Towards a Multi-Omics approach for Disease-Causing Molecules Identification using Graph Convolutional Network

Submitted By : Anjana BHAT

Supervisor : Silvia BOTTINI

Advisor : Michel RIVEILL

Master of Data Science and Artificial Intelligence (MSc1)
University of Cote d'Azur

31 July 2023

Abstract

The rapid advancement in high-throughput biomedical technologies has enabled the collection of various types of omics data with unprecedented details. While each omics technology can only capture part of the biological complexity, integrating multiple types of omics data can provide a more holistic view of the underlying biological processes. Graph deep learning has recently emerged to incorporate graph structures into a deep learning framework to predict disease causing molecules. The recent developments regarding Graph Convolution Network (GCN) models are a promising resource to enhance the application of multi-omics studies on rare diseases. Despite the promising results shown by some studies using GCNs in the field of oncology, there is a scarcity of applications in rare diseases, particularly in mitochondrial diseases. Moreover, intellectual disability, another rare disease category, has also received limited attention in the context of multi-omics studies. This neglect can be attributed to various challenges, such as the limited availability of data, the curse of dimensionality inherent in omics data, and the heterogeneous nature of the data associated with these conditions. Consequently, there is a need to address these challenges and explore the potential of GCNs in advancing research and understanding of rare diseases, including mitochondrial diseases and intellectual disability. In order to address this research gap, I have developed a novel GCN model that combines Graph Convolution and GraphSAGE. This model was applied to analyze ten distinct modules of omics data related to mitochondrial diseases, with the aim of accurately classifying genes as pathogenic or non-pathogenic. By employing the best edge weight method tailored to the complexity of each network, our model achieved impressive accuracy ranging from 98% to 100% across the different modules. To further evaluate the model's performance, I also applied it to intellectual disabilities pathology and obtained a remarkable accuracy of 100% on validation set for JS divergence as the best edge weight method. The success of the implemented model in classifying genes as pathogenic or non-pathogenic in both pathologies further highlights the potential of graph deep learning for advancing multi-omics studies and aiding in disease causing molecules diagnosis and treatment.

Keywords: Multi-omics Integration, Graph Convolutional Networks, Deep Learning, GraphSAGE, mitochondrial diseases, Intellectual Disabilities, disease-causing genes.

Contents

1 Introduction

Modern data-driven biology relies on the interpretation of large-scale molecular measurements (known as omics data) to understand and predict biological phenotype, such as the characteristics of an organism or the state of an individual cell. The 'omics' notion indicates that nearly all instances of the targeted molecules are measured in the assay providing holistic views of the biological system. Omics studies identify, characterize and quantify bio-molecules involved in the structure, function, and dynamics of cells, tissues, or organisms. Compared to single omics interrogations, multi-omics can provide researchers with a greater understanding of the flow of information, from the original cause of disease to the functional consequences or relevant interactions. By combining multiple types of omics data, multi-omics studies can reduce noise, increase statistical power, and enhance the reliability of findings. The integration of complementary information from different data sources helps validate and reinforce the biological insights derived from individual omics data. In this regard, biological networks can be treated as graphs, where nodes represent genes and connections between nodes represent gene–gene interactions, whereas multi-omics data levels can be seen as feature vectors of genes. These networks are useful to predict the disease causing molecules with multi-omics more than using a single type of omics data [1]. In particular, Graph Convolutional Networks (GCNs) are able to classify unlabelled nodes in a network on the basis of both their associated feature vectors, as well as the network's topology, making it possible to integrate graph-based data with feature vectors in a natural way. However, there are still open challenges, especially in the biomedical datasets including missing values, dataset size, curse of dimensionality etc. Such challenges hinder the optimal performance of any model for the classification task. One of these challenges is also the class imbalance which happens when the number of samples across the classes are disproportional [2].

2 Related Works

Graph Convolutional Network is a neural network architecture that works with graph data. The main goal of GCN is to distill graph and node attribute information into the vector node representation aka embedding. GCNs leverage the graph structure to capture relationships and dependencies between nodes, enabling effective learning on graph data. By performing convolution-like operations on the graph, GCNs can learn node representations that incorporate both local and global graph information, making them powerful tools for tasks such as node classification, link prediction, and graph classification. With the help of GCN, many state of the art performances has been achieved that revolves around classification problems in medical field, as mentioned in [3], [4]. Moreover, EMOGI (Explainable Multi-Omics Graph Integration) is a graph convolutional network (GCN) that prioritizes cancer genes by utilizing multi-omics features as network nodes and protein-protein interaction networks for topology. By combining these elements, EMOGI effectively identifies mutations in cancer genes, genes with alterations, and genes interacting with known cancer genes. It has successfully discovered 165 potential new cancer genes that interact with known ones, offering valuable insights into cancer genetics

and potential therapeutic targets [4].
Low-dimensional embedding of nodes in large graphs have proved extremely useful in a variety of prediction tasks, from content recommendation to identifying protein functions. However, most existing approaches require that all nodes in the graph are present during training of the embedding; these previous approaches are inherently transductive and do not naturally generalize to unseen nodes. Graph-SAGE, a general inductive framework that leverages node feature information (e.g., text attributes) to efficiently generate node embedding for previously unseen data. Instead of training individual embeddings for each node, Graph-SAGE learns a function that generates embeddings by sampling and aggregating features from a node's local neighborhood. [5].
Another recent model is SAGE-GCN, that is a novel dynamic Graph Convolutional Network model incorporating a Self-adaptive Stable Gate consisting of a state encoding network and a policy network. SageGCN combines Graph-SAGE's neighborhood aggregation with self-attention mechanisms, which enable nodes to dynamically weigh the importance of their neighbors during aggregation [6].

3 Biomedical Context

3.1 Omics

Omics refers to a broad field of scientific disciplines that focus on the comprehensive study of biological molecules and their interactions within living systems. It involves the analysis of large-scale data sets encompassing various aspects of biological information, such as genomics, transcriptomics, proteomics, metabolomics, and more. Genomics explores the complete set of genes in an organism, while transcriptomics examines the expression of genes through RNA molecules. Proteomics investigates the structure, function, and interactions of proteins, and metabolomics studies the metabolites and small molecules involved in cellular processes. By analyzing these different omics data sets, researchers can gain a holistic understanding of the complex molecular mechanisms underlying biological systems. Omics approaches have revolutionized biological research, enabling insights into diseases, personalized medicine, and the development of novel therapeutic interventions. The cascade of events in omics is depicted in 1 [7].
In addition to the mentioned omics data types, we have also incorporated translatomic data in this research project. Translatomics involves studying the actively translated RNA molecules within a cell or tissue, providing crucial insights into the dynamic process of protein synthesis. Unlike transcriptomics, which examines all RNA molecules, translatomics focuses specifically on the subset of RNA molecules that are actively undergoing translation to produce proteins. With the expanding knowledge and insights gained from various omics fields such as transcriptomics and proteomics, the need for effective data integration arises, allowing researchers to combine and analyze these vast datasets in a cohesive manner. Data integration is the process of combining information from different sources to gain comprehensive insights into biological processes and disease. Integrative multi-omics approaches enhance biological discoveries by highlighting the most relevant features and understanding the combined influence of different omics levels on biological processes,

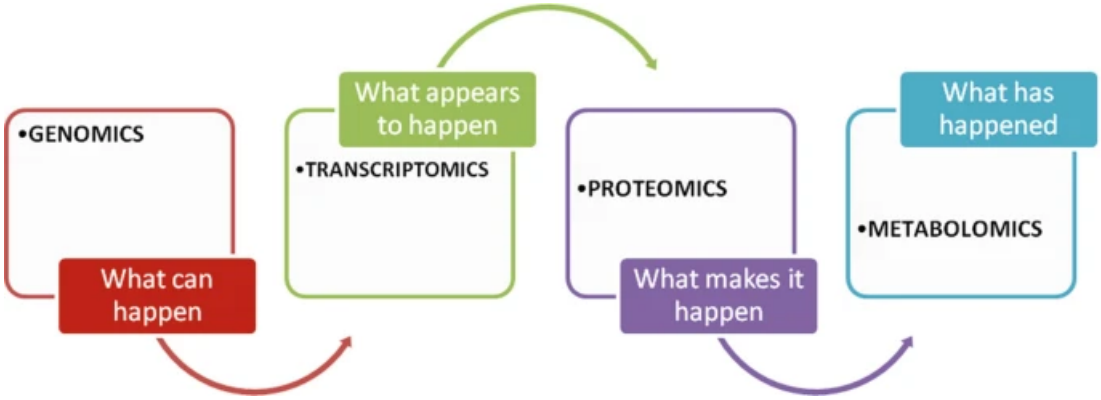


Figure 1: Cascade of Omics Development

although the integration of diverse multi-omics data remains challenging. Moreover, integrating multi-omics data with advanced computational methods, such as GCNs, not only uncovers underlying molecular mechanisms but also fosters knowledge sharing and collaboration within the research community.

3.2 Mitochondrial diseases

Mitochondrial diseases (MD) are rare disorders caused by deficiency of the mitochondrial respiratory chain, which provides energy in each cell. These diseases are extremely heterogeneous, both clinically and genetically, making their diagnosis a real challenge. They are characterized by a high clinical and genetic heterogeneity and in most patients, the responsible gene is unknown. Diagnosis is based on the identification of the causative gene that allows genetic counseling, prenatal diagnosis, understanding of pathological mechanisms, and personalized therapeutic approaches. Disease causing molecules are diagnosed when the gene responsible for the disease's development is identified. The process of identifying possible responsible genes is called gene prioritization. Gene prioritization is done by examining features such as the gene expression and function, its interaction with other genes, and if the gene has some mutations. However, Mitochondrial diseases exhibit genetic heterogeneity, with mutations occurring in multiple genes associated with mitochondrial function, that becomes a challenge to diagnose the diseases. Hence, multi-omics methods are emerging as valuable tools for understanding the functioning of the mitochondria [8]. But the challenge is to integrate the different omics data as they are heterogeneous. One more challenge is the curse of dimensionality since there are many genes but few samples, and this is even worse in the application of mitochondrial diseases, with some studies having very few patients. Even if there is data available, pathogenic or responsible genes are often unknown that leads to unlabeled data. In such cases, either unsupervised methods has to be used or needs to consider semi-supervised methods such as GCN that can learn from minimal train set and manages to generalize the unlabeled data.

3.3 Intellectual Disabilities

Intellectual disability refers to significant limitations in intellectual functioning and adaptive behavior. These limitations manifest during the developmental period and impact an individual's ability to effectively participate in everyday activities [9]. Molecular sub-typing through gene sequencing, gene expression, and other epigenetic and omics data has been used with great success in neurodevelopmental disorders to classify sub-types for more effective treatment, understanding prognosis, and identifying disease mechanisms [10]. However, Intellectual disabilities are relatively rare conditions, and finding an adequate number of participants for research studies can be challenging. Hence, limited sample size and phenotypic variability pose challenges in obtaining statistically robust and generalizable intellectual disability data. Also, integrating and analyzing multi-omics data from diverse omics layers requires specialized computational approaches and bio-informatics pipelines.
Collaborative efforts and data sharing among researchers can effectively address the challenge of limited sample sizes in intellectual disability research. Furthermore, longitudinal studies provide valuable data for understanding the developmental trajectories and variability of these conditions. Additionally, establishing data repositories plays a crucial role in enhancing access, statistical power, and generalizability of research findings.

The Fragile X-Syndrome (FXS) represents the most common inherited form of intellectual disability and the first monogenic cause of Autism Spectrum Disorders, affecting 1/4000 males and 1/7000 females. It is caused by the absence of the FMRP protein, encoded by the FMR1 gene. FMRP plays a crucial role in regulating the translation of thousands of messenger RNAs (mRNAs) involved in synaptic development and function. Despite its primarily cytoplasmic localization, FMRP has also been linked to nuclear functions, including splicing regulation. Given its intricate involvement in various cellular processes and interactions with numerous mRNA targets, the contribution of FMRP to neuronal physiology is complex and involves potentially many biological pathways. To better understand the molecular perturbations underlying FXS, the research community has conducted numerous omics studies, including genomics, transcriptomics, proteomics, and metabolomics. Integrating multi-omics data using bioinformatics and artificial intelligence techniques has become a valuable approach to comprehensively explore complex biological mechanisms and diseases, providing holistic views and insights into FXS pathophysiology.

4 Objectives

This report presents the comprehensive project work conducted at the Medical Data Laboratory (MDLab). MDLab was established through collaboration between data science and medical researchers at the Center of Modeling, Simulation, and Interactions (MSI), as part of the UCAJEDI IDEX project. Located in Nice University Hospital (CHUN), MDLab focuses on processing vast medical and environmental databases to address bio-informatics, mathematical modeling, and biostatistics-related medical research challenges. Supported by a partnership between Centre Hospitalier de Nice - CHUN and Université Côte d'Azur, MDLab integrates

medicine, biology, bio-informatics, and machine learning to implement an integrated multi-omics approach in fields such as genetics, rare illnesses, cancer, health, and the environment. This ongoing project at MDLab aims to introduce new diagnostic tools and alleviate diagnostic uncertainties for patients.

The objective of my internship is to develop and implement a Graph Convolutional Network (GCN) model for multi-omics integration. The key tasks involved in this project include training the model on various biological systems, identifying best weight representations for node interrelationships, and testing its performance on experimental data. Through these activities, the internship aims to enhance competencies in working with omics data, particularly proteomic and transcriptomic data, utilizing established Python libraries for GCN architectures, creating an integrated package for GCN and multi-omics integration, and conducting benchmarking tests to evaluate the performance of different strategies.
To achieve these objectives, a GCN model incorporating Graph Convolution and GraphSAGE layers was implemented. The model's performance was evaluated on two distinct pathologies: mitochondrial diseases, which involved a single omics dataset consisting of transcriptomic data, and intellectual disability, which encompassed multiple omics datasets including transcriptomic, proteomic, and translatomic data. In order to determine the optimal weight representation for node interrelationships, a comprehensive analysis of 14 different approaches was conducted. Further details regarding these investigations can be found in subsequent sections of this report.

5 Dataset

5.1 Mitochondrial diseases

5.1.1 Patient cohort

RNA-sequencing experiments have been done on a set of 20 patients. It allows us measuring and quantifying the levels of gene expression to gain insights into various biological processes, such as development, disease progression, response to treatment, or identification of biomarkers. By performing these experiments, for these 20 patients we have transcriptomic data available for 39202 genes.

5.1.2 Networks

Two kind of networks have been generated.

- **Co-expression network** : A co-expression network is based on the concept that genes with similar expression patterns across different conditions or samples are likely to be functionally related or involved in the same biological processes. Co-expression networks are constructed by measuring the expression levels of genes across multiple samples or experimental conditions and quantifying the degree of correlation or similarity between their expression profiles. Thus, a co-expression network contains the gene-gene links that are connected based on similar expression patterns along with the weight feature. The weight feature is the bi-weight mid-correlation value between the

two genes that are linked. The hosting team have used the program WGCNA (Weighted Correlation Network Analysis) [11] to infer the co-expression network on the available cohort. The co-expression network has been divided into modules, which are groups of genes that show similar expression patterns or have common functions. These modules represent sets of genes that are likely involved in related biological processes or pathways. The creation of modules can be based on gene expression, gene networks, functional characteristics, or a combination of these methods. In our case, we used an expression-based approach, where genes with similar expression profiles, whether positively or negatively correlated, were grouped together into modules. After the analysis, we identified 10 distinct modules within the patient cohort, and we decided to name each module after a color to make them easier to reference.

- **Interaction network** : An interaction network represents the physical or functional interactions between molecules and it provides insights into the complex relationships and communication among different components within a biological system. These interactions can involve proteins, genes, metabolites, or other bio-molecules. By using the set of genes belonging to any particular module, an interaction network dataset is created by passing the list of genes to the StringDB. StringDB is a biological database that focuses on protein-protein interactions (PPIs). This returns interaction links between the genes along with the confidence score that specifies the strength of the interaction between the two genes.

5.1.3 Node Labels

The genes present in the networks mentioned above includes additional features beyond their interactions such as gene name, mitocarta, interactome, labels etc. Gene name and labels are only fetched from this dataset. Moreover, the feature 'Mitocarta' provides information about the presence or absence of a gene in Mitocarta, a comprehensive database of mitochondrial genes. This information is essential because the presence of a gene in mitocarta indicates its potential involvement in mitochondrial functions and associated biological pathways. This knowledge can be valuable for studying mitochondrial-related diseases, investigating mitochondrial gene expression patterns, and exploring potential therapeutic targets related to mitochondrial dysfunction. In the research paper [12], a set of genes associated with mitochondrial diseases was identified and labeled as 1, indicating their potential role in causing the diseases, in our dataset. However, determining which genes are non-pathogenic was a more challenging task. To tackle this, we calculated the correlation between gene expression in patients with mitochondrial diseases and in healthy individuals. Genes with a correlation value greater than 0.216 were labeled as 0, indicating they are less likely to be pathogenic or associated with the diseases. Genes that couldn't be confidently labeled as either pathogenic or non-pathogenic were categorized as 'unlabeled'. It's important to note that the accuracy of these labels is uncertain and requires further investigation to validate their role in mitochondrial diseases.

5.1.4 Edge Weights

The transcriptomic data of the 20 patients of the patient cohort were used to add a weight to each edge between two nodes representing genes in the interaction network.

To calculate the edge weight, several approaches were explored as outlined below.

1. **No Edge Weights** : This approach does not assign any weights to the edges between nodes or genes. It implies that all edges are considered equal, without considering any specific relationship or strength between them.

2. **Correlation** : The correlation measure can be helpful in identifying the strength and direction of the relationship between two genes, indicating potential co-regulation or co-expression patterns.

   - **Canonical Correlation Analysis (CCA)** : CCA examines the linear relationship between two sets of variables, measuring correlation and maximizing correlation between their linear combinations. It is helpful in passing edge weights between genes by identifying shared patterns and relationships, facilitating the estimation of relevant edge weights.
   - **Partial Least Squares Regression (PLS regression)** : PLS regression finds linear combinations of variables to optimize covariance between independent and dependent variables. It aids in passing edge weights between genes by capturing relationships and maximizing shared information, enabling effective estimation of edge weights in high-dimensional scenarios.
   - **Pearson Correlation** : The Pearson correlation coefficient measures linear correlation between variables, quantifying the strength and direction of the relationship. It assists in passing edge weights between genes by assessing similarity and association between gene expression profiles, allowing estimation of edge weights that reflect co-expression patterns.

   It is important to note that all three methods may yield negative values, indicating negative correlations. Negative correlations still provide valuable information about gene interactions. In the experimentation, negative weights were allowed in one part, while in another part, the absolute values of the weights were considered. This approach allows for a comprehensive analysis of both positive and negative interactions between genes.

3. **Gini Index** : The Gini Index is a measure of inequality in a distribution. When used as an edge weight between genes, it helps capture the imbalance of connections within a gene network, highlighting central genes and potential regulatory relationships. This information is valuable for identifying influential genes and understanding network dynamics in gene networks.

4. **Entropy** : Entropy-based measures can provide insights into the diversity or concentration of connections between genes, helping identify hubs or modules within a gene network.

- **Shannon Entropy** : Shannon entropy measures the information content or uncertainty in a probability distribution. When used as edge weights for a gene network, it can be helpful in capturing the diversity or randomness of connections between genes.
- **Tsallis Entropy** : Tsallis entropy is a generalization of Shannon entropy that introduces a parameter to control the sensitivity to rare events. By considering the degree of disorder or diversity in the distribution of edge weights, Tsallis entropy can aid in capturing the heterogeneity of connections between genes in a network.
- **Renyi Entropy** : Renyi entropy is another generalization of Shannon entropy that introduces a parameter to adjust the emphasis on different parts of the distribution. When used as edge weights for a gene network, Renyi entropy can help identify modules or sub-networks within the larger network, based on the concentration or diversity of connections between genes.

5. **Mutual Information** : Mutual information measures the amount of information shared between two variables. In the context of edge weights for a gene network, it can capture the dependence or interaction between genes, aiding in the identification of functionally related genes or regulatory relationships within the network.

6. **Divergence** : Divergence can be useful in quantifying the dissimilarity or difference between gene expression patterns, aiding in the identification of gene clusters with distinct expression profiles or identifying genes that deviate from a common regulatory pattern.

    - **Jensen Shannon Divergence (JS Divergence)** : JS Divergence is a measure of similarity or dissimilarity between probability distributions. When passed as edge weights for a gene network, it can quantify the difference between the distributions of connections for different genes, helping identify clusters or groups of genes with distinct connection patterns.
    - **Kullback Leibler Divergence (KL Divergence)** : KL Divergence measures the difference between two probability distributions. When used as edge weights for a gene network, Kullback-Leibler Divergence can help assess the dissimilarity or discrepancy between the distributions of connections for different genes, providing insights into the uniqueness or distinctiveness of their connectivity patterns.

5.1.5 Final Network Configuration

Table 1 provides a comprehensive overview of the final network configurations for each module, including the varying confidence score thresholds. To mitigate the issue of dense edge connections, a confidence score threshold has been implemented to reduce the number of edges. The specific threshold value varies across modules, taking into consideration the complexity and characteristics of each network.
Moreover, responsible gene information is available only for two modules, named brown and blue modules, respectively. For brown module, 'VPS13D' and 'UFM1' are

identified as responsible genes, and 'C1QBP' as putative candidate. For blue module, 'ADCY5' is identified as responsible gene. The built model should be capable of accurately predicting the responsible genes as pathogenic with high probability.

5.2 Intellectual Disorders

5.2.1 Mouse-model Cohort

We collected multi-omics data from a cohort of mouse samples, including 2 transcriptomic, 3 translatomic, and 1 proteomic dataset. We have integrated these data in a multi-omics composed of 3986 genes for 50 samples. Analysis of this integrated dataset by using a novel model developed in the hosting team, yielded 1299 genes as impaired in the studied phenotype.

5.2.2 Network

We selected the top 453 genes resulted by our multi-omics analysis and the interaction network of the genes was built using StringDB, which provides interaction links between the given list of genes, along with confidence scores as explained in the section 5.1.2. To streamline the interaction network and reduce complexity, only links with a confidence score greater than 0.8 were retained. This filtering process helps focus on the most reliable and significant interactions within the network.

5.2.3 Node Labels

Genes are categorized as pathogenic (1), non-pathogenic (0), or unlabeled. Using the SFARI Gene Database, genes identified as pathogenic are labeled as 1. SFARI is an evolving database for the autism research community that is centered on genes implicated in autism susceptibility. The labeling of genes as non-pathogenic lacks a definitive approach; therefore, we took advantage of the multi-omics analysis carried out. Genes that are not prioritized by the aforementioned multi-omics analysis are labeled as non-pathogenic. This means that these genes did not exhibit significant associations or patterns indicating their direct involvement in the studied condition. The remaining genes are marked as 'unlabeled' and are predicted using the built GCN model.

5.2.4 Edge Weights

The multi-omics data of 50 mouse samples were used to add the edge weight between two nodes representing genes in the interaction network. Edge weights between nodes representing genes were added using the same approaches as mentioned in section 5.1.4, which include correlation, entropy, divergence, and mutual information. These methods were employed to calculate the edge weights, capturing the relationships and dependencies between genes in the network.

5.2.5 Final Network Configuration

Final network of intellectual disabilities consist of 2382 nodes and 15237 edges. Out of these nodes, 26 are labeled as 1 (pathogenic), 2047 as 0 (non-pathogenic) and 309 are unlabeled. Moreover, one responsible gene 'fmr1' is also unlabeled and the

model should be able to predict it as pathogenic with high probability. Additionally, a total of 32 genes have been identified as potential candidates, with 27 of them present in the final network configuration. Among these candidates, 8 genes have been labeled as pathogenic, indicating a strong association with the development of intellectual disabilities. Furthermore, 19 genes remain unlabeled, warranting further investigation to determine their potential pathogenic roles or contributions to the disorder.

6 Method

This section encompasses the general process employed on the omics or multi-omics dataset, as well as the construction of the final GCN model, regardless of the specific focus on the pathology. To evaluate the performance of the GCN model on two distinct pathologies, I have devised a general process and the GCN model for analyzing both the Mitochondrial Diseases dataset and the Intellectual Disabilities dataset.

6.1 Constructing the graph

The omics dataset has undergone a thorough cleaning and structuring process, where the genes are represented as row indices, and the column indices correspond to patients or mouse samples. The values within the dataset represent transcriptomic, proteomic, or translatomic data per sample, depending on the specific disease being considered. The data was first transformed with the QuantileTransformer() to change the features' distributions into a normal distribution since the distributions were skewed. Then the transformed features were scaled with the MinMaxScaler() to get the data in a range of $0 - 1$. This data was then added as a feature vector to the nodes of the graph that was constructed using an interaction network.
Based on the various approaches considered for calculating edge weights, 14 different graphs were constructed by passing respective weights to the edges between nodes. These 14 graphs are passed as input, one by one, to the model to analyze the model's performance.

6.2 Model Architecture

Figure 2 illustrates the implemented GCN model, which comprises four layers, including two Graph Convolution layers and two Graph-SAGE layers alternately. The Graph-SAGE layer utilizes the "pool" aggregation type, which performs graph pooling by aggregating information from neighboring nodes through pooling or downsampling their features [13]. The output layer, consisting of two neurons, classifies the nodes as either pathogenic or non-pathogenic. The size of hidden neurons in the rest of the convolutional layers is set uniformly across the network, considering the complexity of the specific module. Batch normalization is applied in the convolutional layers of the Graph-SAGE layer to enhance training performance. Relu activation function is used between the layers to introduce non-linearity. The Adam optimizer is employed with a learning rate of 0.001, and the model is trained for 1000 epochs. The GCN takes the graph containing the omics information in the nodes and the weight information in the edges as input and then learns to predict

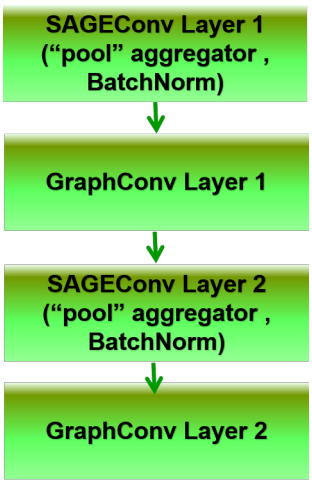


Figure 2: GraphConv + Graph-SAGE model

if unlabeled nodes are pathogenic or not.
During training, the model is exposed to the entire dataset to capture a comprehensive representation of the network. This inclusive approach enables the model to learn from diverse patterns and interactions within the network, facilitating predictions and identification of potential pathogenic nodes based on learned feature vectors and edge weights. To address class imbalance, weights are incorporated into the loss function of the model, ensuring balanced training and gradient computation during back-propagation. Furthermore, hyper-parameter tuning was performed to enhance the overall model performance.

7 Experiments and Observations

The performance of the GCN model was assessed by examining the loss on the validation set at the epoch with the minimum loss value. This is because for different edge weight approaches, the accuracy and F1 score are in the range of 99 - 100% on validation set. Hence, validation loss factor is considered for choosing the best edge weight approach for the specified module. Additionally, for the module associated with a known responsible gene, its performance was evaluated by examining the prediction of the responsible gene with a high probability.
Furthermore, to classify genes as pathogenic, a probability threshold is set based on their predicted pathogenicity. Different threshold values are established for different diseases since they exhibit distinct pathologies. By applying a probability threshold, genes are categorized as pathogenic if their predicted probability of being pathogenic surpasses the threshold. This approach allows for tailored classification criteria that account for the unique characteristics and severity of each disease. To calculate the probability, the log softmax function was applied to the output generated by the GCN model. The log softmax function transforms the model's output into a probability distribution by applying the logarithm to the softmax function's result. This transformation ensures that the predicted probabilities are in the range of 0 to 1 and are suitable for classification purposes.

7.1 Mitochondrial diseases

As mentioned in the section 5.1.2, through the analysis of co-expression data, we identified 10 distinct modules within the patient cohort, each denoted by a specific color as listed in table 1. Initially, we chose to focus on one specific module to thoroughly investigate its characteristics and analyze the model's performance on interaction and co-expression networks. The primary objective was to conclude whether there are any similarities or consistent patterns in the model's predictions that might indicate a potential connection between the interaction and co-expression networks. Therefore, first we focused on brown module that contains total of 1487 genes in which 47 pathogenic, 22 non-pathogenic and 1418 unlabeled genes are present. This data is considered for interaction network with 6364 edge connections. Similarly for the co-expression network, a threshold of 0.2 has been considered for weight feature. The resulting dataset has a total of 112823 edge connections and 1293 genes in which 37 are labeled as pathogenic, 17 are labeled as non-pathogenic and 1239 are unlabeled. Model has been trained by considering all the known genes so that the prediction will be more accurate. In addition, two responsible genes and one putative candidate are unlabeled and the model should be able to predict them as pathogenic with high probability.
However, when applied to co-expression networks, the GCN model exhibits poor performance on predicting genes as pathogenic due to the dense edge connections present in such networks. This is in contrast to interaction networks where the model performs well, as indicated in table 2. As a result, for the remaining modules, the model was exclusively trained using the interaction network data. This decision was made to focus on the network type where the GCN model demonstrated better performance and yielded more reliable results.

Table 3 presents the results for the 10 different modules associated with mitochondrial diseases. To identify pathogenic genes with a high level of confidence, genes that are predicted as pathogenic with a probability exceeding the threshold of 0.95 are considered pathogenic, while the rest are considered non-pathogenic. By setting a higher probability threshold, the classification becomes more stringent, resulting in a smaller set of confidently predicted pathogenic genes. This approach helps to filter out genes with higher uncertainty. Furthermore, the best edge weight method varies across modules due to differences in the complexity of their interaction networks. To provide a general overview, the confidence interval for the overall percentage of predictions is given, considering various edge weight methods. This interval gives an indication of the reliability and consistency of the predictions across different modules. Moreover, the table provides the percentage of pathogenic genes present in mitocarta out of all pathogenic genes, as well as the percentage of mitocarta genes out of all genes. These percentages offer insights into the coverage and representation of pathogenic genes within the mitocarta database. Also, analyzing mitocarta information helps in identifying new potential candidates. In this context, "pathogenic genes" refers to the combined set of known pathogenic genes and predicted pathogenic genes that exceed the probability threshold. With this information, further analysis and interpretation of the results can be conducted.
Additionally, known responsible genes were present only in the blue and brown modules, as indicated in Table 4. Notably, the probability for the responsible gene 'ADCY5' in the blue module is 0.92, which is lower than the probability threshold.

Upon further investigation, it was revealed that the mitochondrial disease is a secondary condition for this patient, and it is not the primary cause of his pathology. This finding suggests that while 'ADCY5' was identified as a responsible gene, its strong association with mitochondrial diseases may not be apparent.

Additionally, Figure 3 illustrates the final networks obtained using the best edge weight method and parameters for each module.

7.2 Intellectual Disorders

In order to assess the GCN model's performance on a different pathology, we used the intellectual disorders dataset, which contains multi-omics data compared to the previous dataset with single omics type. This provided a unique opportunity to showcase the model's adaptability to diverse diseases and its ability to effectively handle multi-omics information. Further, to identify pathogenic genes in the intellectual disorders dataset, we obtained information from the Autism database due to the substantial co-occurrence between intellectual disorders and autism. Notably, approximately 35% of patients with intellectual disabilities also exhibit autism spectrum disorders. This observation prompted our interest in exploring further putative genes that may be associated with both phenotypes, serving as potential bridge between these two diseases. Moreover, the probability threshold set for selecting pathogenic genes in the intellectual disorders dataset was 0.9. Out of the 309 unlabeled genes, 43.04% were predicted as pathogenic with a probability greater than the threshold. Furthermore, the evaluation of different edge weight methods also considered the number of correctly predicted unlabeled candidate genes. Out of the 19 candidate genes, the JS divergence edge weight method correctly predicted 10 of them. This result highlights the effectiveness of the JS divergence method for this particular dataset. It is worth noting that intellectual disabilities are caused by the absence of the 'fmr1' gene due to a mutation in its sequence. Biologists perform experiments using mouse models to study this phenotype, comparing mice with and without this gene, in order to understand the implications of its absence on living organisms. The model's prediction that the 'fmr1' gene is pathogenic with a probability of 1.0 further strengthens the link between 'fmr1' and intellectual disorders, validating the effectiveness of the GCN approach to accurately identify pathogenic genes and highlights its potential as a reliable tool for disease-causing molecules identification. Also, this successful validation strengthens the credibility of the multi-omics integration strategy employed. Hence, JS Divergence is considered to be the best edge weight method.

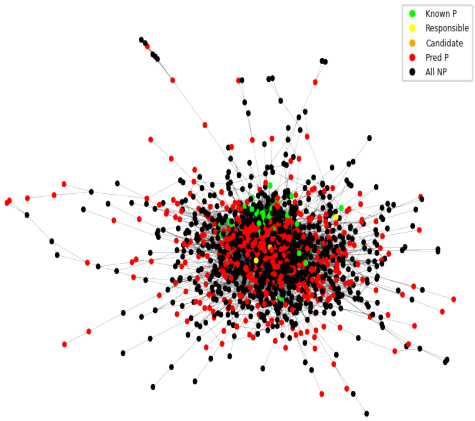
An outline of the results for different edge weight methods can be found in Table 5. This result has been sorted in descending order of number of candidate genes predicted as pathogenic. Moreover, Figure 4 depicts the final networks obtained using the GCN model for the intellectual disorders dataset.

Module	P	NP	Unlabeled genes	Total genes	Edge Score Threshold	Edges
Black	15	28	209	252	>= 0.5	406
Blue	43	10	1603	1656	>= 0.75	10893
Brown	47	22	1418	1487	>= 0.5	6364
Cyan	1	9	38	48	-	43
Green	10	20	217	247	-	436
Green Yellow	12	3	88	103	-	209
Light Cyan	2	4	122	128	-	183
Midnight Blue	3	55	161	219	-	319
Salmon	9	3	55	67	-	102
Yellow	7	8	283	298	>= 0.5	482

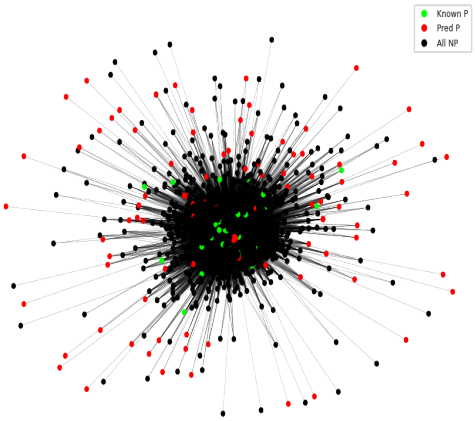
Table 1: Mitochondrial Disease Modules Overview ; P and NP columns represent number of pathogenic and non-pathogenic genes in each module ; Edge Score Threshold corresponds to confidence score threshold for edges

	Networks	
Results	**Interaction**	**Co-Expression**
Total Test Set	1418	1239
Predicted Pathogenic	461	91
Predicted Non-Pathogenic	957	1148
Probability obtained for two responsible and one putative genes		
UFM1	0.96 (P)	0.56 (NP)
VPS13D	0.99 (P)	0.6 (P)
C1QBP	0.99 (P)	0.65 (P)

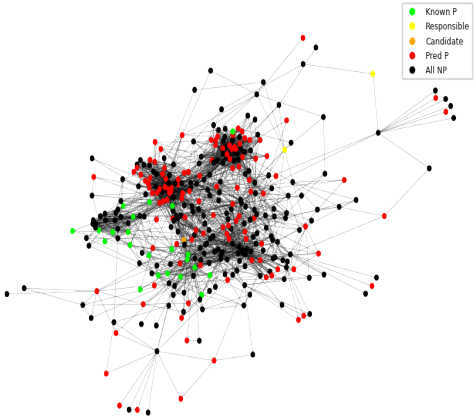
Table 2: Results obtained for Interaction network with TSallis entropy edge weight method and Co-Expression network. P represents gene is predicted as Pathogenic with probability > 0.95 and NP represents rest of the genes as Non-Pathogenic



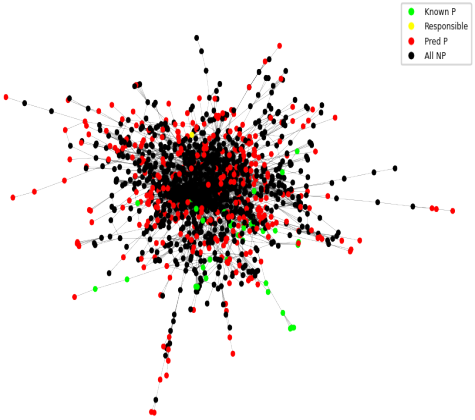
(a) Brown (I)



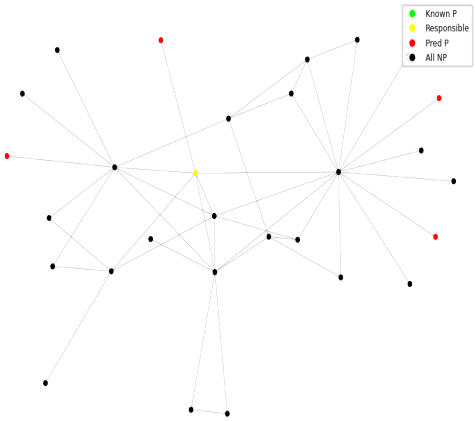
(b) Brown (C)



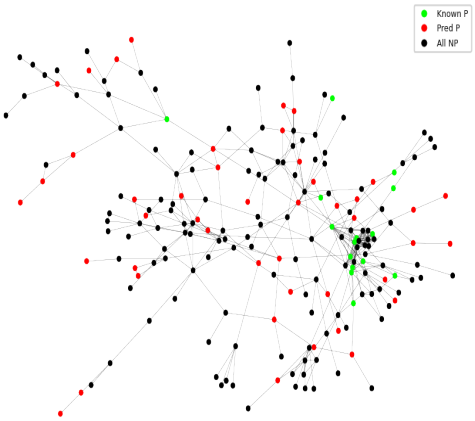
(c) Brown (R)



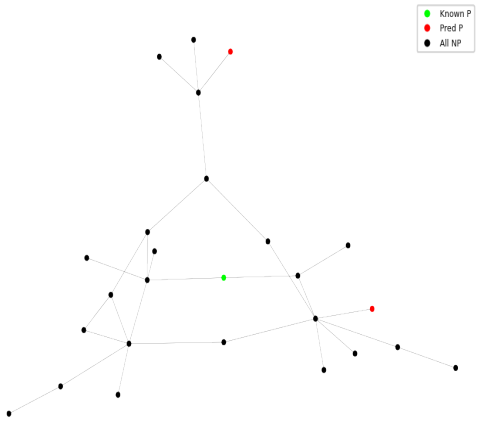
(d) Blue (I)



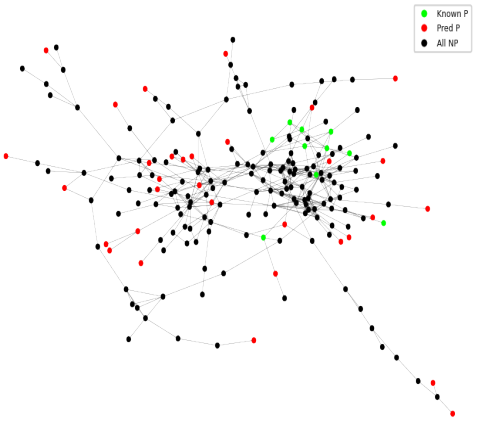
(e) Blue (R)



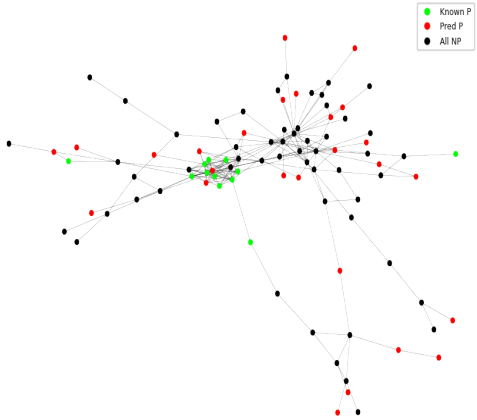
(f) Black (I)



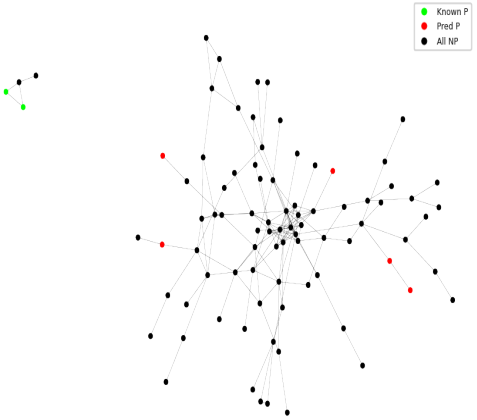
(g) Cyan (I)



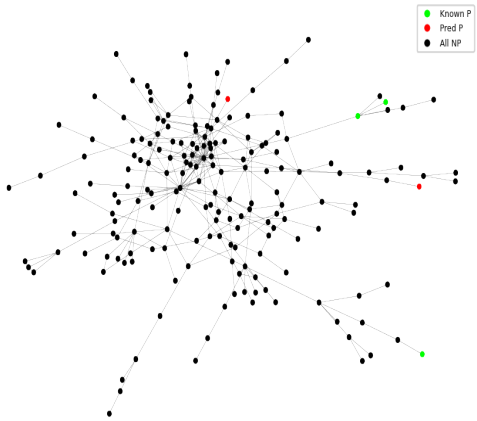
(h) Green (I)



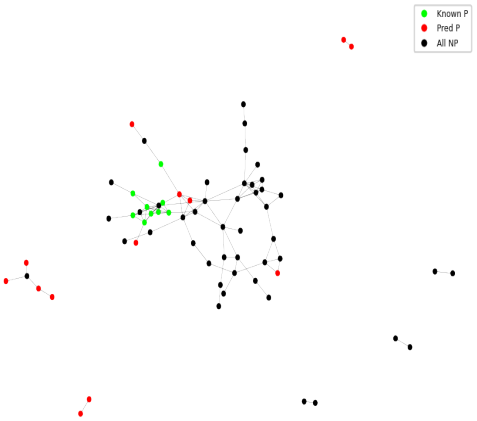
(i) GreenYellow (I)



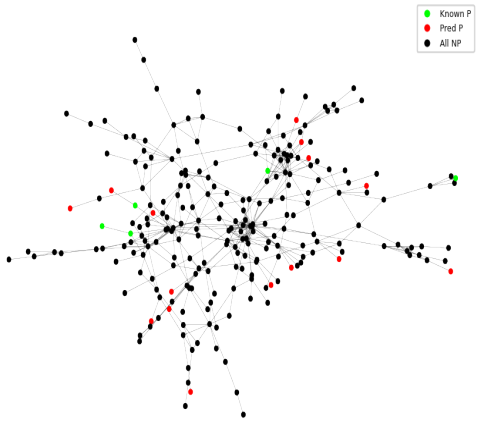
(j) LightCyan (I)



(k) MidnightBlue (I)

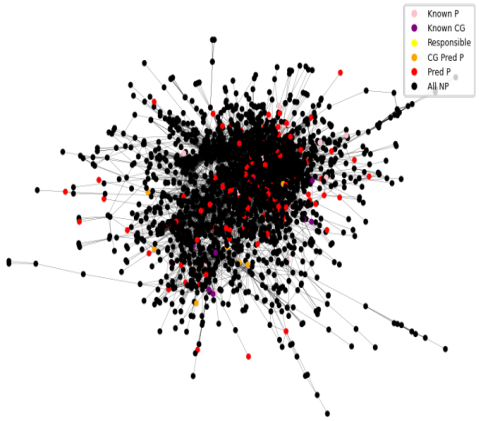


(l) Salmon (I)

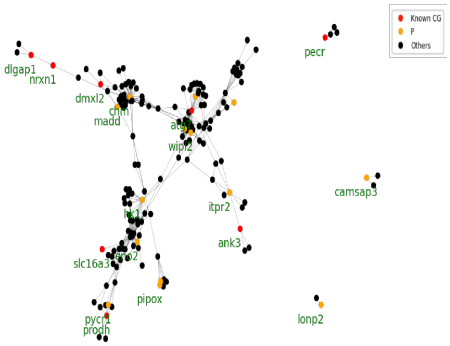


(m) Yellow (I)

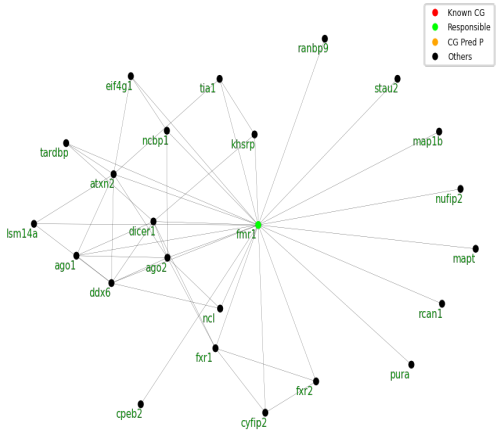
Figure 3: Gene Networks obtained for different modules using GCN Model. (I) represents interaction network. (C) represents co-expression network. (R) represents interaction network of responsible genes with hop-count 2 and this network is fetched from the (I) interaction network of respective model.



(a) Autism (I)



(b) Autism (CI)



(c) Autism (R)

Figure 4: Gene Networks obtained for Intellectual Disorder. (I) represents interaction network. (CI) represents candidate genes interaction network with hop-count 1. (R) represents interaction network of responsible genes with hop-count 1. (CI) and (R) networks are fetched from the (I) interaction network of respective model.

Module	Predicted Patho genes / Total Unlabeled	Edge Weight Method	Predicted Patho genes in %	CI of Edge Weight Methods	% of pathogenic genes in mito-carta	% of mi-tocarta genes in the module
Black	58 / 209	JS Diver-gence	27.75	0 - 45%	28.76	16.67
Blue	399 / 1603	CCA	24.89	19 - 73%	15.38	8.21
Brown	461 / 1418	Tsallis Entropy	32.51	2 - 88%	12.79	6.14
Cyan	2 / 38	JS Diver-gence	5.26	0 - 5%	66.67	14.58
Green	41 / 217	Mutual Infor-mation	18.89	10 - 45%	17.65	7.29
Green Yellow	31 / 88	JS Diver-gence	35.22	6 - 60%	41.86	23.30
Light Cyan	7 / 122	Pearson Correla-tion	5.73	3 - 30%	11.11	3.91
Midnight Blue	2 / 161	Pearson Correla-tion	1.2	1 - 19%	60	6.85
Salmon	13 / 55	No Weights	23.64	20 - 71%	45.45	23.88
Yellow	23 / 283	KL Diver-gence	8.12	3 - 18%	30	5.03

Table 3: Results for Mitochondrial Diseases

Module	Gene $->$ Prediction $->$ Probability
Blue	ADCY5 (responsible) $->$ 1 $->$ 0.92
Brown	VPS13D (responsible) $->$ 1 $->$ 0.99
	UFM1 (responsible) $->$ 1 $->$ 0.96
	C1QBP (putative) $->$ 1 $->$ 0.99

Table 4: Prediction of Responsible and Putative genes for Mitochondrial disease modules ; 1 represents pathogenic

Edge Weight Method	Train Loss	Train Accuracy	Train F1	Validation Loss	Validation Accuracy	Validation F1	Prediction	Candidate Genes	fmr1
Pearson Neg	0.0246	0.99	0.99	0.0013	1.00	1.00	122	10	NP (0.92)
Mutual Info	0.0188	0.99	0.99	0.0049	1.00	1.00	112	10	NP (0.66)
JS	0.0490	0.99	0.99	0.0876	1.00	1.00	133	10	P (1.00)
Tsallis	0.0418	0.99	0.99	0.0781	0.98	0.99	107	8	P (1.00)
No weight	0.0371	0.99	0.99	0.0046	0.99	0.99	102	7	NP (0.94)
Pearson	0.0233	0.99	0.99	0.0026	1.00	1.00	115	7	P (1.00)
CCA Neg	0.0429	1.00	1.00	0.0827	1.00	1.00	105	6	P (0.99)
PLS Neg	0.0429	1.00	1.00	0.0821	1.00	1.00	110	6	P (1.00)
PLS	0.0430	1.00	1.00	0.0828	1.00	1.00	77	5	NP (0.92)
Shannon	0.0263	0.99	0.99	0.0313	1.00	1.00	136	5	P (1.00)
Gini Index	0.2264	0.95	0.96	0.0199	0.99	0.99	64	4	P (0.96)
KL	0.0753	0.99	0.99	0.0619	0.99	0.99	62	4	P (0.98)
CCA	0.0465	1.00	1.00	0.0873	1.00	1.00	77	3	P (1.00)
Renyi	0.0947	0.97	0.97	0.0574	0.99	0.99	0	0	NP (0.60)

Table 5: Performance of GCN model for different edge weight methods in Intellectual Disorders ; Neg represents presence of negative edge weight ; NP represents gene predicted as non-pathogenic ; P represents gene predicted as pathogenic ; In column 'fmr1', probability is mentioned

8 Conclusion and Future works

By employing the proposed GCN model, I have demonstrated its potential to effectively classify genes in the context of mitochondrial diseases and intellectual disabilities. These findings highlight the significance of leveraging advanced computational techniques, such as GCNs, to address the challenges associated with rare diseases and multi-omics data analysis. Our research contributes to bridging the gap in the application of GCNs in rare diseases and underscores the importance of further exploration in this field. Further implementation details can be found on this Github Page.

Despite its successes, the GCN model does face certain limitations. Firstly, it relies solely on interactomic data, which restricts the breadth of information available for analysis. Additionally, in the context of rare diseases, the availability of prior knowledge is limited, which can hinder accurate predictions. Furthermore, the challenge lies in establishing clear criteria for labeling genes as non-pathogenic due to the intricacy of these diseases. It is difficult to determine which gene category would definitively not be involved in the disease given the complex nature of these conditions. One more significant challenge is the absence of a universally applicable method to calculate edge weights that accommodates the diverse range of network structures encountered in different biological systems.

To address these limitations, future implementations could involve integrating additional omics data sources, such as multi-cohorts and multi-organisms data, to enhance the comprehensiveness of the analysis. Developing standardized criteria or computational methods for labeling genes as non-pathogenic would facilitate more reliable predictions. Further, it is important to focus on the development of adaptable edge weight calculation strategies that can effectively capture the nuances of various network topologies. Also, future work should consider enabling access to the features in the feature vectors and their importance, fostering a transparent model that provides insights into how the model arrives at its predictions and promotes understanding of biological processes. By addressing all these challenges, we can enhance the reliability of the proposed GCN model, enabling more precise prioritization of pathogenic genes and facilitating their broader application in rare diseases and other complex biological contexts.

References

[1] T. Wang *et al.*, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," en, *Nature Communications*, vol. 12, no. 1, p. 3445, Jun. 2021, Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/s41467-021-23774-w. [Online]. Available: https://www.nature.com/articles/s41467-021-23774-w (visited on 05/18/2023).

[2] M. Ghorbani, A. Kazi, M. S. Baghshah, H. R. Rabiee, and N. Navab, *RA-GCN: Graph Convolutional Network for Disease Prediction Problems with Imbalanced Data*, arXiv:2103.00221 [cs], Nov. 2021. DOI: 10.1016/j.media.2021.102272. [Online]. Available: http://arxiv.org/abs/2103.00221 (visited on 05/18/2023).

[3] A. Kazi *et al.*, "InceptionGCN: Receptive Field Aware Graph Convolutional Network for Disease Prediction," en, in *Information Processing in Medical Imaging*, A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 73–85. DOI: 10.1007/978-3-030-20351-1_6.

[4] R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico, "Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms," en, *Nature Machine Intelligence*, vol. 3, no. 6, pp. 513–526, Jun. 2021, Number: 6 Publisher: Nature Publishing Group. DOI: 10.1038/s42256-021-00325-y. [Online]. Available: https://www.nature.com/articles/s42256-021-00325-y (visited on 05/18/2023).

[5] W. L. Hamilton, R. Ying, and J. Leskovec, *Inductive Representation Learning on Large Graphs*, arXiv:1706.02216 [cs, stat], Sep. 2018. [Online]. Available: http://arxiv.org/abs/1706.02216 (visited on 05/18/2023).

[6] L. Yang, H. Liu, D. Sun, K. Liu, and C. L. P. Chen, *SAGE-GCN: Graph Convolutional Network Based on Self-adaptive Stable Gates for Link Prediction in Dynamic Complex Networks*, en, Apr. 2023. DOI: 10.36227/techrxiv.22658380.v1. [Online]. Available: https://www.techrxiv.org/articles/preprint/SAGE-GCN_Graph_Convolutional_Network_Based_on_Self-adaptive_Stable_Gates_for_Link_Prediction_in_Dynamic_Complex_Networks/22658380/1 (visited on 05/18/2023).

[7] P. Narad and S. V. Kirthanashri, "Introduction to Omics," en, in *Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data*, P. Arivaradarajan and G. Misra, Eds., Singapore: Springer, 2018, pp. 1–10. DOI: 10.1007/978-981-13-2925-8_1. [Online]. Available: https://doi.org/10.1007/978-981-13-2925-8_1 (visited on 07/17/2023).

[8] J. Labory, M. Fierville, S. Ait-El-Mkadem, S. Bannwarth, V. Paquis-Flucklinger, and S. Bottini, "Multi-Omics Approaches to Improve Mitochondrial Disease Diagnosis: Challenges, Advances, and Perspectives," *Frontiers in Molecular Biosciences*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmolb.2020.590842 (visited on 05/26/2023).

[9] K. McKenzie, D. Metcalfe, and A. L. Murray, "Screening for intellectual disability in autistic people: A brief report," en, *Research in Autism Spectrum Disorders*, vol. 100, p. 102 076, Feb. 2023. DOI: 10.1016/j.rasd.2022.102076. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1750946722001635 (visited on 07/06/2023).

[10] R. Higdon *et al.*, "The Promise of Multi-Omics and Clinical Data Integration to Identify and Target Personalized Healthcare Approaches in Autism Spectrum Disorders," *OMICS : a Journal of Integrative Biology*, vol. 19, no. 4, pp. 197–208, Apr. 2015. DOI: 10.1089/omi.2015.0020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4389910/ (visited on 07/06/2023).

[11] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, Dec. 2008. DOI: 10.1186/1471-2105-9-559. [Online]. Available: https://doi.org/10.1186/1471-2105-9-559 (visited on 05/30/2023).

[12] L. D. Schlieben and H. Prokisch, "The Dimensions of Primary Mitochondrial Disorders," *Frontiers in Cell and Developmental Biology*, vol. 8, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcell.2020.600079 (visited on 05/31/2023).

[13] M. Labonne, *GraphSAGE: Scaling up Graph Neural Networks*, en, Aug. 2022. [Online]. Available: https://towardsdatascience.com/introduction-to-graphsage-in-python-a9e7f9ecf9d7 (visited on 05/18/2023).