miRGD: miRNA-Gene-Disease Association Prediction Based on GCN

1. database (NAR)
2. miRNA similarity improvement (consider 5p/3p): david bartal (seed sites decide the target gene; only 2nd-7th bp is important；targetScan; match between seed and target is strict) can also calculate similarity based on their target genes
3. input: RNA-seq(patient/control) -> output: g-d or d-d association(DESeq2->DEgene)

rex’s comments：

1. add more data/ type of data/
2. pre-training - mask objective
3. interpretability approach - attention, <https://arxiv.org/pdf/2201.12987.pdf>

Jan 27

Mark:

1. filling triangle that is missing an edge, find mirna-gene connection
2. text mining? - find missing edge
3. mirna-gene based on sequence

## Introduction

### 1.1 Define the problem

A microRNA (abbreviated miRNA) is a small single-stranded non-coding RNA molecule (containing about 22 nucleotides) found in all eukaryotic cells, that functions in RNA silencing and post-transcriptional regulation of gene expression. [1] One-third of the human genome is estimated to be regulated by miRNAs. It is estimated that about 50% of miRNAs expressed in the genome are transcribed from non-protein-coding genes and the remaining miRNAs are coded in the introns of coding genes.[2]

miRNA expression patterns are tissue-specific [3] and have been shown to play an important role in a wide range of developmental processes, including metabolism, cell proliferation, apoptosis, development time, and neuronal cell fate [4-8]. Other regulatory effects include neuronal gene expression [9], brain morphogenesis [10], muscle differentiation [10], and stem cell division [11]. The relationship between miRNAs and human diseases has been widely investigated. It has been reported that microRNAs’ deficiencies or excesses have been linked to a number of other clinically important diseases, including cancer, cardiovascular disease, neurodevelopmental disease, autoimmune disease, etc.[2][12]

Work has been done to associate miRNA with their target genes, and associate miRNA with diseases. However, there is still a missing piece to put the picture together. Here, we are trying to use a graph structure-based deep learning framework to model the regulatory network of miRNAs and their target genes as well as the relational graph of human disease and genes to reveal the whole landscape of miRNA, target genes, and associated diseases.

### 1.2 Summarize the method and results

We created a heterogeneous graph with 3 types of nodes (miRNA, gene, disease) and 6 types of relationships (miRNA-miRNA similarity, gene-gene similarity, disease-disease similarity, miRNA-gene association, gene-disease association, and miRNA-disease association). Based on this heterogeneous graph, we built a 2 layer Graph Convolution Neural Network to learn the embedding of each node. The training dataset included the first 5 types of links as well as the m-d links with labels. Our goal is to predict the links between miRNA and diseases.

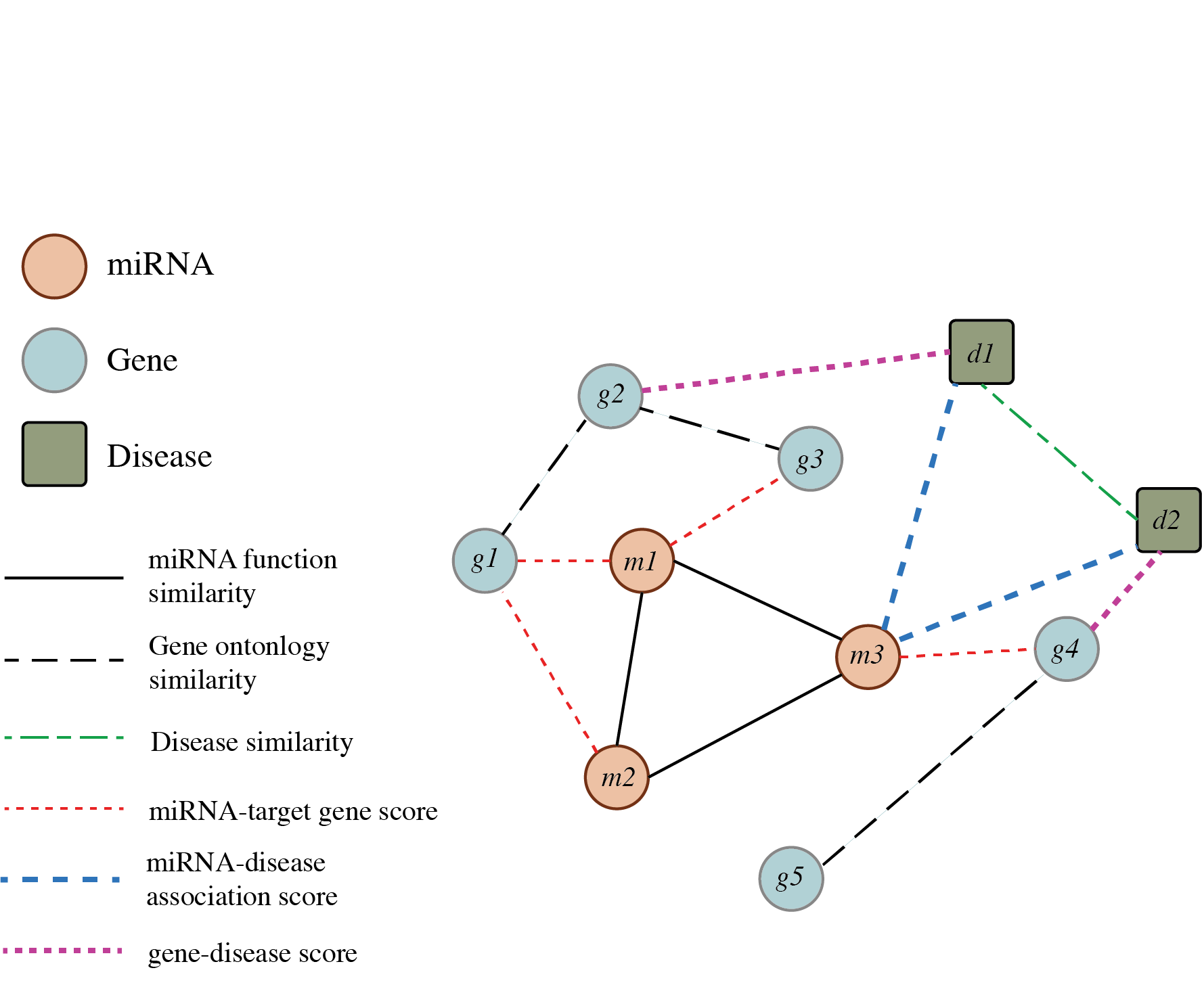


Figure 1. Schematic diagram of the miRNA-gene-disease heterogeneous graph

### 1.3 Conclusion

We created a heterogeneous graph with 3 node types: miRNA, gene, disease, and 6 edge types to represent the similarity or association between 2 nodes. We encoded the node types by one-hot encoding to distinguish the different node types and built 2 GCN layers to learn node embeddings. Then use the output of the last GCN layer to predict whether there is an edge between two nodes. The accuracy of the 3 tests is 100% +/- 0. We also manually collected disease leukemia and acquired immunodeficiency syndrome-related miRNAs as further validation. In summary, our method works pretty well on this miRNA-gene-disease prediction task.

## Related Work

The association of miRNA and diseases has been studied thoroughly at the biological level. miR-155 and let-7a have been corroborated to be associated with lung cancer [14]. miR-29b-1, miR-29a, and miR-9 showed an association with Alzheimer’s disease [15]. Previous work miRTargetLink 2.0 [16] presented an interactive miRNA target gene and target pathway networks but lacked the disease information. NIMCGCN [17] proposed a graph convolutional network for predicting miRNA-disease association. Ji et.al. [18] reported a miRNA-disease association from a heterogeneous information network with the GraRep embedding model. However, both of these 2 works skip the target gene which leaves out information about a complicated network between miRNA and diseases which can provide more insights into the regulation related to miRNA. The global investigation to reveal what are the target genes of the disease-associated miRNAs and how they regulate disease-related pathways is still blank. By predicting miRNA-disease relations by combining the miRNA-gene regulatory network and gene-disease knowledge graph, we can compensate for the lack of gene regulation level by tracing the existence of indirect pathways linked through the gene nodes for the predicted link between miRNA and disease.

Moreover, two mature miRNA species may be generated from the 5’ and 3’ arms of a pre-miRNA precursor. In most cases, only one species remains while the complementary species is degraded. The miRNA-3p and miRNA-5p are mutually complementary sequences, thus their target genes could be very different. So from a biological perspective, it is meaningful to distinguish miRNA-3p or miRNA-5p when you associate a target gene with a miRNA and the same for disease and miRNA. However, the current computational predictions of miRNA-diseases relation neglected this point. Our work bridged this gap by predicting miRNA-disease relations with the 3p or 5p subtype information.

## Method

In this section, we talk about two different architectures we used in this project.

### 3.1 Simple GCN

The graph convolutional network (GCN) is a type of deep learning model that is used to analyze graph structure data. We applied the simple GCN on our built miRNA-gene-disease graph. Each node represents the entity in the data, which are different miRNA, genes, and diseases type with initial features based on their node type. The node type is one-hot encoded. We only add edges to the graph if two nodes have similarity or association scores with high confidence or are experimentally verified. The input graph is represented as a matrix of node features and a set of edges. The convolutional layers are used to extract features from the input graph and combine them in a hierarchical manner using a graph convolution operation. The output layer is the hidden layer of the last graph convolutional layer which we considered as the learned representation of each node. Based on the learned representation, we use linear classifiers and similarity scores to predict the relationship between nodes.

### 3.2 Heterogeneous Graph

Besides the simple GCN model, we design a heterogeneous graph to learn the node representation. A heterogeneous graph is a type of graph that includes nodes and edges with different labels or attributes. Similar to simple GCN, we have node type (miRNA, gene, disease) one-hot encoded as the initial node features. The differences are on the edges and the convolutional layers. In the heterogeneous graph, we have 6 different types of edges. This is to say, we have 6 different relationships between nodes. The relationship is solely based on the type of the two nodes in a node pair. Given three node types, we have relationships as follows: miRNA-miRNA, gene-gene, disease-disease, miRNA-gene, miRNA-disease, and disease-gene. The input graph is also a matrix of node features, and six different edge sets for each of the relationships. The convolutional layers take the input graph and pass the input graph to six different graph convolutional operators. For each of the graph convolutional layers, the model gathers the GCN output and uses a linear transformation to generate a combined output. This output will be fed to the next convolutional layer. In the end, we use linear classifiers and similarity scores to predict the relationship between nodes.

### 3.3 miRNA Nomenclature

Existing computational methods that build computational graphs and learning relationships between miRNAs, genes, and diseases only use a broad category of miRNA. However, each miRNA can be differentiated into two categories: miRNA-3p and miRNA-5p. Though they share the same prefix, they function differently in real situations and should not be considered the same. Work and publications in the biology and genetics field that study miRNAs, always specifically indicate if it's a miRNA-3p or miRNA-5p, while in the computational field, most of the work regardless of -3p and -5p information. There is no existing study in computational biology that summarizes the miRNA-disease relationship that includes 3p and 5p information, but there do exist studies about miRNA-3p and miRNA-5p’s target gene. Therefore, by including genes in our graph, we want to infer the relationship between miRNA-disease.

## Experiments

### 4.1 Data preprocess

We first collected the 6 relations dataset from the following source:

1. miRNA-miRNA functional similarity: <http://www.csbio.sjtu.edu.cn/bioinf/MiRGOFS/>
2. disease-disease similarity: <https://github.com/ljatynu/NIMCGCN/tree/master/datasets>
3. gene-disease association validated by experimental data: <https://diseases.jensenlab.org/Downloads>
4. miRNA-gene association validated by reporter assay or Western blot: <https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/download.php>
5. miRNA-disease association: <https://github.com/ljatynu/NIMCGCN/tree/master/datasets>
6. gene-gene semantic similarity between gene ontology terms

For 1-5, we downloaded the off-shelf data. For 6, we first generated the human gene Ensemble ID list and the corresponding Gene Ontology(GO) terms by R package biomaRt. Next, converted the Ensemble ID in the gene-GO table to gene names by Python package mygene. In this way, we derived the gene-GO table with gene names as the external keys.

Then we matched the 6 different datasets with the shared external keys. The external keys could be miRNA ID, gene names, and disease names (Fig. 1). Given that in Datasets 1 and 4, the miRNA are named in the same format, including 3p/5p, which indicated it generated from 5’ end or 3’ end. To uniform the use of miRNA ID with 3p/5p for all the datasets that included miRNA, we matched the miRNA IDs in Dataset 5 with the general nomenclature (without 3p/5p) to the detailed nomenclature (with 3p/5p) by the following rules: if the miRNA IDs in Dataset 5 already named by detailed nomenclature, just kept the original way. If the miRNA IDs are named by general nomenclature, for example, hsa-miR-485b, we matched it to both hsa-miR-485b-3p and hsa-miR-485-5b, if both of them exist. In most cases, hsa-miR-485b-3p and hsa-miR-485-5b won’t exist at the same time, we only need to match hsa-miR-485b with the existing one. We kept the shared miRNAs among all the 3 datasets and the unique miRNAs in dataset 1 since we would like to make some novel predictions. There are 1691 miRNAs reserved in total.

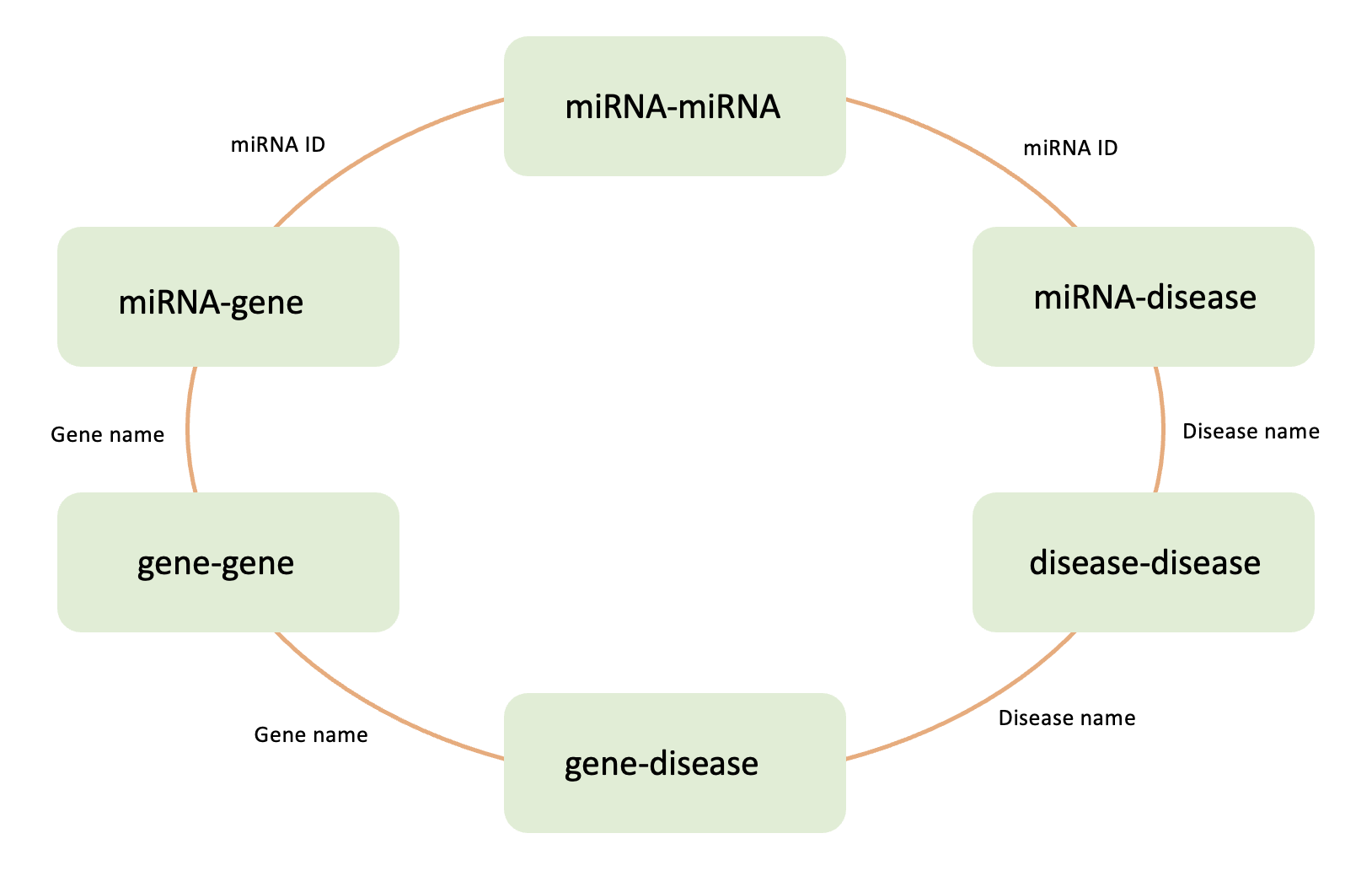


Figure 2. Overview of data preprocess

To match the disease names, we first applied a string semantic similarity scoring using gestalt pattern matching (Racliff, 1980) implemented in difflib to allocate each disease name pair a score between Datasets 2 and 3 (2 and 5 have the exactly same disease names). For the disease name pairs with score 1, we directly kept them. For the disease name pairs with scores lower than 1, we manually determined if they are the same disease. 113 diseases were reserved after filtering.

Given that there are over 26,000 genes in humans, to reduce the calculation burden of gene-gene Ontology semantic similarity, we arranged it at the last step. After merging the data across datasets 1-5 using the shared external keys, there were only 1299 shared genes left, so the computation resource and time would be significantly reduced. The ‘Calculate pairwise functional similarities between a list of genes’ function of GOGO [19] calculated 3 dimensions of similarities of each pair of genes, Biological Process Ontology (BPO), Cellular Component Ontology (CCO), and Molecular Function Ontology (MFO). In order to keep it simple, we calculated the average of those 3 similarity scores as the final similarity between gene pairs.

In summary, we end up with 1109 different miRNAs with precursor information, 113 different diseases, and 9311 different genes.

### 4.2 Data statistics

We found that the different datasets show different patterns. Gene-gene Ontology semantic similarity should exhibit gene clustering since the GOGO also generated another output for the gene clustering. However, we could hardly tell a clear clustering structure from Fig. 3a, which indicates that taking the average was probably not an ideal way to combine the 3-dimensional similarities. The clustering structure presented in the miRNA-miRNA similarity heatmap (Fig. 3b) suggested a good agreement with the fact that miRNAs are clustered by miRNA families.

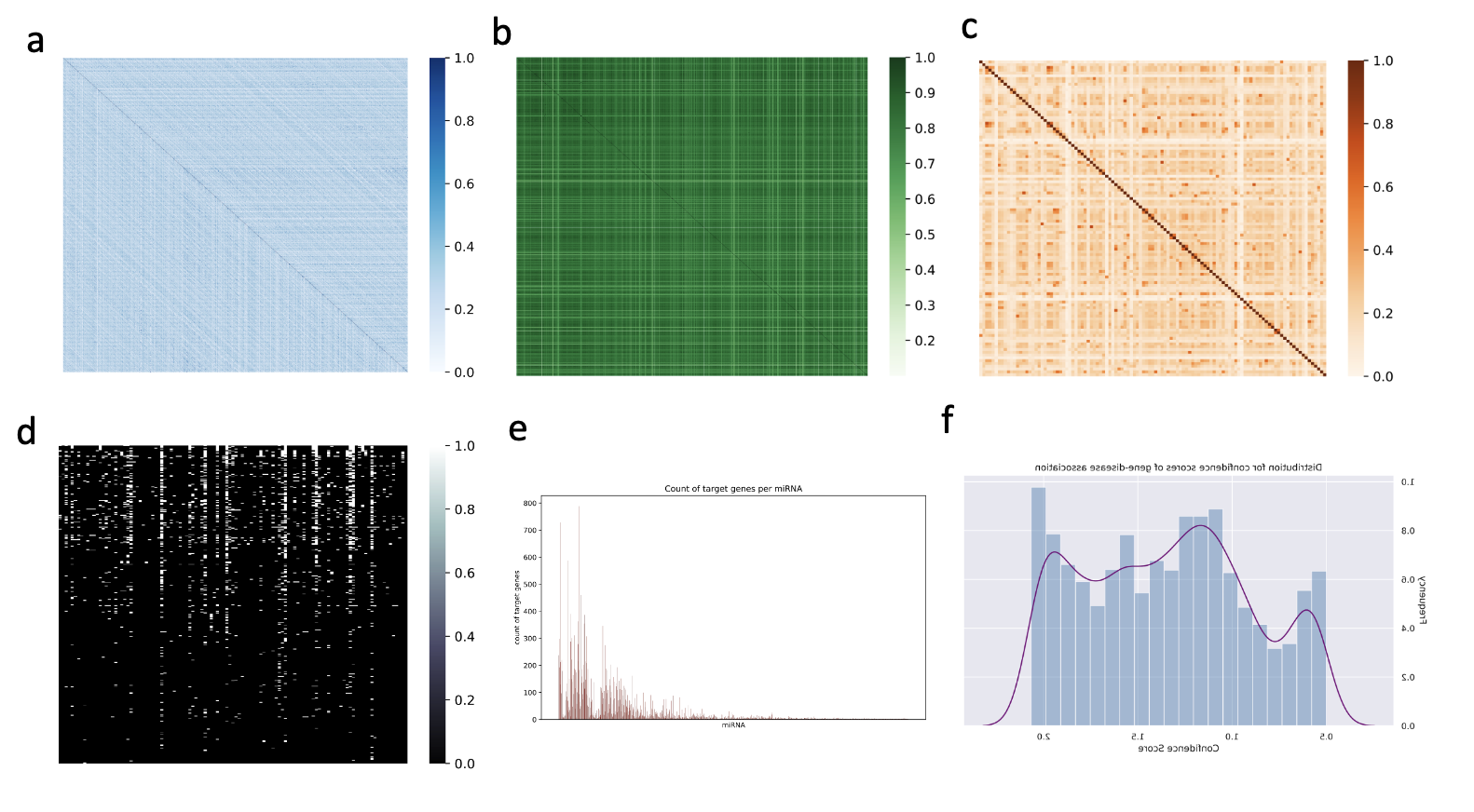


Figure 3. Dataset statistics

3a. Heatmap of gene-gene Ontology semantic similarity

3b. Heatmap of miRNA-miRNA similarity

3c. Heatmap of disease-disease similarity

3d. Heatmap of miRNA-disease association

3e. Count of target genes per miRNA

3f. Distribution for confidence scores of gene-disease association

### 4.3 Training

Our dataset includes 10532 nodes and 1594438 edges. We used all six kinds of edges for message passing but we only update our model based on miRNA and disease edges. We have 146623 edges corresponding to the miRNA-disease relationship, and we split edges into training, validation, and testing datasets. In the end, we have 259967 training edges, 17331 validation edges, and 69325 testing edges. To prevent the model from overfitting and to increase the generalizability, we add a relatively small amount of negative edges between some miRNA and diseases. This is called negative sampling and we hope by using negative sampling, our model can learn to make more accurate predictions on a wider range of data. There have been many cases that negative sampling can be especially useful in large graphs, where the number of potential connections can be very large.

Hyperparameter tuning is crucial for achieving good results with deep learning models, it allows us to adjust the various settings or parameters of the model. While training the graph model, we performed hyperparameter tuning over the following grid of values to pick the best model architecture: the size of hidden state {8, 16, 32, 64, 128} and the number of graph convolutional layers {1, 2, 3}, the size of output layer {32, 64, 128}, the learning rate {0.1,0.05, 0.01, 0.001}. We used Adam optimize and BCELoss function, with random seed 0.

### 4.4 Results

The GCN model we train gets the best training and validation results with a hidden state size of 32, an output size of 64, a learning rate of 0.01, and 20 epochs. After 10 epochs of training, the training loss and validation loss stabilized at the lowest level. The model gives 1.0 accuracy on the testing set. By running the model three times, we get the testing accuracy of 1.0 +-0.

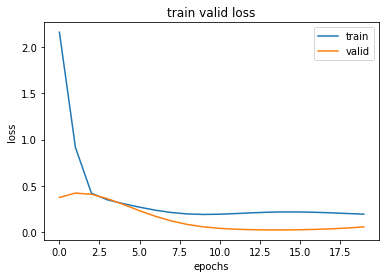


Figure 4. Training and validation loss

In the second part of the project, we built a heterogeneous graph model to include different kinds of relationships between types of nodes. However, the model performance is not very satisfying. The later part of this report will focus on the result of the GCN model.

### 4.5 Comparison with potential baselines

The baseline we choose for this project is a miRNA-disease network for association score prediction. They have 5,430 verified associations between 495 miRNAs and 383 diseases. We have fewer diseases in our dataset due to the loss of samples during the process of combining different datasets, but we have far more miRNA given that we include the miRNA-gene relationship in our model. Their model performance evaluated using AUROC is 0.9751.

### 4.6 Analysis

##### 4.6.1 Model works well in manually collected out-of-network relationships

To validate the model, we manually collect 12 miRNA-disease relationships from two recent experimental studies. The diseases and miRNA are included in our dataset for training and testing, but none of these relationships(edge) exist in our graph. All of them can be correctly predicted by our model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Disease | PubMed ID | number of miRNAs | number of edges exist in graph | Correctly predicted edges |
| leukemia | PMID: [34281210](https://pubmed.ncbi.nlm.nih.gov/34281210) | 7 | 0 | 7 |
| acquired immunodeficiency syndrome | PMID: [35601502](https://pubmed.ncbi.nlm.nih.gov/35601502) | 5 | 0 | 7 |

##### Table 1. Validation on manually collected out-of-network relationships

Given the 100% accuracy in both the test sets and manually collected cases, we examined our model by reducing the number of training epochs. It turns out that the model is still given a perfect number for the testing set, but the prediction performance for manually collected out-of-network edges is reduced. This indicates that the model learned the representation and relationships between nodes.

##### 4.6.2 Check node embedding in tSNE

In order to check whether the node embeddings really learn something, we used tSNE to plot the node embeddings. In general, the same type of nodes is clustering together, which

means the same types of nodes have similar embeddings.

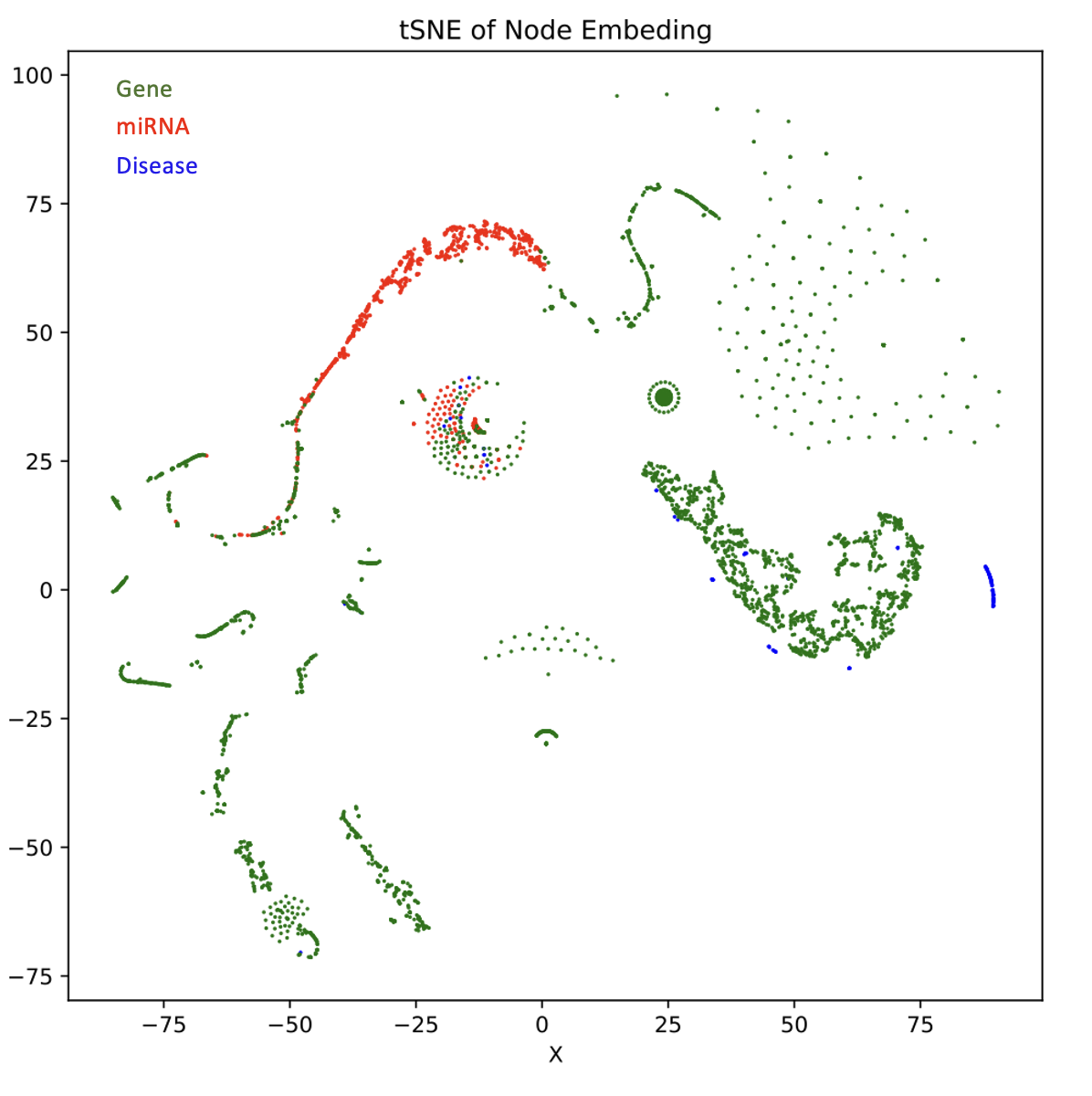


Figure 5. tSNE of node embeddings

##### 4.6.3 Comparison of original similarity score and embedding similarity

We calculated the similarities of embeddings of each node type, then compare the embedding similarity with the original similarity. The similarity of embedding is defined as the inner dot of 2 node embeddings. We expected to see a positive correlation between the two similarities, however, as figure 6 shows, regardless of the values of the original similarities, all embedding pairs show large embedding similarities. That could be a potential interpretation of the 100% accuracy of our training and test.

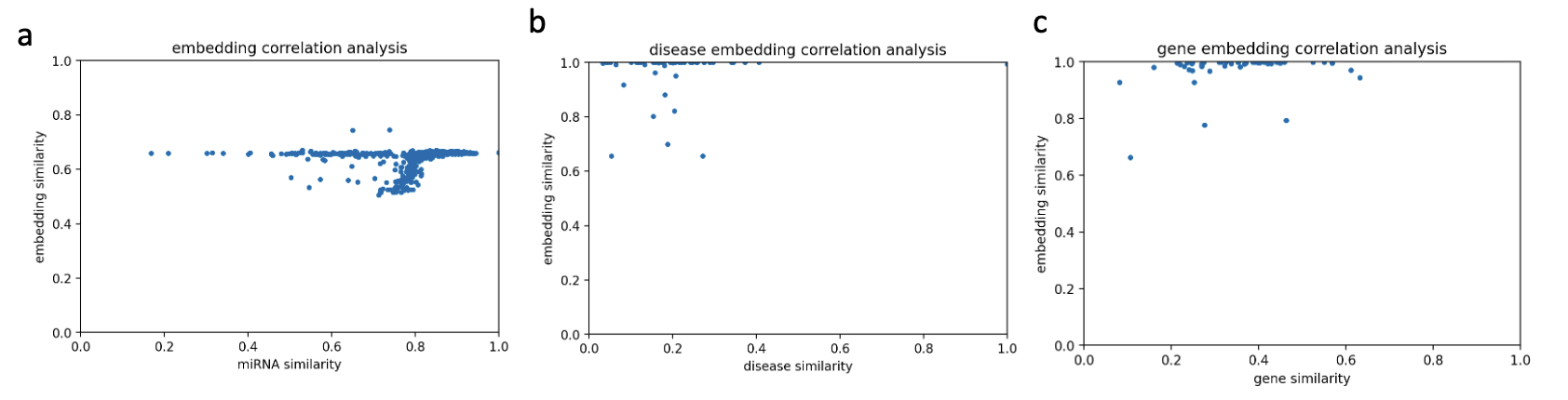


Figure 6. Node embedding similarity v.s. original similarity

6a. miRNA node embedding similarity v.s. original similarity

6b. Disease node embedding similarity v.s. original similarity

6c. Gene node embedding similarity v.s. original similarity

## 5. Conclusion

We created a heterogeneous graph with 3 node types: miRNA, gene, disease, and 6 edge types to represent the similarity or association between 2 nodes. We encoded the node types by one-hot encoding to distinguish the different node types and built 2 GCN layers to learn node embeddings. Then use the output of the last GCN layer to predict whether there is an edge between two nodes. The accuracy of the 3 tests is 100% +/- 0. We also manually collected disease leukemia and acquired immunodeficiency syndrome-related miRNAs as further validation. In summary, our method works pretty well on this miRNA-gene-disease prediction task.

## Reference

[1] Wikipedia contributors. (2022, October 9). MicroRNA. Retrieved October 14, 2022, from https://en.wikipedia.org/wiki/MicroRNA

[2] Ardekani AM, Naeini MM. The Role of MicroRNAs in Human Diseases. Avicenna J Med Biotechnol. 2010 Oct;2(4):161-79. PMID: 23407304; PMCID: PMC3558168.

[3] Lagos-Quintana M., Rauhut R., Yalcin A., Meyer J., Lendeckel W., Tuschl T. Identification of tissue-specific microRNAs from mouse. Curr Biol. 2002;12:735–739.

[4] Mattick JS, Makunin IV. Small regulatory RNAs in mammals. Hum Mol Genet. 2005;14(Suppl 1):R121–R132.

[5] Berezikov E, Plasterk RH. Camels and zebrafish, viruses and cancer: a microRNA update. Hum Mol Genet. 2005;14(Suppl 2):R183–R190.

[6] Bartel B. MicroRNAs directing siRNA biogenesis. Nat Struct Mol Biol. 2005;12(7):569–571.

[7] Zamore PD, Haley B. Ribo-genome: the big world of small RNAs. Science. 2005;309(5740):1519–1524.

[8] Croce CM, Calin GA. miRNAs, cancer, and stem cell division. Cell. 2005;122(1):6–7.

[9] Klein ME, Impey S, Goodman RH. Role reversal: the regulation of neuronal gene expression by microRNAs. Curr Opin Neurobiol. 2005;15(5):507–513.

[10] Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, et al. MicroRNAs regulate brain morphogenesis in zebrafish. Science. 2005;308(5732):833–838.

[11] Naguibneva I, Ameyar-Zazoua M, Polesskaya A, Ait-Si-Ali S, Groisman R, Souidi M, et al. The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation. Nat Cell Biol. 2006;8:278–284.

[12] Li Y, Kowdley KV. MicroRNAs in common human diseases. Genomics Proteomics Bioinformatics. 2012 Oct;10(5):246-53. doi: 10.1016/j.gpb.2012.07.005. Epub 2012 Sep 29. PMID: 23200134; PMCID: PMC3611977.

[13] Babbi, G., Martelli, P.L., Profiti, G. et al. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. BMC Genomics 18 (Suppl 5), 554 (2017). https://doi.org/10.1186/s12864-017-3911-3

[14] Yanaihara N., Caplen N., Bowman E., Seike M., Kumamoto K., Yi M. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell. 2006;9:189–198.

[15] Nunez-Iglesias J., Liu C.C., Morgan T.E., Finch C.E., Zhou X.J. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer’s disease cortex reveals altered miRNA regulation. PLoS One. 2010;5:e8898.

[16] Fabian Kern et.al., miRTargetLink 2.0—interactive miRNA target gene and target pathway networks, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W409–W416

[17] Jin Li, Sai Zhang, Tao Liu, Chenxi Ning, Zhuoxuan Zhang, Wei Zhou, Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction, Bioinformatics, Volume 36, Issue 8, 15 April 2020, Pages 2538–2546, <https://doi.org/10.1093/bioinformatics/btz965>

[18] Ji, BY., You, ZH., Cheng, L. et al. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. Sci Rep 10, 6658 (2020). <https://doi.org/10.1038/s41598-020-63735-9>

[19] Zhao, C., Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. Sci Rep 8, 15107 (2018). https://doi.org/10.1038/s41598-018-33219-y