhancing predictive accuracy in disease differentiation [26].

Looking ahead, there is significant potential to enhance disease prediction models by integrating graph-based machine learning with electronic health data. As traditional ML methods evolve to handle the complexities of non-Euclidean data, advanced GNN architectures, such as FDAGC, could prove invaluable. Combining these models with electronic health records, disease networks, and multi-omics data—including methylation patterns—will enable deeper insights into disease mechanisms and therapeutic discoveries. [13]. By further refining these models with real-world healthcare datasets and expanding their application to diverse biological data, new predictive capabilities can be unlocked, and previously hidden patterns in disease prediction and drug repositioning may emerge.

In summary, our findings demonstrate that while classical models can provide reasonable performance in predicting disease from PPI networks, GNNs are far better suited for this task, particularly as network size and complexity increase. The GAT model, in particular, showed the highest accuracy across all network sizes, highlighting its potential for future applications in disease prediction. Moving forward, we expect that improvements such as incorporating additional external datasets, enhancing feature selection, and exploring more advanced GNN architectures will further refine the predictive capabilities of these models. The combination of graph-based methodologies with rich healthcare datasets offers a promising avenue for enhancing disease prediction and advancing precision medicine.

5. Conclusion

The exploration into the application of machine learning for studying protein-protein interactions (PPIs) signifies a significant advancement in disease prediction. This merging of computational biology and artificial intelligence has enhanced our understanding of diseases at the molecular level and provided new opportunities for early diagnosis and treatment options.

It is clear that machine learning models, with their advanced ability to process and learn from large amounts of complex data, are well-suited for unraveling PPI networks. These interactions are typically indicative of many diseases, making their analysis crucial for early diagnosis and prognosis.

Different ML models were applied to networks of PPIs related to Microsomia, Microtia, and Fetal Alcohol Spectrum Disorder, three childhood

diseases. These models included Random Forest, XGBoost, AdaBoost, and a stacked model combining the previous three. Through k-fold cross-validation, all four models achieved similar predictive accuracies in the 50-60 percent range. While these results indicate that further refinement is needed, they still show promise in the ability of ML models to predict diseases.

After developing these classical ML models, the focus shifted to Graph Neural Networks (GNNs). Given the graphical nature of protein networks, it was hypothesized that these models would produce higher accuracy results. This was confirmed with the small network, where accuracy scores for the two GNN models were 5-10% higher compared to the classical ML models. While the scores were not high enough to make conclusive interpretations with the small network, they did show promise for the effectiveness of graphical models.

Focusing on feature importance, network-related features were the most impactful on the ML models. Features such as degree, closeness, eccentricity, and clustering coefficient ranked high in both the Random Forest feature importance chart and the XGBoost SHAP models. These observations suggest that FAS, Microtia, and Microsomia are caused by networks of proteins rather than single proteins. The relationships between proteins are more influential in the presence of these diseases than specific information about individual proteins. This conclusion also explains why GNNs were the highest scoring models, as they are built to handle complex graphical data such as PPI networks.

The next step involved greatly expanding the PPI network. Using StringDB to create an overarching PPI network of individual disease networks (FAS, Microtia, and Microsomia), a 1,215-protein network was developed. Using the same general pipeline created in Python, the same classical machine learning models were developed. Scores across the board for these four models hovered around 65-75% accuracy, a 10-20% increase over the small, 52-protein network. Furthermore, GNNs achieved scores of up to 80%.

This noticeable improvement in predictive accuracy for the larger protein-protein network showcased the true potential of these models. This improvement is expected as more information is used to train the ML models. The careful data collection and processing steps ensured that the additional information was relevant and beneficial to the models.

However, while GNNs have demonstrated superior accuracy in predicting diseases from protein-protein interaction (PPI) networks, these models are often under-interpreted in terms of their biological significance. High-

performing models must be accompanied by insights into the underlying biological mechanisms they are modeling, especially in medical and healthcare applications. Simply achieving high accuracy is not enough; the outputs must be understandable to medical professionals for them to be applied in real-world clinical settings.

To enhance the biological interpretability of GNNs, we have employed feature importance analysis to identify critical protein features contributing to predictions, and we have created visual aids that illustrate key interactions within the PPI networks. These tools can effectively highlight the most relevant components and pathways implicated in the diseases studied.

In addition to these methods, linking predictions back to known biological pathways and interactions is essential. Collaborating with domain experts can further ground the GNN outputs in biological reality, ensuring the predictions are scientifically sound and clinically relevant. Additionally, establishing feedback loops where predictions are validated against experimental data or clinical outcomes can help refine the model's applicability.

Integrating these approaches will significantly improve the interpretability of GNN predictions, making them not only accurate but also actionable in healthcare contexts. This enhanced understanding will be key to translating the promising performance of GNNs into real-world applications, ultimately facilitating advancements in disease prediction and treatment.

In the final iteration of this experiment, there was a noticeable drop-off in most models' performance. While surprising at first, this drop in accuracy is understandable given the lack of external database information due to API issues. However, the network shape and features were still available, explaining why the top-performing GNN model still rivaled that of the large network.

Other challenges experienced throughout the process included dealing with the vastness of PPI data and determining which pieces are relevant, potential biases within the data, and the inevitable noise that comes with biological data. Despite these challenges, innovations in data handling and modeling continue to improve ML model accuracy. As PPI datasets become increasingly available and comprehensive, the integration of machine learning will continue to advance the field of disease prediction.

Moving forward, it is important for researchers worldwide to acknowledge the great potential of machine learning in revolutionizing disease prediction and treatment. The information learned from analyzing PPIs of childhood diseases proves this potential. Additionally, the benefit of analyzing PPIs

is extensive, enhancing the ability to cater treatments to patients' needs and leading to the development of new treatment plans. However, it is also important to note the potential pitfalls and limitations of applying machine learning in disease prediction.

To conclude, the application of machine learning in analyzing protein-protein interactions presents a unique and promising opportunity for predicting and combating diseases. It represents a significant step towards personalized medicine, where treatments and disease predictions can be tailored to individual genetic profiles, ushering in a new era of healthcare.

5.1. Repository

All working codes, data, and documents related to this study can be found in `https://github.com/CEL-lab/PPI_GNN`

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT 4o in order to proofread, paraphrase, and code edits. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgement

This research is supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103442.

References

[1] Michael Ashburner, Catherine A Ball, and Judith A Blake. Gene ontology: tool for the unification of biology. *Nat Genet*, 2011.

[2] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009.

[3] R.R.J. Cousley and M.L. Calvert. Current concepts in the understanding and management of hemifacial microsomia. *British Journal of Plastic Surgery*, 50(7):536–551, 1997.

[4] Centers for Disease Control and Prevention. 2023.

[5] Emily L Geisler, Saloni Agarwal, Rami R Hallac, Ovidiu Daescu, and Alex A Kane. A role for artificial intelligence in the classification of craniofacial anomalies. *Journal of Craniofacial Surgery*, 32(3):967–969, 2021.

[6] Sneha Grampurohit and Chetan Sagarnal. Disease prediction using machine learning algorithms. pages 1–7, 2020.

[7] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[8] Jeremy A Hunt and Craig P Hobar. Common craniofacial anomalies: the facial dysostoses. *Plastic and reconstructive surgery*, 110(7):1714–1725, 2002.

[9] Xianghu Jia, Weiwen Luo, Jiaqi Li, Jieqi Xing, Hongjie Sun, Shunyao Wu, and Xiaoquan Su. A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation. 25(1):214.

[10] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13:1–23, 2021.

[11] Michael Kohl, Sebastian Wiese, and Bettina Warscheid. Cytoscape: software for visualization and analysis of biological networks. *Data mining in proteomics: from standards to applications*, pages 291–303, 2011.

[12] Gautam Kunapuli. *Ensemble Methods for Machine Learning*. Simon and Schuster, 2023.

[13] Haohui Lu and Shahadat Uddin. Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends. In *Healthcare*, volume 11, page 1031. MDPI, 2023.

[14] Varsha Nemade and Vishal Fegade. Machine learning techniques for breast cancer prediction. *Procedia Computer Science*, 218:1314–1320, 2023.

[15] Edward P Riley, M. Alejandra Infante, and Kenneth R. Warren. Fetal alcohol spectrum disorders: An overview. *Neuropsychology Review*, 2011.

[16] Debasree Sarkar and Sudipto Saha. Machine-learning techniques for the prediction of protein–protein interactions. *Journal of biosciences*, 44(4):104, 2019.

[17] Zhenfeng Shao, Muhammad Nasar Ahmad, and Akib Javed. Comparison of random forest and xgboost classifiers using integrated optical and sar features for mapping urban impervious surface. *Remote Sensing*, 16(4), 2024.

[18] Riccardo Smeriglio, Joana Rosell-Mirmi, Petia Radeva, and Jordi Abante. Leveraging protein-protein interactions in phenotype prediction through graph neural networks. *bioRxiv*, 2024.

[19] Zhenchao Sun, Hongzhi Yin, Hongxu Chen, Tong Chen, Lizhen Cui, and Fan Yang. Disease prediction via graph neural networks. *IEEE Journal of Biomedical and Health Informatics*, 25(3):818–826, 2020.

[20] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 11 2022.

[21] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 2019.

[22] Daan P.F. van Nunen, Mischka N. Kolodzynski, Marie-José H. van den Boogaard, Moshe Kon, and Corstiaan C. Breugem. Microtia in the netherlands: Clinical characteristics and associated anomalies. *International Journal of Pediatric Otorhinolaryngology*, 78(6):954–959, 2014.

[23] R. Wang, W. Guo, Y. Wang, X. Zhou, J.C. Leung, S. Yan, and L. Cui. Hybrid multimodal fusion for graph learning in disease prediction. *Methods*, 229:41–48, Sep 2024. Epub 2024 Jun 14.

[24] Junwei Wu, Jingwei Sun, Hao Sun, and Guangzhong Sun. Performance analysis of graph neural network frameworks. In *2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 118–127. IEEE, 2021.

[25] Yuan Xie, Jian Peng, and Yuan Zhou. Integrating protein-protein interaction information into drug response prediction by graph neural encoding. 2019.

[26] Yue Yang, Xiaorui Su, Bowei Zhao, GuoDong Li, Pengwei Hu, Jun Zhang, and Lun Hu. Fuzzy-based deep attributed graph clustering. *IEEE Transactions on Fuzzy Systems*, 32(4):1951–1964, 2024.

[27] Xin Zeng, Fan-Fang Meng, Meng-Liang Wen, Shu-Juan Li, and Yi Li. Gnngl-ppi: multi-category prediction of protein-protein interactions using graph neural networks based on global graphs and local subgraphs. *BMC Genomics*, 25(1):406, May 2024.

[28] Fan Zhang, Sheng Chang, Binjie Wang, and Xinhong Zhang. Dssgnn-ppi: A protein–protein interactions prediction model based on double structure and sequence graph neural networks. *Computers in Biology and Medicine*, 177:108669, 2024.

[29] Hai-Tao Zou, Bo-Ya Ji, and Xiao-Lan Xie. A multi-source molecular network representation model for protein–protein interactions prediction. *Scientific Reports*, 14(1):6184, March 2024.