*Briefings in Bioinformatics,* 2023, pp. 1–13

**doi: DOI HERE**
Advance Access Publication Date: Day Month Year
Paper

PAPER

# Generalizable prediction of potential miRNA-disease associations based on heterogeneous graph learning

Yi Zhou, Meixuan Wu, Chengzhou Ouyang, Xinyi Wang and Min Zhu

College of Computer Science, Sichuan University, No.24 South Section 1 Yihuan Road, 610065, Chengdu, China
Corresponding author: Min Zhu. Tel: 86-13980020500; Fax: 86-028-85469560; E-mail: zhumin@scu.edu.cn
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

## Abstract

Biomedical studies have revealed the crucial role of miRNAs in the progression of many diseases, and computational prediction methods are increasingly proposed for assisting biological experiments to verify miRNA-disease associations (MDAs). The generalizability is a significant issue, the prediction ought to be available for entities with fewer or without existing MDAs, while it is previously underemphasized. In this study, we work on the stages of data, model, and result analysis. First, we integrate multi-source data into a miRNA-PCG-disease graph, embracing all authoritative recorded human miRNAs and diseases, and the verified MDAs are split by time and known degree as a benchmark. Second, we propose an end-to-end data-driven model that avoids taking the existing MDAs as an input feature. It performs node feature encoding, graph structure learning, and binary prediction centered on a heterogeneous graph transformer. Finally, computational experiments indicate that our method achieves state-of-the-art performance on basic metrics and effectively alleviates the neglect of less and zero known miRNAs and diseases. Predictions are conducted on all human miRNA-disease pairs, case studies further demonstrate that we can make reliable MDA detections on unseen diseases, and the prediction basis is instance-level explainable.

**Key words:** miRNA-disease association, heterogeneous graph neural networks, explainable AI

## Introduction

MicroRNAs (miRNAs) are a class of endogenous non-coding RNAs that can play important regulatory roles in animals and plants by targeting messenger RNAs (mRNAs) for cleavage or translational repression, which can influence the output of many protein-coding genes (PCGs) [1, 2]. Since causing the unexpected output of PCGs, miRNAs can participate in the process of disease onset and progression, and the identification of miRNA-disease associations (MDAs) is of great value for disease diagnosis and treatment. For example, the mRNA expressions of serum miR-155-5p was gradually increased with the aggravation of sepsis [3]. And miR-155-5p could be useful biomarkers for detecting sepsis, a clinical serum specimen is also indicated for diagnostic purposes [4].

To date, lots of MDAs have been verified through biological experiments. However, the advancement of biological experiments is largely limited by being expensive and time-consuming, and computational prediction methods are increasingly proposed to detect potential MDAs for assistance. We summarize that there are three main stages in the research of MDA predictions: (i) How to organize sufficient data to describe the MDA issue. (ii) How to propose an applicable model to make effective MDA predictions. And (iii) how to understand and explain prediction results.

In stage (i), the most commonly adopted input features are the existing MDAs and disease father-son relations, presented as miRNA-disease adjacency matrix and directed acyclic graph of diseases [5, 6, 7, 8, 9, 10]. And studies [11, 12, 13] transform miRNA families into associations since miRNAs belonging to the same family tend to execute similar biological functions. Based on the process of miRNAs regulating diseases through PCGs, studies [14, 15, 16, 12, 13] introduce miRNA-PCG and PCG-disease associations to further describe the MDA issue by PCG-mediated pathways. However, while integrating all these inter- and intra-class associations into a unified miRNA-PCG-disease graph, node features with biomedical semantics are always ignored. Although miRNA sequences are considered by some studies [16, 17, 18], they are out of the heterogeneous graph and processed separately.

In addition to the comprehensiveness of the input feature types, the available learning and prediction scope matters more. Organizing the heterogeneous data as extensible, and allowable to embrace growing entities and associations should be encouraged. Especially for the prediction target, based on a public database like HMDD [19] that curated experiment-supported evidence for human MDAs, most existing studies make predictions within the entities mentioned in the database only, while the rest greater proportion of human miRNAs and diseases are excluded. As a precondition stage, a good benchmark

dataset covering more miRNAs and diseases, containing more MDAs, and with a reasonable split for evaluating the basic performance and generalizability of models is urgently needed. Studies [12, 13] take the subsets through several set operations between two versions of HMDD for computational experiments and allow for a wider scope of predictions. But they can still be improved by making all authoritative recorded human miRNAs and diseases available and merging more verified MDAs from different databases into the dataset.

In stage (ii), mainstream computational models conduct secondary feature extraction first [20, 21, 5]. Various similarities are calculated to extract key information, for example, disease semantic similarity [22] represents the father-son relations of diseases, miRNA sequence similarity [23] compares sequences among miRNAs, Gaussian interaction profile kernel similarity [24] encodes the miRNA-disease adjacency matrix, and miRNA functional similarity [25] relies on the similarities of miRNA-associated disease sets. And some embedding algorithms like Node2Vec [26, 18] and SDNE [27, 13] are employed to extract surrounding topologies of nodes. The well-designed heuristic similarities and network embeddings are all proven to help in downstream link predictions. However, since the adaptability and expressiveness are limited for these handcrafted or shallow encodings extractions, they can be the bottleneck if there is a higher goal, e.g. generalizability.

With the extracted secondary features presenting key information, machine learning models are employed for prediction. NEMII [11] and DFELMDA [6] make classification by random forest. ERMDA [7] adopts an ensemble learning framework. In particular, graph neural networks (GNNs) are increasingly exploited in the MDA prediction which is naturally described by graphs, to further fuse and enhance representations. NIMCGCN [8] converts similarities into similarity networks and applies graph convolutional networks (GCNs) on which. MMGCN [16] implements GCNs on several similarity networks and combines them with a multichannel attention mechanism. MINIMDA [9] implements modified GCNs on similarity and association multimodal networks. AGAEMD [10] builds a graph autoencoder on the miRNA-disease bipartite graph where the encoder includes graph attention networks and a jumping knowledge module. Between the requirement to input node features into GNNs and the actuality of missing biomedical semantic node features, the above studies perform augmentation by filling node features with randomly initialed vectors or corresponding slices of adjacency and similarity matrices.

Message-passing GNNs pass and aggregate information from neighbors, thereby learning from the graph structure [28]. It's convoluted to conclude graph structures into similarities, newly construct similarity graphs, and then perform GNNs that essentially learn graph structures. On the one hand, it relies on the existing MDAs as an input feature, leading to the neglect of entities with fewer or without existing MDAs. On the other hand, there are more obstacles to the explanation of predictions. Finally, it leads to a strong demand for stage (iii): What types of features contribute more to the performance? Can the prediction basis be revealed for each miRNA-disease pair? While generalizing into unknown regions, besides prediction scores, case-by-case explanations are more needed. Existing studies are inadequate to make prediction results comprehensible.

To address the above shortcomings, (i) we construct a dataset containing all authoritative recorded human miRNAs and diseases, relevant information is organized as a heterogeneous graph with biomedical semantic node features. (ii) We propose a data-driven model that learns the heterogeneous graph sufficiently, which does not extract secondary features or take existing MDAs as input. And (iii) we present analysis by metrics comparisons, visualizations, and case studies to demonstrate the effectiveness of our method. Main contributions can be summarized as follow:

- We integrate a miRNA-PCG-disease graph from multi-source databases through a reliability-guaranteed and extension-friendly process. We employ a massive MDA database and design a time and known degree-based split as a benchmark.
- We propose an intuitive prediction model EGPMDA that end-to-end performs node feature Encoding, Graph structure learning, and binary Prediction sequentially, where the central module is a heterogeneous graph transformer.
- Computational experiments indicate that EGPMDA achieves state-of-the-art performance on basic metrics and effectively alleviates the neglect of less and zero known nodes. Case studies further illustrate that it can make instance-level explainable MDA detections on unseen diseases.

## Materials and Data Processing

### Dataset Construction

To better formulate the MDA prediction problem, we integrate rich relevant information from multiple sources into a miRNA-PCG-disease graph. It is stated as undirected here and would be processed into directed for modeling. To better evaluate computational models, we split the verified MDAs into the training, validation, and test sets by publication year of literature evidence, and the test set is grouped into subsets by the known degree of miRNAs and diseases. We describe the dataset construction process in steps as follows, and Table 1 illustrates the source and volume statistics. For more details, please visit the data and codes at `https://github.com/EchoChou990919/EGPMDA`.

**Step 1**. Determine standard nodes. All miRNAs, diseases, and PCGs of homo sapiens are included in our dataset as standard nodes based on the authoritative databases miRBase [29, 30], MeSH[1], and HGNC[2]. And we take the miRBase "Accession", MeSH "UI", and HGNC "Entrez ID" as primary IDs, respectively. Meanwhile, possible alternative symbols (aliases) of nodes are recorded serving for subsequent edge acquisition.

**Step 2**. Obtain biomedical semantic node features. We take miRNA stem-loop and mature sequences, disease heading and scope note, PCG name and belonging group name as the features of three types of nodes.

**Step 3**. Obtain intra-class edges. MiRNA-miRNA edges are transformed from the miRNA families where miRNAs belonging to the same family are fully connected. Disease-disease edges present the semantic father–son relations among diseases. Similarly to miRNAs, PCGs belonging to the same group are linked to each other, forming PCG-PCG edges.

**Step 4**. Obtain inter-class edges. MiRNA-PCG associations are downloaded from ENCORI [31], presenting the degradome between miRNAs and mRNAs (transcripted by PCGs). Disease-PCG associations are acquired from DisGeNet [32], and we only preserve the ones with the evidence-level score $\geq 0.1$. Through direct primary ID matching or indirect "aliases matching - primary ID transfer", both ends of these associations are mapped to standard nodes, forming inter-class edges after de-duplication.

[1] `https://www.nlm.nih.gov/mesh/`

[2] `https://www.genenames.org/`

**Table 1.** Statistics of our miRNA-PCG-disease graph

| | | Source | $\|miRNA\|$ | $\|PCG\|$ | $\|Disease\|$ | $\|Edge\|$ |
|---|---|---|---|---|---|---|
| Nodes | miRNA | miRBase | **1917** | / | / | / |
| | PCG | HGNC | / | 19229 | / | / |
| | disease | MsSH | / | / | **4933** | / |
| Intra-class Edges | miRNA-miRNA | miRBase | 576 | / | / | 4500 |
| | PCG-PCG | HGNC | / | 14281 | / | 1219564 |
| | disease-disease | MsSH | / | / | 4933 | 7678 |
| Inter-class Edges | miRNA-PCG | ENCORI | 1855 | 14452 | / | 144636 |
| | PCG-disease | DisGeNet | / | 11316 | 2911 | 134805 |
| | **miRNA-disease** | **RNADisease** | **1874** | / | **968** | **57298** |

**Step 5**. Obtain miRNA-disease associations. MDAs are derived from RNADisease [33], which integrates manual curation of numerous literature and other experimentally verified databases, covering the widely used HMDD. And records without available PMID are filtered out, ensuring that every MDA can be backdated to its literature evidence. Similarly, these verified MDAs are mapped to standard miRNA and disease nodes. In addition, we get the publication year of literature by PMID through the "Bio.Entrez" package.

**Step 6**. Split MDAs by time and known degree. We use 39991 MDA samples first verified before 2019 as the training set (69.79%), 6268 samples between 2019 and 2020 as the validation set (10.94%), and 11039 samples between 2021 and 2022 as the test set (19.27%). According to MDAs in the training and validation sets that can be "known" during model training, we group miRNAs and diseases into the zero, less, and more known, and further group the test set into corresponding subsets.

Eventually, we construct a dataset containing information on miRNA sequences, disease description texts, gene name texts, miRNA families, disease father-son relations, gene groups, miRNA-gene associations, disease-gene associations, and MDAs. All are unified into a heterogeneous graph, where the nodes are entities recorded by authoritative databases, and the edges are supported by trustworthy evidence. The construction process is extension-friendly, for example, the PCG-PCG edges can be replaced by the gene functional networks from HumanNet [34].

## Data Observation

Visualizations of the miRNA-disease adjacency matrix (Figure 1) show the distribution of samples. Compared to the blank to be explored, verified MDAs take only a small proportion, which emphasizes the significance of generalizable prediction. And there are distributional differences among the training, validation, and test samples, the 11039 test samples locate a bit more dispersed, partially toward the less and zero known nodes. Biomedical experimental researches are progressive, verified MDAs "spread" gradually from traditionally more known entities to others over time with new coming research interests. Therefore, our time-based dataset split is reasonable and propitious to evaluating both basic performance and generalizability, especially through analyzing successful MDA detections on the almost-blank subsets.

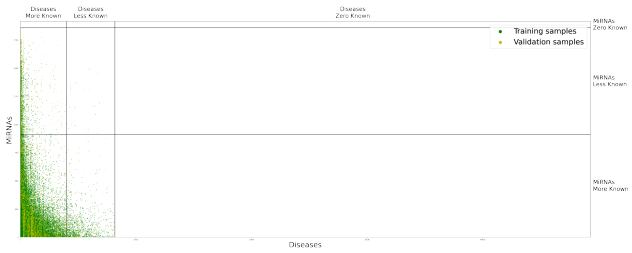
Moreover, as seen in Supplementary File SF.1, we make an observation of the graph structures around miRNA-disease pairs via common neighbor statistics. It has driven our model design that there is an instinctive need to learn the surrounding graph structures in some way like GNNs. However, the critical information hides in noises, which requires the GNNs to estimate critical neighbors accurately, aggregating important messages with higher weights.

## Notations and Preprocessing

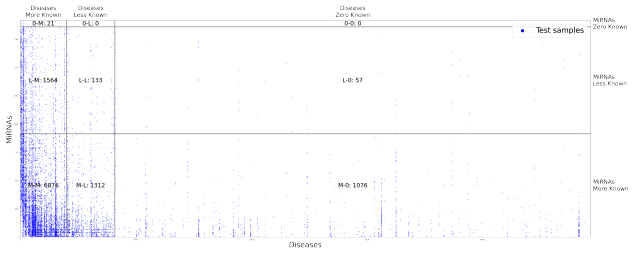
The **Heterogeneous Graph** is defined as a directed graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \tau, x, \phi)$, where node $v \in \boldsymbol{V}$ with type $\tau(v)$ and feature vector $x_v$, and edge $e \in \boldsymbol{E}$ with type $\phi(e)$. The **Meta-Relation** describes the edge $e = (s, t)$ linked from source node $s$ to target node $t$ as a tuple $\langle \tau(s), \phi(e), \tau(t) \rangle$.

For subsequent modeling, we preprocess the miRNA-PCG-disease graph as follows:

RNA sequence consists of four bases 'A' (adenine), 'U' (uracil), 'C' (cytosine), and 'G' (guanine) naturally. We simply adopt the classic 1-mer representation to transfer an RNA sequence to a vector: unify the sequence to an upper limit length by filling the end with placeholder base 'N'. Along the sequence, map 'A', 'U', 'C', 'G' to one-hot vectors and




(a) Training and validation samples




(b) Test samples

**Fig. 1.** The miRNA-disease adjacency matrix. (a) Describing the occurrence number of miRNAs or diseases in the training and validation sets as the degree of "known", the median is 8 for miRNAs and 10 for diseases. Sorting miRNAs and diseases by the degree, the combination of zero (0), less (0 to median), and more (median to max) known miRNAs and diseases divide the adjacency matrix into nine regions. (b) Splitting the test set into corresponding subsets, the sample sizes are labeled in the figure. We take the three in the bottom left as the sparse, and the remaining four with non-zero sample size as the almost-blank.

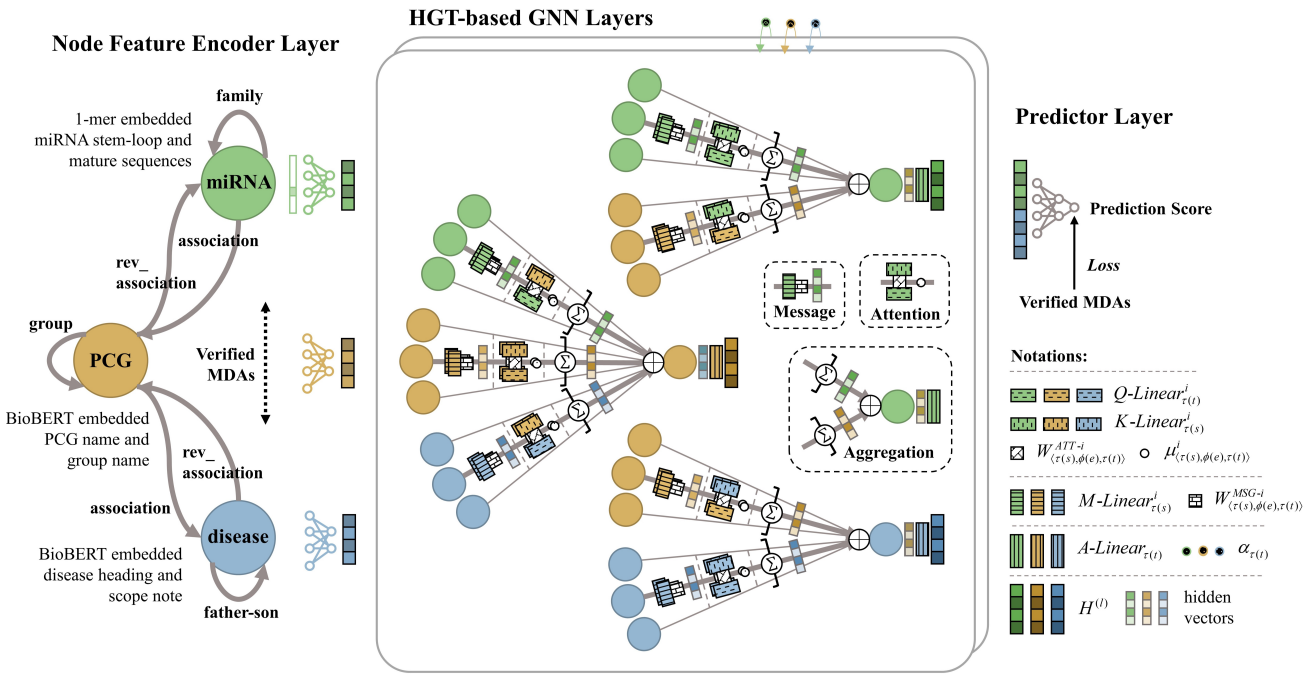



**Fig. 2.** The flowchart of EGPMDA. Relevant data is integrated into a miRNA-PCG-disease graph with biomedical semantic node features, where we exclude existing MDAs from the graph structure. In the end-to-end prediction process, there is an encoder layer learning node features, stacked GNN layers learning graph structures, and a predictor layer deriving classifications. The central module is a heterogeneous graph transformer (HGT), by performing heterogeneous mutual attention, heterogeneous message passing and target-specific aggregation, it focuses more on critical information, and we can extract the attention and residual scores for explanation.

'N' to $[0.25, 0.25, 0.25, 0.25]^T$. For each miRNA, there are one stem-loop sequence and up to two mature sequences. Perform the above 1-mer representation separately and concatenate the vectors together, we get $x_m \in \mathbb{R}^{(l_s+l_{m_1}+l_{m_2})\times 4}$, where $l_s$, $l_{m_1}$ and $l_{m_2}$ denote the corresponding maximum lengths.

BioBERT [35] is a domain-specific language model pre-trained on large-scale biomedical corpora, which can be invoked to derive contextualized embeddings from biomedical concepts-related texts. For each disease, we process the heading and the scope note by BioBERT and concatenate vectors together as $x_d \in \mathbb{R}^{2*d_B}$, where $d_B$ is the default embedding size. Similarly, for each PCG, we get $x_g \in \mathbb{R}^{2*d_B}$ from the name and the belonging group name.

For the sake of adequate GNN message passing, we add "reverse" connections for edges and add self-loops to the graph. In conclusion, the node types $\tau(v) \in \{$ `miRNA, disease, PCG` $\}$, and the meta-relations $\langle\tau(s), \phi(e), \tau(t)\rangle \in \{$ `<miRNA, family, miRNA>`, `<disease, father-son, disease>`, `<PCG, group, PCG>`, `<miRNA, association, PCG>`, `<PCG, rev_association, miRNA>`, `<PCG, association, disease>`, `<disease, rev_association, PCG>` $\}$.

## Methods

In this work, we define the MDA prediction as a binary link prediction towards miRNA-disease pairs on the miRNA-PCG-disease graph. As shown in Figure 2, inspired by the design space of GNNs [36], we propose an end-to-end model **EGP**MDA, which includes three intuitive parts: a node feature **E**ncoder layer to learn biomedical semantics, a stack of **G**raph neural network layers to learn heterogeneous graph structures, and a **P**redictor layer to derive classifications.

We denote the output of $(l)$-th layer as $H^{(l)}$, where $l = 0$ for the encoder layer and $l \in [1, L]$ for GNN layers. The output of the current layer is the input to the next layer until the last layer makes the prediction. For simplicity, sth.-Linear $(x)$ below denotes a linear transformation $xW + b$, with a parameterized weight matrix $W$ and a learnable bias $b$.

### Node Feature Encoder Layer

As formulated by equation 1, the first step is to encode the feature vectors of nodes. For miRNA, a single convolution shaped $k \times 4$ filters the subsequences of length $k$ along the 1-mer embedded miRNA sequences $x_m$ with stride 1 and no padding, followed by a linear transformation to project it to $\mathbb{R}^d$. And for disease and PCG, $x_d$ and $x_g$ are linearly transformed in parallel into $\mathbb{R}^d$ as well.

$$\begin{aligned} H^{(0)}[m] &= \text{ME-Linear}\left(\text{Conv}\left(x_m, (k, 4)\right)\right) \\ H^{(0)}[d] &= \text{DE-Linear}\left(x_d\right) \\ H^{(0)}[g] &= \text{GE-Linear}\left(x_g\right) \end{aligned} \tag{1}$$

### Graph Neural Network Layers

The stacked $L$ GNN layers learn subgraph structure around miRNA-disease pairs of up to $L$-hop. Meeting the observation of data, we adopt the Heterogeneous Graph Transformer (HGT) [37] to extract critical information. By exploiting the powerful Transformer architecture, HGT learns heterogeneous mutual attention weights for source-to-target aggregation while modeling message passing.

Each HGT layer begins with the **Heterogeneous Mutual Attention**. As formulated by Equation 2, it calculates $h$ heads of attention for each edge $e = (s, t)$. For the $i$-th head attention

$\mathrm{ATT}^i(s,e,t)$, there are four main groups of learnable parameters: First, Q-Linear$^i_{\tau(t)}$ transforms the $\tau(t)$-type target node $t$ into a Query vector $Q^i(t)$, mapping $\mathbb{R}^d$ to $\mathbb{R}^{\frac{d}{h}}$. Parallelly, K-Linear$^i_{\tau(s)}$ projects $\tau(s)$-type source node $s$ into a Key vector $K^i(s) \in \mathbb{R}^{\frac{d}{h}}$. Matrix $W^{ATT\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle} \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ describes the distinct semantic of each meta-relation from $s$ to $t$, and the similarity between Query and Key vectors is calculated through two continuous matrix multiplications. Especially, $\mu^i_{\langle\tau(s),\phi(e),\tau(t)\rangle}$ scales to the attention adaptively, estimating the general importance of each meta-relation. It is initialized to 1, and the learned variations could help us understand what meta-relations are focused more or less on.

Finally, $h$ heads of attention vectors are concatenated together. Grouping the attentions by meta-relations, given a target node $t$, $s \in N_{meta}(t)$ denotes all of its source neighbors along the meta-relation $\langle\tau(s),\phi(e),\tau(t)\rangle$. Then for each target node, the heterogeneous mutual attention fulfills $\sum_{s\in N_{meta}(t)} \mathbf{ATT}_{\mathbf{HGT}}(s,e,t) = 1_{h\times 1}$ through a softmax.

$$
\begin{aligned}
\mathbf{ATT}_{\mathbf{HGT}}(s,e,t) &= \underset{\forall s\in N_{meta}(t)}{\mathrm{Softmax}}\left(\Big\|_{i\in[1,h]} \mathrm{ATT}^i(s,e,t)\right)\\
\mathrm{ATT}^i(s,e,t) &= \left(K^i(s) W^{ATT\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle} Q^i(t)^T\right)\cdot \frac{\mu^i_{\langle\tau(s),\phi(e),\tau(t)\rangle}}{\sqrt{d}}\\
K^i(s) &= \text{K-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right)\\
Q^i(t) &= \text{Q-Linear}^i_{\tau(t)}\left(H^{(l-1)}[t]\right)
\end{aligned}
\tag{2}
$$

Parallel to the attention calculation, the HGT layer performs the **Heterogeneous Message Passing**. As formulated by Equation 3, it learns the information passing from the source nodes by each meta-relation. For the $i$-th head message $\mathrm{MSG}^i(s,e,t)$, there are two main parts of learnable parameters: M-Linear$^i_{\tau(s)}$ linearly transforms the $\tau(s)$-type source node $s$ from $\mathbb{R}^d$ to $\mathbb{R}^{\frac{d}{h}}$. Followed another meta-relation-based matrix $W^{MSG\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle} \in \mathbb{R}^{\frac{d}{h}\times\frac{d}{h}}$ incorporates the meta-relation semantic during message passing. Furthermore, all $h$ heads of message vectors of $e=(s,t)$ are concatenated together to derive $\mathbf{MSG}_{\mathbf{HGT}}(s,e,t)$.

$$
\begin{aligned}
\mathbf{MSG}_{\mathbf{HGT}}(s,e,t) &= \Big\|_{i\in[1,h]} \mathrm{MSG}^i(s,e,t)\\
\mathrm{MSG}^i(s,e,t) &= \text{M-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) W^{MSG\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle}
\end{aligned}
\tag{3}
$$

After the heterogeneous mutual attention and messages are calculated, the HGT layer conducts **Target-Specific Aggregation**, aggregating the messages from all source nodes to the target node within and across meta-relations. As formulated by Equation 4, within each meta-relation, the attention scores summing to 1 are the weights to average corresponding messages. And Meta-Agg denotes the function for aggregating different meta-relations, set as Sum by default. There are also two sets of learnable parameters: A-Linear$_{\tau(t)}$ projects the heterogeneously aggregated vector with GELU activation back to the $\tau(t)$-type specific feature distribution. And $\alpha_{\tau(t)}$ with Sigmoid activation controls residual connections, deriving the output $H^{(l)}[t]$.

$$
\begin{aligned}
H^{(l)}[t] =& \sigma_s\left(\alpha_{\tau(t)}\right)\cdot \text{A-Linear}_{\tau(t)}\left(\sigma_g\left(\widetilde{H}^{(l)}[t]\right)\right) +\\
&\left(1-\sigma_s\left(\alpha_{\tau(t)}\right)\right)\cdot H^{(l-1)}[t]\\
\widetilde{H}^{(l)}[t] =& \underset{\forall\langle\tau(s),\phi(e),\tau(t)\rangle}{\text{Meta-Agg}}\\
&\left(\sum_{s\in N_{meta}(t)} \mathbf{ATT}_{\mathbf{HGT}}(s,e,t)\cdot \mathbf{MSG}_{\mathbf{HGT}}(s,e,t)\right)
\end{aligned}
\tag{4}
$$

## Predictor Layer

As formulated by Equation 5, for each candidate miRNA-disease pair, the corresponding miRNA and disease vectors are concatenated together. Two sequential linear transformations map it from $\mathbb{R}^{2d}$ to $\mathbb{R}^d$ and to $\mathbb{R}^1$. Ultimately, EGPMDA produces the MDA prediction score $y$ through a sigmoid activation, where a score closer to 1 indicates a higher probability of association.

$$
y = \sigma_s\left(\text{P}_2\text{-Linear}\left(\text{P}_1\text{-Linear}\left(H^{(L)}[m]\,\|\,H^{(L)}[d]\right)\right)\right) \tag{5}
$$

As formulated by Equation 6, we adopt binary cross entropy as the loss function for optimizing our model. Where $\hat{y}$ denotes the supervision label, $\hat{y}=1$ if the miRNA-disease pair is verified associated, otherwise $\hat{y}=0$.

$$
Loss = -\sum\left[\hat{y}\cdot log(y) + (1-\hat{y})\cdot log(1-y)\right] \tag{6}
$$

# Results

In this section, we evaluate our method comprehensively, aiming at investigating the following questions:

- **RQ1**. How does EGPMDA perform vs. state-of-the-art baseline methods overall?
- **RQ2**. How is the generalizability? Does EGPMDA alleviate the neglect of entities with fewer or without existing MDAs?
- **RQ3**. How does each part of the miRNA-PCG-disease graph affect our method in performance?
- **RQ4**. Is our method explainable? Can we derive the prediction basis for each candidate miRNA-disease pair?

## Experimental Settings

**Implementation Details.** EGPMDA is implemented on PyTorch and PyTorch Geometric, and the codes are available at `https://github.com/EchoChou990919/EGPMDA`. Taking verified MDAs as positive samples and random miRNA-disease pairs without verified association as negative samples. For training and validation sets, we keep an equal size for positive and negative samples. And for the test set, in addition to the balanced case, we further increase the negative sample size to 100 times. In selecting hyperparameters, the model is trained on the training set and evaluated on the validation set. After fixing all hyperparameters, it is trained on the union of training and validation sets to receive more supervision, and evaluated on the test set. All the computational experiments are repeated five times, more details are stated in Supplementary File SF.2.

**Evaluation Metrics.** The threshold for deciding the positive or negative is critical for binary classification that outputs a prediction score. AUC and AUPR are threshold-insensitive metrics, denoting the areas under the receiver operating characteristic and precision-recall curves, respectively. Accuracy,

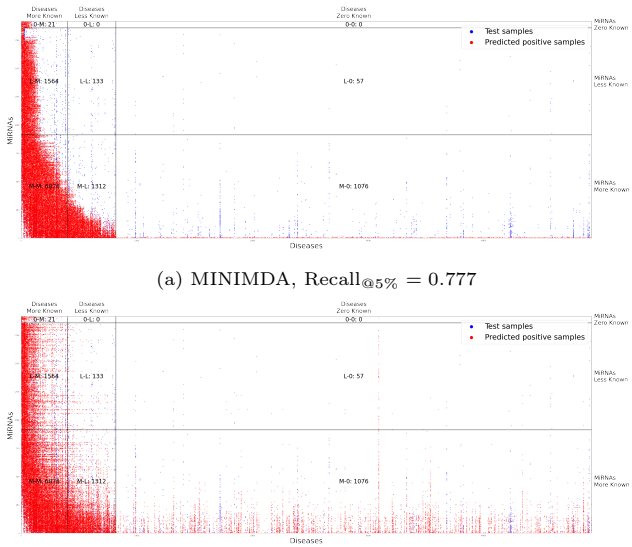
**Table 2.** Comparisons with Baseline Methods

| Method | Positive & Negative Samples Balanced | | | | | | Imbalanced | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | AUPR | Accuracy | Precision | Recall | F1-score | AUC | AUPR | $\text{Recall}_{@5\%}$ | $\text{Recall}_{@10\%}$ |
| NIMCGCN | 0.824 | 0.848 | - | - | - | - | 0.824 | 0.091 | 0.443 | 0.616 |
| MMGCN | 0.867 | 0.882 | - | - | - | - | 0.867 | 0.135 | 0.558 | 0.708 |
| DFELMDA | .948 | .950 | .847 | 0.957 | 0.728 | 0.827 | - | - | - | - |
| MINIMDA | 0.951 | 0.951 | .842 | **0.961** | 0.713 | 0.819 | 0.950 | 0.286 | **0.778** | 0.862 |
| AGAEMD | 0.945 | 0.945 | 0.839 | 0.917 | 0.745 | 0.822 | 0.945 | 0.292 | 0.736 | 0.818 |
| **EGPMDA** | **0.954** | **0.952** | **0.864** | 0.943 | **0.774** | **0.850** | **0.953** | **0.295** | 0.758 | **0.872** |

Precision, Recall, and F1-score are threshold-sensitive metrics, for models trained by balanced supervision and making 0-1 symmetric prediction, the threshold is set as 0.5 by default. Furthermore, the top N-ranked samples can be considered as predicted positive and the others negative, metrics can be calculated as well, e.g. $\text{Recall}_{@N}$.

Noting true positive predictions as successful detections, Prediction or Recall indicates the proportion of successful detections among all positive predictions or all verified MDAs, respectively. Considering the MDA prediction aims at finding potential MDAs that have not yet been verified, and there can be actual MDAs but the evidence not included in the dataset, we take Recall as more important than Precision: the verified MDAs should be successfully detected as much as possible, and it's acceptable to predict a bit of "excess" miRNA-disease pairs as associated.

Therefore, we adopt average AUC, AUPR, Accuracy, Precision, Recall, and F1-score as evaluation metrics in the balanced case, and calculate average AUC, AUPR, $\text{Recall}_{@5\%}$, and $\text{Recall}_{@10\%}$ in the imbalanced case. Especially, for analysis of generalizability, we compare Recall on the almost-blank and sparse test subsets (introduced in Figure 1b) and visualize the distribution of top-ranked predictions.




(a) MINIMDA, $\text{Recall}_{@5\%} = 0.777$

(b) EGPMDA, $\text{Recall}_{@5\%} = 0.757$

**Fig. 3.** Top@5% predictions from MINIMDA and EGPMDA.

## Comparison to Baseline Methods

To demonstrate the effectiveness of our EGPMDA, we compare it with five baseline methods: NIMCGCN [8], MMGCN [16], DFELMDA [6], MINIMDA [9], and AGAEMD [10]. They each adopt and combine several types of similarity elaborately, which are all recalculated on our data. Information leakage is carefully avoided with only training and validation MDAs used in the similarity calculation, GNN message propagation, and supervised training processes. It should be noticed that our reproduction is already an extension of the original methods, their scopes of prediction are extended to all human miRNAs and diseases based on our dataset.

The comparison of EGPMDA and baseline methods is seen in Table 2. NIMCGCN and MMGCN are relatively underperforming on all available metrics. Since required to take the whole adjacency matrix for supervision, superabundant negative label information restrains their high score outputs, leading to unassociated predictions only. At a competitive level, our EGPMDA performs superior to DFELMDA, MINIMDA, and AGAEMD. In the balanced case, EGPMDA achieves the highest AUC, AUPR, Accuracy, Recall, and F1-score. Especially for Recall, it is 0.029 higher than the second-best AGAEMD, indicating approximately 320 ($0.029 \times 11039$) more successful MDA detections. In the imbalanced case, our method is best in AUC, AUPR, and $\text{Recall}_{@10\%}$, and is the second highest to MINIMDA in $\text{Recall}_{@5\%}$.

Further analysis of generalizability is shown in Table 3, demonstrating the average Recall by the percentage on four almost-blank and three sparse subsets. DFELMDA achieves the highest metric on *M-M* and AGAEMD performs the best on *M-L*, while they both make hardly any successful MDA detection in the almost-blank regions. In contrast, our EGPMDA not only performs well in sparse regions, achieving the highest Recall on *L-M* and the second highest on *M-L* and *M-M*. In particular, it derives the most effective MDA predictions on three of four almost-blank subsets. There are 1076 MDA samples in the *M-0* with the disease end never seen during the model training, and EGPMDA successfully detects 17.0% of them, far beyond the second-best MINIMDA which does only 2.9%, i.e. approximately 152 samples more.

Rethinking the imbalanced case, how is EGPMFA performing better overall but secondary to MINIMDA on $\text{Recall}_{@5\%}$? Figure 3 are visualizations of the Top@5% predictions (one of five repeats). It shows that MINIMDA's predictions tend to gather in sparse regions, but EGPMDA's predictions appear relatively more in the almost-blank regions. While taking an equal number of predictions as positive, the more "bravely" distribute in the almost-blank regions, the less opportunity for successful detection on sparse subsets which own the majority of positive labels. The visualizations once again demonstrate the generalizability of our EGPMDA, while at a small expense of $\text{Recall}_{@5\%}$.

**Table 3.** Comparisons with Baseline Methods: Analysis on Almost-blank or Sparse Subsets

| Recall | Almost-blank Subsets | | | | Sparse Subsets | | |
|---|---|---|---|---|---|---|---|
| Method | *0-M*$_{(21)}$ | *L-0*$_{(57)}$ | *M-0*$_{(1076)}$ | *L-L*$_{(133)}$ | *L-M*$_{(1564)}$ | *M-L*$_{(1312)}$ | *M-M*$_{(6878)}$ |
| DFELMDA | 0.0% | 0.0% | 0.0% | 0.0% | 50.2% | 52.4% | **95.4%**$_{(\uparrow 131)}$ |
| MINIMDA | 24.8% | 0.0% | 2.9% | 0.0% | 55.5% | 46.4% | 92.5% |
| AGAEMD | 4.8% | 0.0% | 0.0% | 0.0% | 52.5% | **78.7%**$_{(\uparrow 147)}$ | 92.7% |
| **EGPMDA** | **26.7%**$_{(\uparrow 1)}$ | 0.0% | **17.0%**$_{(\uparrow 152)}$ | **1.5%**$_{(\uparrow 2)}$ | **66.9%**$_{(\uparrow 179)}$ | 67.5% | 93.5% |

In conclusion, we can answer **RQ1** and **RQ2**: EGPMDA outperforms state-of-the-art baseline methods overall and shows good generalizability, alleviating the neglect of less and zero known miRNAs and diseases.

## Ablation study

We discuss the effect of each part in the miRNA-PCG-disease graph incrementally by the following five conditions, and for the **3** that we mainly adopt, the number of GNN layers $\boldsymbol{L}$ is also analyzed.

0 Utilize the supervision information only where node features of miRNAs and diseases are initialized randomly. Since without graph structure, the GNN layers are skipped.
1 Further utilize the biomedical semantic node ***Fea***tures: the embeddings representing miRNA sequences and disease description texts. Ditto, the GNN layers are excluded.
2 Further utilize the ***Intra***-class edges: the graph structure representing miRNA families and disease father-son relations.
**3** Further utilize the information relevant to ***PCG***s: the node features transferred from PCG name texts, and the graph structure representing PCG groups, miRNA-PCG associations, and disease-PCG associations.
4 Further utilize the existing ***MDA***s: the graph structure representing miRNA-disease associations.

Table 4 summarizes the results of our ablation study. From 0 to **3**, for almost all metrics, each additional utilization of a portion of the miRNA-PCG-disease graph brings a gain in performance. And stacking two GNN layers is optimal since the 2-hop subgraph "reception field" can cover the information about PCGs while still free from the over-smoothing. There is an explainable attention parameter $\mu^{i}_{\langle\tau(s),\phi(e),\tau(t)\rangle}$ (see Equation 2) describing the overall importance of meta-relations. As shown in Figure 4, the most important 2-hop paths are `<miRNA, family, miRNA, family, miRNA>`, `<PCG, rev_association, miRNA, family, miRNA>`, and `<disease, father-son, disease, father-son, disease>`. And from **3** to 4, it's generally competitive and even achieves higher balanced Precision and imbalanced AUPR and Recall$_{@5\%}$, but as discussed in Supplementary File SF.3, the reliance on miRNA-disease edges hurts the generalizability.

Therefore, we can answer **RQ3** and half of **RQ4**: every part of the miRNA-PCG-disease graph does contribute to the prediction performance, but it's better not to take existing MDAs into the GNN message passing for the pursuit of generalizability. And we can explain which meta-relations matter more in general.

## Case study

We take all verified MDAs in the dataset to retrain the model, predictions are performed on all miRNA-disease pairs, and two diseases without any MDA records are investigated for case studies. Latent autoimmune diabetes in adults (LADA, MeSH ID: D000071698) is discussed as follows, and Teratozoospermia is reported in Supplementary File SF.4.1.

LADA is a heterogeneous disease characterized by a less intensive autoimmune process and a broad clinical phenotype compared to classical type 1 diabetes mellitus (T1DM) [38]. We search for literature and collect the following specific descriptions about the associations of miRNAs with LADA. Table 5 shows the prediction results, and seven of the eight are successfully detected:

- Study [39]: "Quantitative real-time PCR (qRT–PCR) showed that hsa-miR-146a-5p, hsa-miR-21-5p and hsa-miR-223-3p were significantly upregulated in LADA patients compared with healthy controls."
- Study [40]: "People with LADA were best distinguished based on the levels of miR-34a, miR-24, and miR-21." and "miRNAs like miR-34a, miR-30d, and miR-24 could be useful to classify subjects with LADA."
- Study [41]: "microRNA-143-3p contributes to inflammatory reactions by targeting FOSL2 in PBMCs from patients with autoimmune diabetes mellitus (T1DM and LADA)."
- Study [42]: "The qRT-PCR results further suggest the capability of circulating miRNAs, at least hsa-miR-517b-3p, as the LADA biomarker."

For each miRNA-disease pair, we can explain the prediction by restoring the surrounding subgraph and extracting corresponding attention and residual values, i.e. the $\mathbf{ATT}_{\text{HGT}}(s,e,t)$ for meta-relations and $\sigma_s\left(\alpha_{\tau(t)}\right)$ for nodes in each HGT layer. The case of "hsa-miR-143-3p - LADA" is visualized in Figure 5. It shows that the semantics of "LADA" is mainly aggregated from "Diabetes Mellitus", "Diabetes Mellitus, Type 1", "Autoimmune Diseases" and itself. The attention mechanism of HGT captures the closer relations correctly. Since there is a verified association "hsa-miR-143-3p - Diabetes Mellitus" included in the dataset, our method recommends that hsa-miR-143-3p is potentially associated with LADA. The explanations of "miR-34a - LADA" and "hsa-miR-517b-3p - LADA" are shown in the Supplementary File SF.4.2, we can see how is the basis of a higher (0.836) or lower (0.098) score prediction. It still follows the classic assumption that similar diseases tend to be associated with similar miRNAs, but the "similar" is not measured by heuristic similarity calculations, it implies in the message passing and aggregation of GNNs.

So we can answer **RQ4**: our method is explainable, we can explain the prediction basis on the instance level.

## Conclusion

The identification of miRNA-disease associations (MDAs) is of great value for disease diagnosis and treatment, and computational prediction methods are now increasingly proposed

**Table 4.** Ablation Study

| | Ablation Types | | | | | Positive & Negative Samples Balanced | | | | | | Imbalanced | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Fea* | *Intra* | *PCG* | *MDA* | *L* | AUC | AUPR | Accuracy | Precision | Recall | F1-score | AUC | AUPR | $\text{Recall}_{@5\%}$ | $\text{Recall}_{@10\%}$ |
| 0 | × | × | × | × | - | 0.884 | 0.890 | 0.816 | 0.902 | 0.710 | 0.794 | 0.881 | 0.126 | 0.582 | 0.734 |
| 1 | ✓ | × | × | × | - | 0.914 | 0.917 | 0.836 | 0.919 | 0.738 | 0.818 | 0.914 | 0.182 | 0.661 | 0.793 |
| 2 | ✓ | ✓ | × | × | 2 | 0.949 | 0.948 | 0.850 | 0.944 | 0.745 | 0.833 | 0.949 | 0.274 | 0.747 | 0.860 |
| | ✓ | ✓ | ✓ | × | 1 | 0.954 | 0.952 | 0.860 | 0.946 | 0.763 | 0.844 | 0.953 | 0.292 | 0.761 | **0.872** |
| **3** | ✓ | ✓ | ✓ | × | 2 | **0.954** | **0.952** | **0.864** | 0.943 | **0.774** | **0.850** | **0.953** | 0.295 | 0.758 | 0.872 |
| | ✓ | ✓ | ✓ | × | 3 | 0.952 | 0.951 | 0.861 | 0.944 | 0.768 | 0.847 | 0.952 | 0.286 | 0.759 | 0.870 |
| | ✓ | ✓ | ✓ | × | 4 | 0.951 | 0.949 | 0.841 | 0.952 | 0.719 | 0.819 | 0.951 | 0.274 | 0.749 | 0.865 |
| 4 | ✓ | ✓ | ✓ | ✓ | 2 | 0.954 | 0.952 | 0.843 | **0.957** | 0.719 | 0.821 | 0.953 | **0.301** | **0.766** | 0.861 |

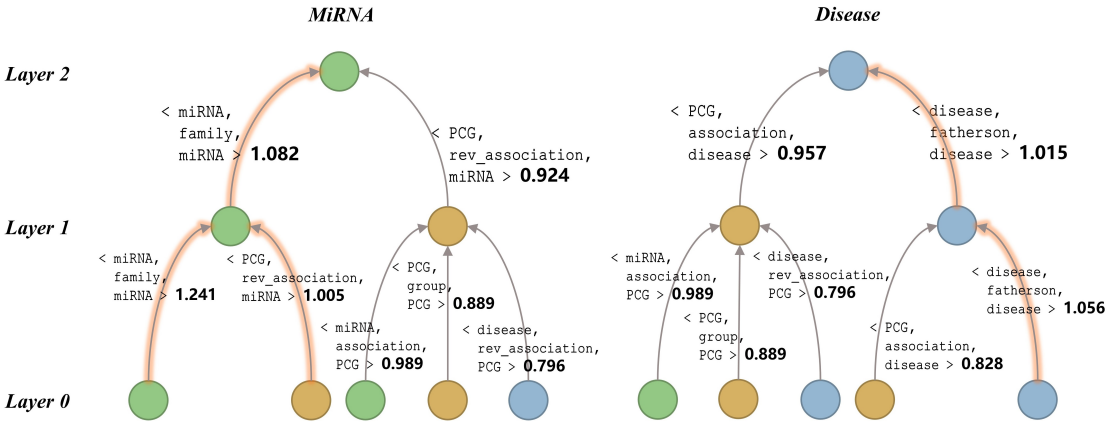



**Fig. 4.** Hierarchy of the learned meta-relations. Important meta-relations with average attention value $\mu^{i}_{\langle\tau(s),\phi(e),\tau(t)\rangle}$ larger than 1 are highlighted.

**Table 5.** MiRNAs associated with LADA in the collected literature

| PMID | Mature Name | Accession | Pred. Score |
|---|---|---|---|
| 36746199 | hsa-miR-146a-5p | MI0000477 | 0.986 |
| | hsa-miR-21-5p, miR-21 | MI0000077 | 0.965 |
| | hsa-miR-223-3p | MI0000300 | 0.935 |
| 27558530 | miR-34a | MI0000268 | 0.836 |
| | miR-24.1 | MI0000080 | 0.596 |
| | miR-30d | MI0000255 | 0.703 |
| 32815005 | hsa-miR-143-3p | MI0000459 | 0.584 |
| 31383887 | hsa-miR-517b-3p | MI0003165 | *0.098* |

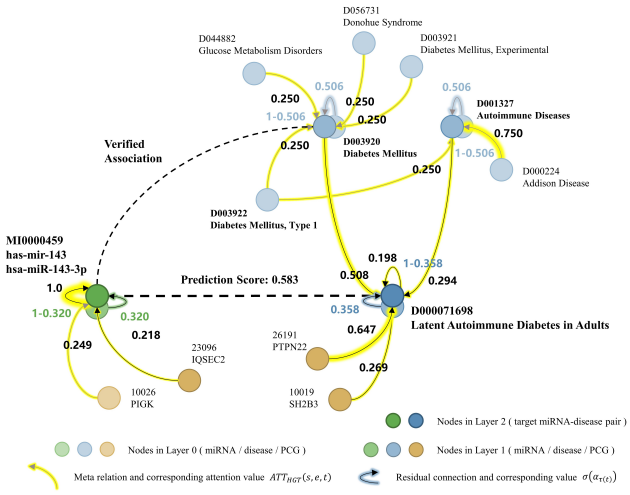



**Fig. 5.** Explanation of the prediction of "hsa-miR-143-3p - LADA".

for assisting biological experiments. It's significant to generalize the effective predictions to entities with fewer or without existing MDAs. Firstly in the stage of data, we construct a miRNA-PCG-disease graph containing all authoritative recorded human miRNAs and diseases, and split the verified MDAs reasonably for better evaluation. Then in the stage of the model, we propose an end-to-end MDA prediction model EGPMDA, stacking a node feature Encoder layer, HGT-based GNN layers, and a Predictor layer sequentially. Finally, in the stage of result analysis, computational experiments show that EGPMDA outperforms state-of-the-art methods on both basic metrics and generalizability. Case studies further demonstrate that our method can detect potential MDAs reliably for unseen diseases, we can explain the overall contribution of input features and the prediction basis of instances.

However, there is still plenty of space for progress. First, the heterogeneous graph can be continuously extended by introducing more biological entities and associations like protein interactions. Then, it's in demand to get trustworthy negative samples rather than random miRNA-disease pairs for better training of models. In addition, future studies can bring more benefits by improving the human-computer interaction experience, such as integrating the data, prediction results, and explanations into a visual analysis system.

**Key Points**

- We integrate multi-source data into a heterogeneous graph with a wider learning and prediction scope, and we split massive verified MDAs into independent training, validation, and test sets as a benchmark.
- We construct an end-to-end data-driven model that avoids taking the existing MDAs as an input feature. It performs node feature encoding, graph structure learning, and prediction sequentially, with a heterogeneous graph transformer as the central module.
- Computational experiments illustrate that our method outperforms existing state-of-the-art methods, it achieves better evaluation metrics and alleviates the neglect of entities with less or without existing MDAs effectively.
- Case studies demonstrate that we can make reliable MDA detections on unseen diseases, and the predictions can be explained in general and case by case.

## Data availability

The data and source codes are available at `https://github.com/EchoChou990919/EGPMDA`

## Funding

This work has been supported by the National Natural Science Foundation of China (Nos.62172289).

## References

1. David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.
2. Victor Ambros. The functions of animal micrornas. *Nature*, 431(7006):350–355, 2004.
3. Chao Lan, Xiaopeng Shi, Nannan Guo, Hui Pei, and Huali Zhang. Value of serum mir-155-5p and mir-133a-3p expression for the diagnosis and prognosis evaluation of sepsis. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*, 28(8):694–698, 2016.
4. Xiaolan Zheng, Yue Zhang, Sha Lin, Yifei Li, Yimin Hua, and Kaiyu Zhou. Diagnostic significance of microrna in sepsis. *Plos one*, 18(2):e0279726, 2023.
5. Liang Yu, Yujia Zheng, Bingyi Ju, Chunyan Ao, and Lin Gao. Research progress of mirna–disease association prediction and comparison of related algorithms. *Briefings in Bioinformatics*, 23(3):bbac066, 2022.
6. Wei Liu, Hui Lin, Li Huang, Li Peng, Ting Tang, Qi Zhao, and Li Yang. Identification of mirna–disease associations via deep forest ensemble learning based on autoencoder. *Briefings in Bioinformatics*, 23(3), 2022.
7. Qiguo Dai, Zhaowei Wang, Ziqiang Liu, Xiaodong Duan, Jinmiao Song, and Maozu Guo. Predicting mirna-disease associations using an ensemble learning framework with resampling method. *Briefings in Bioinformatics*, 23(1):bbab543, 2022.
8. Jin Li, Sai Zhang, Tao Liu, Chenxi Ning, Zhuoxuan Zhang, and Wei Zhou. Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. *Bioinformatics*, 36(8):2538–2546, 2020.
9. Zhengzheng Lou, Zhaoxu Cheng, Hui Li, Zhixia Teng, Yang Liu, and Zhen Tian. Predicting mirna–disease associations via learning multimodal networks and fusing mixed neighborhood information. *Briefings in Bioinformatics*, 23(5), 2022.
10. Huizhe Zhang, Juntao Fang, Yuping Sun, Guobo Xie, Zhiyi Lin, and Guosheng Gu. Predicting mirna-disease associations via node-level attention graph auto-encoder. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.
11. Yuchong Gong, Yanqing Niu, Wen Zhang, and Xiaohong Li. A network embedding-based multiple information integration method for the mirna-disease association prediction. *BMC bioinformatics*, 20:1–13, 2019.
12. Ngan Dong, Stefanie Mücke, and Megha Khosla. Mucomid: A multitask graph convolutional learning framework for mirna-disease association prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3081–3092, 2022.
13. Thi Ngan Dong, Johanna Schrader, Stefanie Mücke, and Megha Khosla. A message passing framework with multiple data integration for mirna-disease association prediction. *Scientific Reports*, 12(1):16259, 2022.
14. Liang Yu, Yujia Zheng, and Lin Gao. Mirna–disease association prediction based on meta-paths. *Briefings in Bioinformatics*, 23(2), 2022.
15. Wei Peng, Zicheng Che, Wei Dai, Shoulin Wei, and Wei Lan. Predicting mirna-disease associations from mirna-gene-disease heterogeneous network with multi-relational graph convolutional network model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.
16. Xinru Tang, Jiawei Luo, Cong Shen, and Zihan Lai. Multi-view multichannel attention graph convolutional network for mirna–disease association prediction. *Briefings in Bioinformatics*, 22(6):bbab174, 2021.
17. Cheng Yan, Guihua Duan, Na Li, Lishen Zhang, Fang-Xiang Wu, and Jianxin Wang. Pdmda: predicting deep-level mirna–disease associations with graph neural networks and sequence features. *Bioinformatics*, 38(8):2226–2234, 2022.
18. Wenxiang Zhang, Hang Wei, and Bin Liu. idenmd-nrf: a ranking framework for mirna-disease association identification. *Briefings in Bioinformatics*, 23(4), 2022.
19. Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, and Qinghua Cui. Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. *Nucleic acids research*, 47(D1):D1013–D1017, 2019.
20. Xing Chen, Di Xie, Qi Zhao, and Zhu-Hong You. Micrornas and complex diseases: from experimental results to computational models. *Briefings in bioinformatics*, 20(2):515–539, 2019.
21. Xiujuan Lei, Thosini Bamunu Mudiyanselage, Yuchen Zhang, Chen Bian, Wei Lan, Ning Yu, and Yi Pan. A comprehensive survey on computational methods of non-coding rna and disease association prediction. *Briefings in bioinformatics*, 22(4):bbaa350, 2021.
22. James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.
23. Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

24. Twan Van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.
25. Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics*, 26(13):1644–1650, 2010.
26. Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
27. Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234, 2016.
28. Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
29. Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.
30. Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019.
31. Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, 42(D1):D92–D97, 2014.
32. Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.
33. Jia Chen, Jiahao Lin, Yongfei Hu, Meijun Ye, Linhui Yao, Le Wu, Wenhai Zhang, Meiyi Wang, Tingting Deng, Feng Guo, et al. Rnadisease v4. 0: an updated resource of rna-associated diseases, providing rna-disease analysis, enrichment and prediction. *Nucleic Acids Research*, 51(D1):D1397–D1404, 2023.
34. Chan Yeong Kim, Seungbyn Baek, Junha Cha, Sunmo Yang, Eiru Kim, Edward M Marcotte, Traver Hart, and Insuk Lee. Humannet v3: an improved database of human gene networks for disease research. *Nucleic acids research*, 50(D1):D632–D639, 2022.
35. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
36. Jiaxuan You, Zhitao Ying, and Jure Leskovec. Design space for graph neural networks. *Advances in Neural Information Processing Systems*, 33:17009–17021, 2020.
37. Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.
38. Silvia Pieralice and Paolo Pozzilli. Latent autoimmune diabetes in adults: a review on clinical implications and management. *Diabetes & Metabolism Journal*, 42(6):451, 2018.
39. Wenqi Fan, Haipeng Pang, Xia Li, Zhiguo Xie, Gan Huang, and Zhiguang Zhou. Plasma-derived exosomal mirnas as potentially novel biomarkers for latent autoimmune diabetes in adults. *Diabetes Research and Clinical Practice*, 197:110570, 2023.
40. Attila A Seyhan, Yury O Nunez Lopez, Hui Xie, Fanchao Yi, Clayton Mathews, Magdalena Pasarica, and Richard E Pratley. Pancreas-enriched mirnas are altered in the circulation of subjects with diabetes: a pilot cross-sectional study. *Scientific reports*, 6(1):31479, 2016.
41. Shan Pan, Mengyu Li, Haibo Yu, Zhiguo Xie, Xia Li, Xianlan Duan, Gan Huang, and Zhiguang Zhou. microrna-143-3p contributes to inflammatory reactions by targeting fosl2 in pbmcs from patients with autoimmune diabetes mellitus. *Acta Diabetologica*, 58:63–72, 2021.
42. Ke Yu, Zhou Huang, Jing Zhou, Jianan Lang, Yan Wang, Xingqi Yin, Yuan Zhou, and Dong Zhao. Transcriptome profiling of micrornas associated with latent autoimmune diabetes in adults (lada). *Scientific Reports*, 9(1):1–9, 2019.
43. Marc De Braekeleer, Minh Huong Nguyen, Frédéric Morel, and Aurore Perrin. Genetic aspects of monomorphic teratozoospermia: a review. *Journal of assisted reproduction and genetics*, 32:615–623, 2015.
44. Maja Tomic, Luka Bolha, Joze Pizem, Helena Ban-Frangez, Eda Vrtacnik-Bokal, and Martin Stimpfel. Association between sperm morphology and altered sperm microrna expression. *Biology*, 11(11):1671, 2022.
45. Delnya Gholami, Farzane Amirmahani, Reza Salman Yazdi, Tahereh Hasheminia, and Hossein Teimori. Mir-182-5p, mir-192-5p, and mir-493-5p constitute a regulatory network with crisp3 in seminal plasma fluid of teratozoospermia patients. *Reproductive Sciences*, 28:2060–2069, 2021.
46. Ling-Yu Yeh, Robert Kuo-Kuang Lee, Ming-Huei Lin, Chih-Hung Huang, and Sheng-Hsiang Li. Correlation between sperm micro ribonucleic acid-34b and-34c levels and clinical outcomes of intracytoplasmic sperm injection in men with male factor infertility. *International Journal of Molecular Sciences*, 23(20):12381, 2022.
47. Delnya Gholami, Reza Salman Yazdi, Mohammad-Saeid Jami, Sorayya Ghasemi, Mohammad-Ali Sadighi Gilani, Shaghayegh Sadeghinia, and Hossien Teimori. The expression of cysteine-rich secretory protein 2 (crisp2) and mir-582-5p in seminal plasma fluid and spermatozoa of infertile men. *Gene*, 730:144261, 2020.

**Yi Zhou** is a master's student at the College of Computer Science, Sichuan University. Her research interests include data-centric AI, graph data mining, and bioinformatics.

**Meixuan Wu** is a PhD student at the College of Computer Science, Sichuan University. Her research interests include machine learning and bioinformatics.

**Chengzhou Ouyang** is an undergraduate student at the College of Life Sciences, Sichuan University. His research interests include system biology and bioinformatics.

**Xinyi Wang** is a master's student at the College of Computer Science, Sichuan University. His research interests include deep learning and bioinformatics.

**Min Zhu** is a professor at the College of Computer Science, Sichuan University. Her research interests include bioinformatics, visual analysis, and image processing.

# Supplementary File

## SF.1 Data Observation: Common Neighbor Statistics

For miRNA-disease pairs with or without verified association, we make an observation of the surrounding graph structure via common neighbor statistics. Taking miRNA-disease pairs with verified associations and equal randomly sampled pairs into account, considering all MDAs as edges momentarily, we calculate the proportion of common neighbors formulated as

$$CM^t(m,d) = \begin{cases} \frac{|N_m^t \cap N_d^t|}{|N_m^t \cup N_d^t|}, & |N_m^t \cup N_d^t| > 0 \\ \text{Not a Number}, & |N_m^t \cup N_d^t| = 0 \end{cases}, \quad (7)$$

where $N_m^t$ and $N_d^t$ denote the $t$-type neighbor nodes of miRNAs and diseases, and $t$ can be miRNA, disease, PCG, or all unified.

Figure 6 shows the distribution of the proportion, and we can make the following observations: 1) MiRNA-disease pairs with verified associations share more common neighbors all over the three neighbor node types. Our multi-source data integration should be helpful, the 1-hop subgraph structures (along various inter- and intra-class edges) differ a lot. 2) On the whole, the proportion of common neighbors is very low, observation 1 is a comparison between the few and the hardly any. Critical information hides in noises.

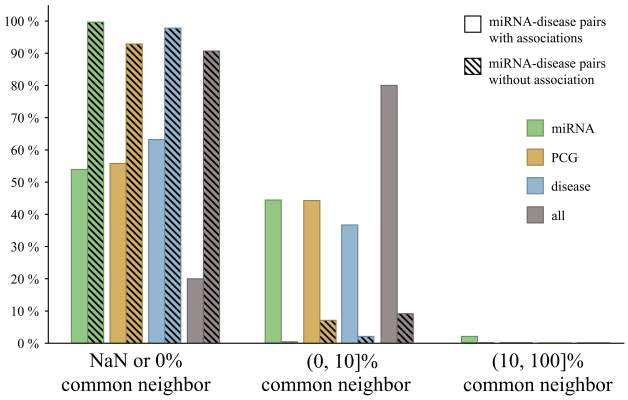


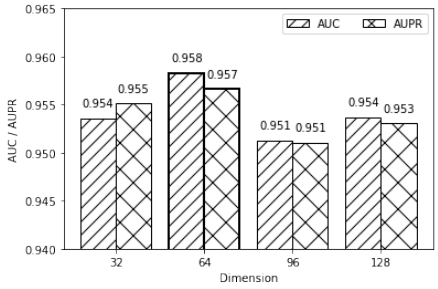

**Fig. 6.** Statistic of common neighbors for miRNA-disease pairs. For miRNA-disease pairs with verified associations, there are 53.9%, 55.7%, 63.2%, and 20.0% of them do not have any common miRNA, PCG, disease, or overall neighbors, while for pairs without verified association, it's 99.6%, 92.9%, 97.8%, and 90.7%. And when there are common neighbors, for all neighbor node types and for both samples with or without verified association, the proportion of common neighbors is mostly lower than 10%.

## SF.2 Model Training and Hyperparameter Selection

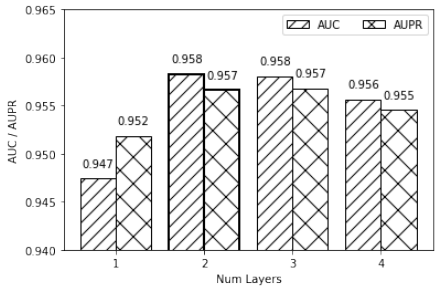
All the computational experiments are conducted on an NVIDIA RTX 3090 GPU with 24GB memory. We adopt the Adam optimizer with a learning rate of 0.001 and set an early stop with a maximum of 50 epochs. In selecting hyperparameters, the model is trained on the training set and evaluated on the validation set, returning the weights that achieve the highest validation accuracy with the patience set to five. After fixing all hyperparameters, the model is trained on the union of training and validation sets to utilize more supervision information, and such training process is stopped when the loss does not drop twice in a row.

There are three main hyperparameters: the dimension of intermediate vectors $d$, the number of GNN layers $L$, and the number of multi-heads $h$. A grid search is conducted where $d \in \{32, 64, 96, 128\}$, $L \in \{1, 2, 3, 4\}$ and $h \in \{1, 2, 4, 8\}$. In the histograms Fig. 7, the x-axis represents one unfixed hyperparameter and the y-axis shows the average AUC and AUPR on the validation set.

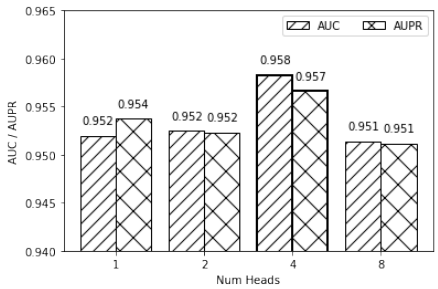
Eventually, we choose $d = 64$, $L = 2$ and $h = 4$.




(a) Dimension $d$




(b) Number of Layers $L$




(c) Number of Heads $h$

**Fig. 7.** Hyperparameter selection.

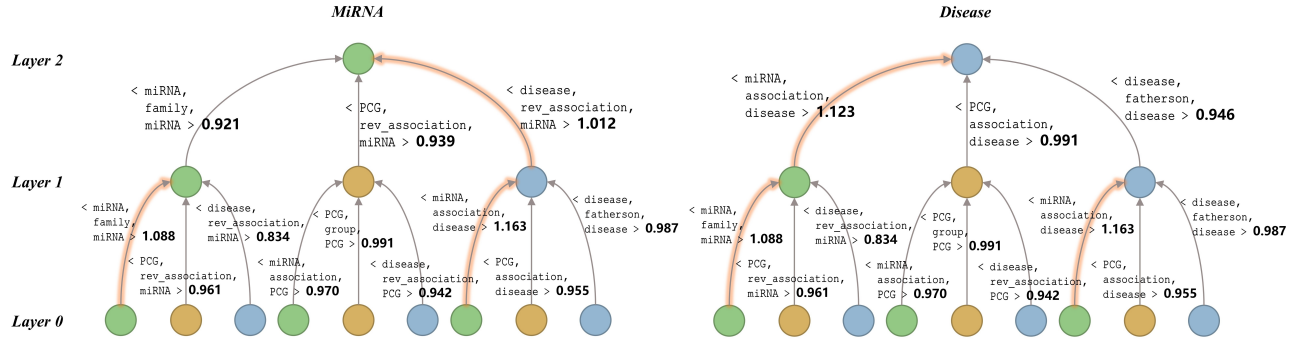



**Fig. 8.** Hierarchy of the learned meta-relations while taking existing MDAs into message passing. For each meta-relation and each layer, we calculate the average attention value $\mu^{i}_{(\tau(a),\phi(e),\tau(t))}$ of four heads and five repeats. Important meta-relations with attention scores larger than 1 are highlighted.

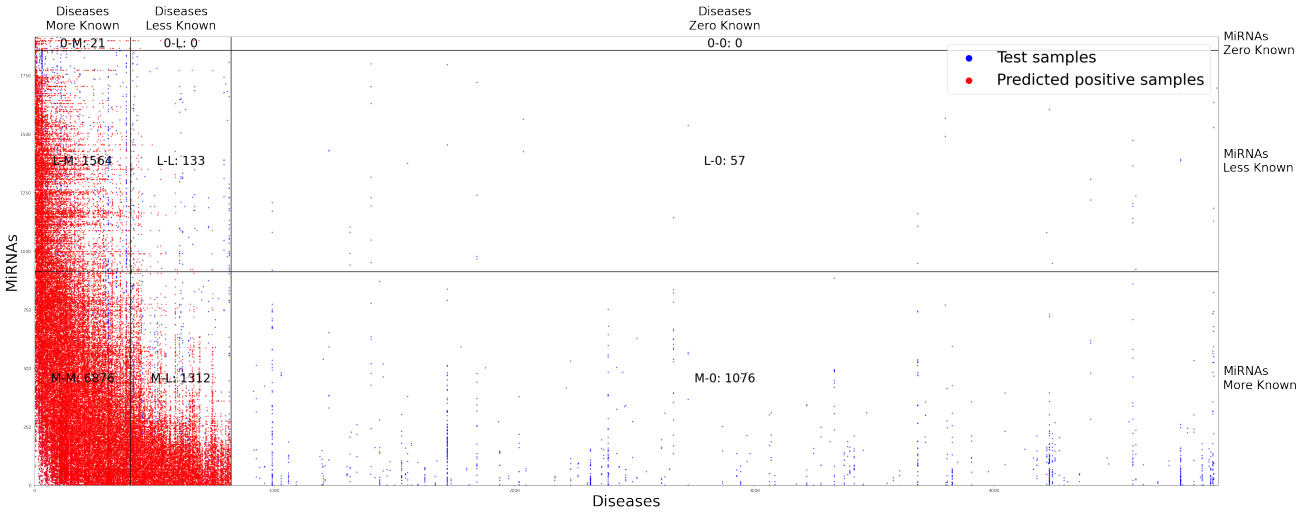



**Fig. 9.** Top@5% predictions from EGPMDA (+MDAs), $\text{Recall}_{@5\%} = 0.771$.

## SF.3 Ablation Study: Case 4

As shown in Fig 8, once taking existing MDAs into message passing, the model focuses more on **<miRNA, association, disease, rev_association, miRNA>** and **<miRNA, family, miRNA, association, disease>**. Since relying heavily on miRNA-disease edges, the model almost loses the ability to make successful MDA detections for less and zero known miRNAs and diseases. Figure 9 shows the Top@5% predictions under this case, obviously all top-ranked predictions are undesirably distributed in the left bottom regions.

## SF.4 Case Study

### *SF.4.1 Teratozoospermia*

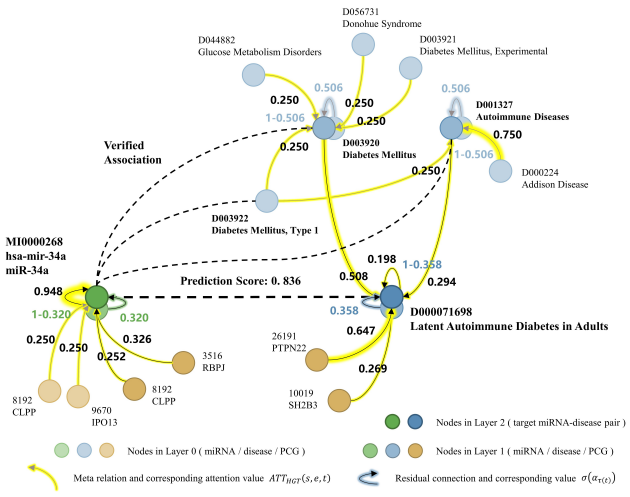
Teratozoospermia (MeSH ID: D000072660) is a type of male infertility, it is characterized by the presence of spermatozoa with abnormal morphology over 85% in sperm [43]. There are miRNAs associated with Teratozoospermia, and we've obtained the following descriptions. Table 6 shows the prediction results of these MDAs, and twelve of the fourteen are successfully detected:

**Table 6.** MiRNAs associated with Teratozoospermia in the collected literature

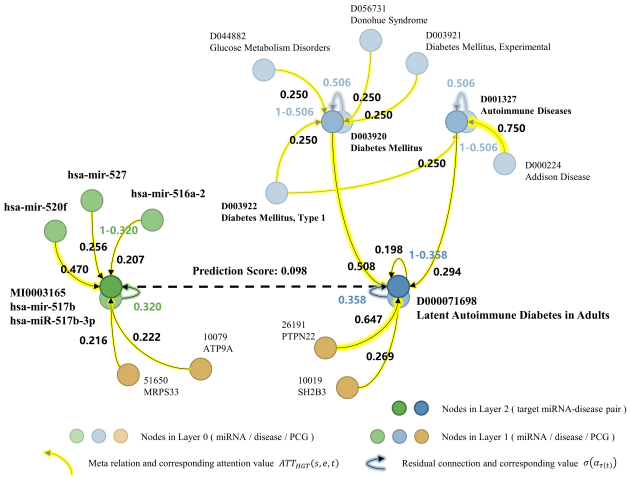
| PMID | Mature Name | Accession | Pred. Score |
|---|---|---|---|
| 36421385 | miR-10a-5p | MI0000266 | 0.715 |
| | miR-15b-5p | MI0000438 | 0.949 |
| | miR-26a-5p | MI0000083 | 0.824 |
| | miR-34b-3p, miR-34b | MI0000742 | 0.586 |
| | miR-122-5p | MI0000442 | 0.880 |
| | miR-125b-5p | MI0000446 | 0.919 |
| | miR-191-5p | MI0000465 | 0.502 |
| | miR-296-5p | MI0000747 | 0.585 |
| | let-7a-5p | MI0000062 | 0.833 |
| 33620707 | miR-182-5p | MI0000272 | 0.691 |
| | miR-192-5p | MI0000234 | 0.898 |
| | miR-493-5p | MI0003132 | *0.412* |
| 36293237 | miR-34c | MI0000743 | 0.625 |
| 31778754 | miR-582-5p | MI0003589 | *0.350* |

- Study [44]: "we determined significant under-expression of nine miRNAs (miR-10a-5p/-15b-5p/-26a-5p/-34b-3p/-122-5p/-125b-5p/-191-5p/-296-5p and let-7a-5p) in spermatozoa from patients with teratozoospermia compared to the controls."
- Study [45]: "miR-182-5p, miR-192-5p, and miR-493-5p constitute a regulatory network with CRISP3 in seminal plasma fluid of teratozoospermia patients."
- Study [46]: "miR-34b and miR-34c were significantly associated with intracytoplasmic sperm injection (ICSI) clinical outcomes in male factor infertility, especially teratozoospermia."
- study [47]: "miR-582-5p expression significantly increased in teratozoospermia patients."

### *SF.4.2 Instance-level Explanations*




(a) "miR-34a - LADA", Prediction Score = 0.836




(b) "hsa-miR-517b-3p - LADA", Prediction Score = 0.098

**Fig. 10.** Explanations of the predictions of "miR-34a - LADA" and "hsa-miR-517b-3p - LADA". We extract the $\mathbf{ATT}_{\text{HGT}}(s,e,t)$ and calculate the average four heads. The obtained attention and residual values are labeled on the corresponding edges.