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

Yi Zhou[a], Meixuan Wu[a], Chengzhou Ouyang[b] and Min Zhu[a]

[a]College of Computer Science, Sichuan University, Chengdu, 610065, China
[b]College of Life Sciences, Sichuan University, Chengdu, 610065, China

ARTICLE INFO

Keywords:
miRNA-disease association
heterogeneous graph neural networks
explainable AI

ABSTRACT

Biomedical research has 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). However, the generalizability and explainability are currently underemphasized. It's significant to generalize effective predictions to entities with fewer or no existing MDAs and reveal how the prediction scores are derived. In this study, our work contributes to data, model, and result analysis. First, for better formulation of the MDA issue, we integrate multi-source data into a heterogeneous graph with a broader learning and prediction scope, and we split massive verified MDAs into independent training, validation, and test sets as a benchmark. Second, we construct an end-to-end data-driven model that performs node feature encoding, graph structure learning, and binary prediction sequentially, with a heterogeneous graph transformer as the central module. Finally, computational experiments illustrate that our method outperforms existing state-of-the-art methods, achieving better evaluation metrics and alleviating the neglect of unknown miRNAs and diseases effectively. Case studies further demonstrate that we can make reliable MDA detections on diseases without MDA records, and the predictions can be explained in general and case by case.

1. 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, influencing the output of many protein-coding genes (PCGs) [1, 2]. Due to their ability to disrupt the normal functioning of PCGs, miRNAs can participate in the process of disease onset and progression. Consequently, the identification of miRNA-disease associations (MDAs) holds significant value for the diagnosis and treatment of diseases. For instance, the mRNA expressions of serum miR-155-5p and miR-133a-3p were gradually increased with the aggravation of sepsis [3]. And meta-analysis revealed that miRNAs, specifically miR-155-5p, could be useful biomarkers for detecting sepsis, a clinical serum specimen is also indicated for diagnostic purposes [4].

To date, numerous MDAs have been verified through biological experiments. However, the advancement of such experiments is hindered by their high cost and time-consuming nature. Consequently, computational prediction methods are increasingly proposed to detect potential MDAs for assistance. In this regard, we summarize the research of MDA predictions into three main stages: (i) Data organization to adequately describe miRNA-disease associations. (ii) Development of applicable models for accurate MDA predictions, and (iii) Analysis and explanation of prediction results.

In stage (i), the most commonly used input features include the existing MDAs and disease father-son relations, presented as miRNA-disease adjacency matrix and directed acyclic graph of diseases, respectively [5, 6, 7, 8, 9, 10]. Additionally, studies [11, 12, 13] transform miRNA families into associations since miRNAs belonging to the same family usually have highly similar sequence secondary structures and tend to execute similar biological functions. To capture the regulatory process of miRNAs on diseases through PCGs, studies [14, 15, 16, 12, 13] incorporate miRNA-PCG and PCG-disease associations to further describe miRNAs and diseases via PCG-mediated pathways. However, when integrating these inter- and intra-class associations into a unified miRNA-PCG-disease graph, node features with biomedical semantics are always overlooked. Although some studies [16, 17, 18] have considered miRNA sequences, they are processed separately and are not integrated into the heterogeneous graph.

In addition to considering the comprehensiveness of input feature categories, the available scope of learning and predictions matters more. It is crucial to organize heterogeneous data in an extensible manner that allows for the inclusion of growing entities and associations, which is an aspect that has been previously underemphasized. Most of the existing studies focus on making predictions within the entities listed in a specific database, such as HMDD [19], which curates experiment-supported evidence for human MDAs. But the rest greater proportion of the human miRNAs and diseases are excluded. As a prerequisite stage, it is essential to construct a benchmark dataset that covers a broader range of miRNAs and diseases, contains more MDAs, and is suitable for the evaluation of both the basic performance and generalizability of models. Few studies [12, 13] allow for predictions on a larger scale and employ subsets created through set operations between two versions of HMDD for training and evaluating models. However, further improvements can be made by including all authoritatively recorded human miRNAs and diseases and merging

*Corresponding author
zhumin@scu.edu.cn (M. Zhu)
ORCID(s):

additional verified MDAs from different databases into the dataset.

In stage (ii), mainstream computational models conduct secondary feature extraction first [20, 21, 5]. Various similarity measures are employed to extract key information. For instance, disease semantic similarity [22] captures the father-son relations between 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. Additionally, embedding algorithms like Node2Vec [26, 18] and SDNE [27, 13] are utilized to extract the surrounding topologies of nodes. These carefully designed heuristic similarities and network embeddings have been shown to be beneficial for downstream link predictions. However, the limitations of handcrafted or shallow encoding extractions, in terms of adaptability and expressiveness, can become a bottleneck for MDA predictions when aiming for higher goals, such as improved generalizability.

With the extracted secondary features presenting key information, machine learning algorithms are employed for MDA predictions. NEMII [11] and DFELMDA [6] use random forest for classification. SMAP [7] conducts matrix factorization with a specific goal. SRJP [28] performs sparse regularized joint projection to derive a prediction matrix.

In particular, with the inspiring development in recent years, graph neural networks (GNNs) are increasingly exploited in MDA predictions which inherently exhibit a graph-like structure, to further enhance feature representations prior to the final classification. NIMCGCN [8] utilizes similarity measures to construct similarity networks and applies graph convolutional networks (GCNs) [29] to extract latent representations. MMGCN [16] utilizes GCNs as encoders on multiple similarity networks and combines them by using a multichannel attention mechanism. MINIMDA [9] employs modified GCNs on multimodal networks consisting of similarities and associations. AGAEMD [10] constructs a node-level attention graph autoencoder on a miRNA-disease bipartite graph, where the encoder includes graph attention networks and an LSTM-based jumping knowledge module. To address both the necessity of inputting node features into GNNs and the actuality of missing biomedical semantic node features, the above studies perform graph augmentation by filling node features with randomly initialed vectors or the corresponding slices of the concatenated adjacency and similarity matrix.

Message-passing GNNs pass and aggregate information from neighboring nodes, thereby learning from the graph structure [30]. Therefore, it's convoluted to conclude graph structures into similarities, newly construct similarity graphs, and subsequently apply GNNs that essentially learn from graph structures. Such a prediction process heavily relies on the presence of existing MDAs, leading to the ignorance of entities with fewer or no existing MDAs. Additionally, explaining the predictions becomes more challenging.

Consequently, there is a pressing need for stage (iii): identifying which features significantly contribute to the prediction performance and revealing the prediction basis for each miRNA-disease pair. As the predictions generalize to unknown regions, there is a greater demand for explanations on a case-by-case basis, beyond prediction scores alone. Current studies are inadequate to make prediction results explainable.

To address the aforementioned shortcomings, (i) we construct a comprehensive dataset comprising all authoritatively recorded human miRNAs and diseases. The relevant information is organized into a heterogeneous graph, incorporating biomedical semantic node features. (ii) We propose a data-driven model that sufficiently learns from the heterogeneous graph, which does not perform secondary feature extraction and avoids taking the existing MDAs as an input feature. And (iii) we present thorough analysis by metrics comparisons, visualizations, and case studies to demonstrate the effectiveness of our method. Our main contributions can be summarized as follows:

- We integrate a miRNA-PCG-disease graph from multi-source databases through a reliability-guaranteed and extension-friendly process. A massive MDA database is split by time as a benchmark.
- We propose an intuitive model EGPMDA that end-to-end performs MDA predictions through 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 unknown nodes. Case studies further illustrate that our model can make reliable MDA detections on diseases without MDA records, which is instance-level explainable.

2. Materials and Data Processing

2.1. Dataset Construction

To better formulate the MDA prediction problem, we integrate abundant 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. Additionally, to facilitate the evaluation of computational models, we split the experimentally verified MDAs into training, validation, and test sets based on the publication year of literature evidence.

Table 1 provides an overview of the data volumes and respective sources. To the best of our knowledge, it should be the largest dataset in existing MDA prediction studies. The construction process is presented in the following steps. For more details, please visit the data and codes at https://github.com/EchoChou990919/EGPMDA.

Step 1. Determine standard nodes. Based on the authoritative databases miRBase [31, 32], MeSH[1], and HGNC[2], we

[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**

|sth.| denotes the count

include all miRNAs, diseases, and PCGs of homo sapiens in our dataset as standard nodes. 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 with an in-depth understanding of the source data, serving for the subsequent obtainment of edges.

Step 2. Obtain biomedical semantic node features. The node features consist of the stem-loop and mature sequences for miRNA, the name and scope note for disease, and the name and belonging group name for PCG. Hence, the node features are represented by RNA sequences for miRNAs and text for diseases and PCGs.

Step 3. Obtain intra-class edges. MiRNA-miRNA edges are derived from miRNA families, wherein miRNAs belonging to the same family are fully connected. Disease-disease edges present the semantic hierarchical (father–son) relations among diseases. Similar to miRNAs, PCGs within the same group are linked to each other, resulting in PCG-PCG edges.

Step 4. Obtain inter-class edges. MiRNA-PCG associations are downloaded from ENCORI[3] [33], presenting experimental results of the degradome between miRNAs and mRNAs (transcripted by PCGs). Disease-PCG associations are acquired from DisGeNet [34], and we only preserve the ones with the evidence-level score ≥ 0.1. Through direct primary ID matching or indirect "aliases matching - primary ID transfer", both ends of these associations are aligned to the standard nodes, and inter-class edges are established after de-duplication.

Step 5. Obtain miRNA-disease associations. MDAs are derived from RNADisease [35], a comprehensive database that incorporates manual curation of numerous literature and other experimentally verified databases, including HMDD [19] and dbDEMC [36] that are widely used in other MDA prediction studies. To pursue the reliability of the associations, records without available PMID are filtered out, ensuring that each MDA is supported by at least one piece of literature evidence. Similarly, these verified MDAs are aligned to standard miRNA and disease nodes. In addition, the publication year of the literature corresponding to each PMID is acquired by using the "Bio.Entrez" package [4].

Step 6. Split MDA samples by time. We use 39991 MDAs 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%).

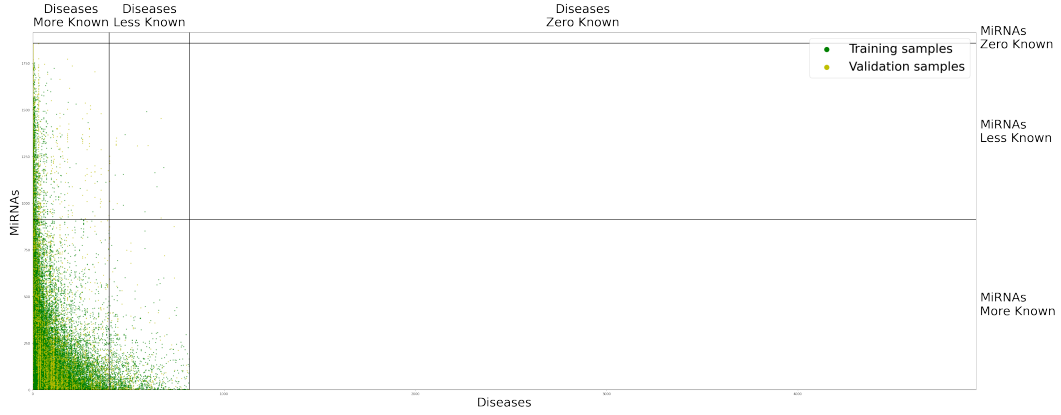
Eventually, we construct a dataset that encompasses the following information: miRNA sequences, disease description texts, gene name texts, miRNA families, disease father-son relations, gene groups, miRNA-gene associations, disease-gene associations, and verified miRNA-disease associations. All information is integrated into a heterogeneous graph, where the nodes represent 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 substituted with the gene functional networks sourced from HumanNet [37]. And the nearest verified MDAs are used to evaluate the effectiveness of MDA predictions.

2.2. Data Observation

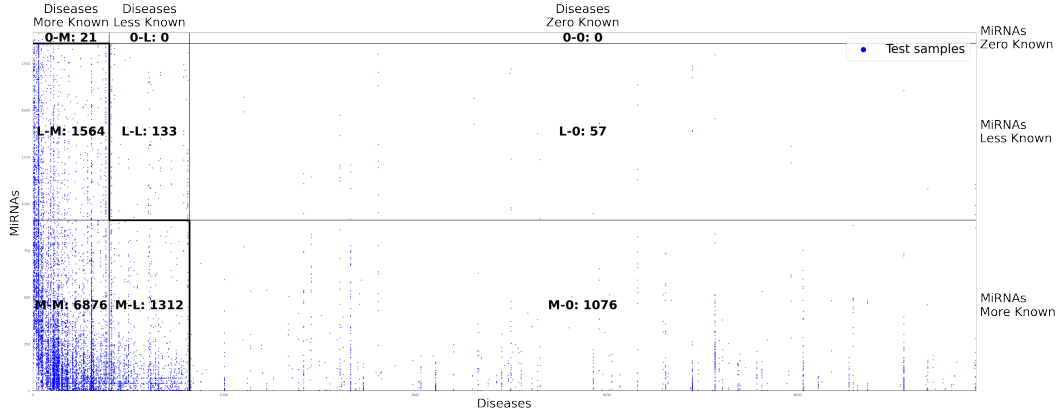
Visualizations of the miRNA-disease adjacency matrix (Figure 1) show the distribution of MDA samples. Verified MDAs only occupy a small proportion compared to the unexplored region, highlighting the importance of generalizable predictions. Additionally, there are distribution shifts observed among the training, validation, and test samples. The 11039 test samples exhibit a slightly more dispersed distribution, particularly towards the less known and zero known nodes. It can be attributed to the progressive nature of biomedical experimental research. Verified MDAs gradually "spread" from traditionally more known entities to others over time, reflecting evolving research interests and discoveries. Therefore, our time-based dataset split is reasonable and beneficial for evaluating both the basic performance and generalizability of models. In particular, analyzing successful MDA detections on the almost blank subsets can provide valuable insights into the effectiveness of prediction methods in challenging real-world scenarios.

[3] https://rna.sysu.edu.cn/encori/

[4] https://biopython.org/



(a) Training and validation samples



(b) Test samples

Figure 1: The miRNA-disease adjacency matrix. (a) The occurrence number of miRNAs or diseases in the training and validation sets can be considered as the degree of "known". The median degree is 8 for miRNAs and 10 for diseases. Sorting miRNAs and diseases based on their known degree, the adjacency matrix can be divided into nine regions by considering the combinations of zero (0), less (0 to median), and more (median to max) known miRNAs and diseases. (b) Meanwhile, the test set is split into corresponding subsets for further evaluation, with the sample sizes indicated in the figure. The three in the bottom left (*L-M*, *M-L* and *M-M*) are sparse ($sparsity \approx 0.866\%$), and the remaining four with non-zero sample size (*0-M*, *L-0*, *M-0* and *L-L*) are almost blank ($sparsity \approx 0.016\%$).

Moreover, we analyze the graph structures around miRNA-disease pairs via common neighbor statistics. Taking miRNA-disease pairs with verified associations and an equal number of randomly sampled pairs into account, and considering all MDAs as edges momentarily, we calculate the proportion of common neighbors formulated as

$$CM^{\tau}(m,d) = \begin{cases} \frac{|N_m^{\tau} \cap N_d^{\tau}|}{|N_m^{\tau} \cup N_d^{\tau}|}, & |N_m^{\tau} \cup N_d^{\tau}| > 0 \\ \text{NaN}, & |N_m^{\tau} \cup N_d^{\tau}| = 0 \end{cases},$$

where N_m^{τ} and N_d^{τ} denote the τ-type neighboring nodes of miRNAs and diseases, and τ can be miRNA, disease, PCG, or all without distinction.

Figure 2 shows the distribution of $CM^{\tau}(m,d)$, and we can make the following observations: 1) miRNA-disease pairs with verified associations exhibit a higher proportion of common neighbors across all three classes of neighboring nodes. The 1-hop subgraph structures formed by inter- and intra-class edges vary a lot. It suggests that our multi-source data integration should be significant. 2) The overall proportion of common neighbors is relatively low—observation 1 highlights a comparison between a small number of common neighbors and an even smaller number. It emphasizes the challenge posed by the scarcity of critical information.

These observations have driven our model design, highlighting the inherent need to learn the surrounding graph structures. Therefore, we can utilize Graph Neural Networks. However, critical information can be obscured by less relevant or noisy information. It requires the GNNs to estimate critical neighboring nodes accurately, and one effective approach is to incorporate attention mechanisms within GNNs to emphasize the aggregation of important messages with higher weights.

2.3. 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

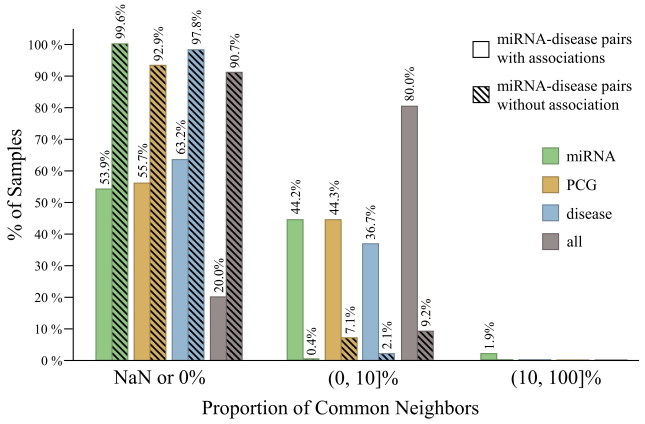


Figure 2: Common neighbor statistic for miRNA-disease pairs. 1) For miRNA-disease pairs with verified associations, it is observed that 53.9%, 55.7%, 63.2%, and 20.0% of them do not share any common miRNA, PCG, disease, or overall neighbors, respectively. On the other hand, for miRNA-disease pairs without verified associations, the corresponding percentages are 99.6%, 92.9%, 97.8%, and 90.7%. 2) When common neighbors are present, the proportion of common neighbors is predominantly lower than 10% for all types of neighboring nodes, regardless of whether the miRNA-disease pairs have verified associations or not.

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 model learning, the node features are embedded into vectors and the edges are preprocessed into directed by the following:

RNA sequences are composed of four natural bases: 'A' (adenine), 'U' (uracil), 'C' (cytosine), and 'G' (guanine). To represent an RNA sequence as a vector, we employ the 1-mer representation method. Firstly, the sequence is normalized to a predetermined maximum length by padding the end with a placeholder base 'N'. Along the sequence, 'A', 'U', 'C' and 'G' are mapped to one-hot vectors representing 0, 1, 2, and 3, respectively. The placeholder base 'N' is mapped to a vector $[0.25, 0.25, 0.25, 0.25]^T$.

For each miRNA, there is one stem-loop sequence and one or two mature sequences. And we perform the aforementioned mapping separately for each sequence type and then concatenate the resulting vectors. Thus, we obtain a vector $x_m \in \mathbb{R}^{(l_s + l_{m_1} + l_{m_2}) \times 4}$, where l_s, l_{m_1} and l_{m_2} denote the maximum lengths of the stem-loop sequence, the first mature sequence, and the second mature sequence of all miRNAs, respectively.

BioBERT [38] is a specialized language representation model designed for the biomedical domain and trained on a large collection of biomedical corpora. By utilizing BioBERT, we can generate contextualized embeddings for texts related to biomedical concepts. In the case of diseases, we employ BioBERT to process the disease name and scope note, resulting in concatenated vectors denoted as $x_d \in \mathbb{R}^{2d_B}$, where d_B represents the default embedding size. Similarly, for each PCG, we obtain $x_g \in \mathbb{R}^{2d_B}$ from the PCG name and the belonging group name.

To ensure effective message passing in the GNNs, we introduce "reverse" connections for all edges in the graph and include self-loops. In conclusion, the node types $\tau(v) \in \{\texttt{miRNA}, \texttt{disease}, \texttt{PCG}\}$, and the meta-relations $\langle \tau(s), \phi(e), \tau(t) \rangle \in \{<\texttt{miRNA}, \texttt{family}, \texttt{miRNA}>, <\texttt{disease}, \texttt{father-son}, \texttt{disease}>, <\texttt{PCG}, \texttt{group}, \texttt{PCG}>, <\texttt{miRNA}, \texttt{association}, \texttt{PCG}>, <\texttt{PCG}, \texttt{rev\_association}, \texttt{miRNA}>, <\texttt{PCG}, \texttt{association}, \texttt{disease}>, <\texttt{disease}, \texttt{rev\_association}, \texttt{PCG}>\}$.

3. Methods

In this work, we define the MDA prediction as a binary link prediction towards miRNA-disease pairs within the miRNA-PCG-disease graph. Figure 3 illustrates our method, which is based on the design space of GNNs [39]. Our proposed model, referred to as **EGP**MDA, consists of three main modules: an **E**ncoder layer for learning biomedical semantics from node features, a stack of **G**raph neural network layers for capturing heterogeneous graph structures, and a **P**redictor layer for generating predictions.

We denote the output of (l)-th layer as $H^{(l)}$, where $l = 0$ corresponds to the encoder layer, and $l \in [1, L]$ corresponds to the GNN layers. The output of the current layer serves as the input to the next layer, and this process continues until the predictor layer generates the final prediction. To simplify notation, we use sth.-Linear(x) to denote a linear transformation $xW + b$, where x is the input vector, W is a parameterized weight matrix, and b is a learnable bias.

3.1. Node Feature Encoder Layer

As formulated by equation 1, the first layer encodes the feature vectors of nodes:

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

For miRNA, a convolution operation is applied to the 1-mer embedded miRNA sequences x_m, there is a single filter of shape $k \times 4$ with a stride of 1 and no padding. The output of the convolution is then linearly transformed to $\mathbb{R}^{dim}$. Similarly, for disease and PCG, parallel linear transformations are applied to project x_d and x_g into $\mathbb{R}^{dim}$. Here, dim represents a hyperparameter determining the dimensionality of the hidden layers.

3.2. Graph Neural Network Layers

The stacked L GNN layers can capture the subgraph structure within a maximum L-hop neighborhood of miRNA-disease pairs.

Based on the aforementioned data observations, we first introduce the general attention-based message-passing GNNs [30, 40] briefly with the following equation:

$$H^{(l)}[t] \leftarrow \mathbf{Agg}\left(\mathbf{ATT}(s, t) \cdot \mathbf{MSG}(s)\right), \tag{2}$$

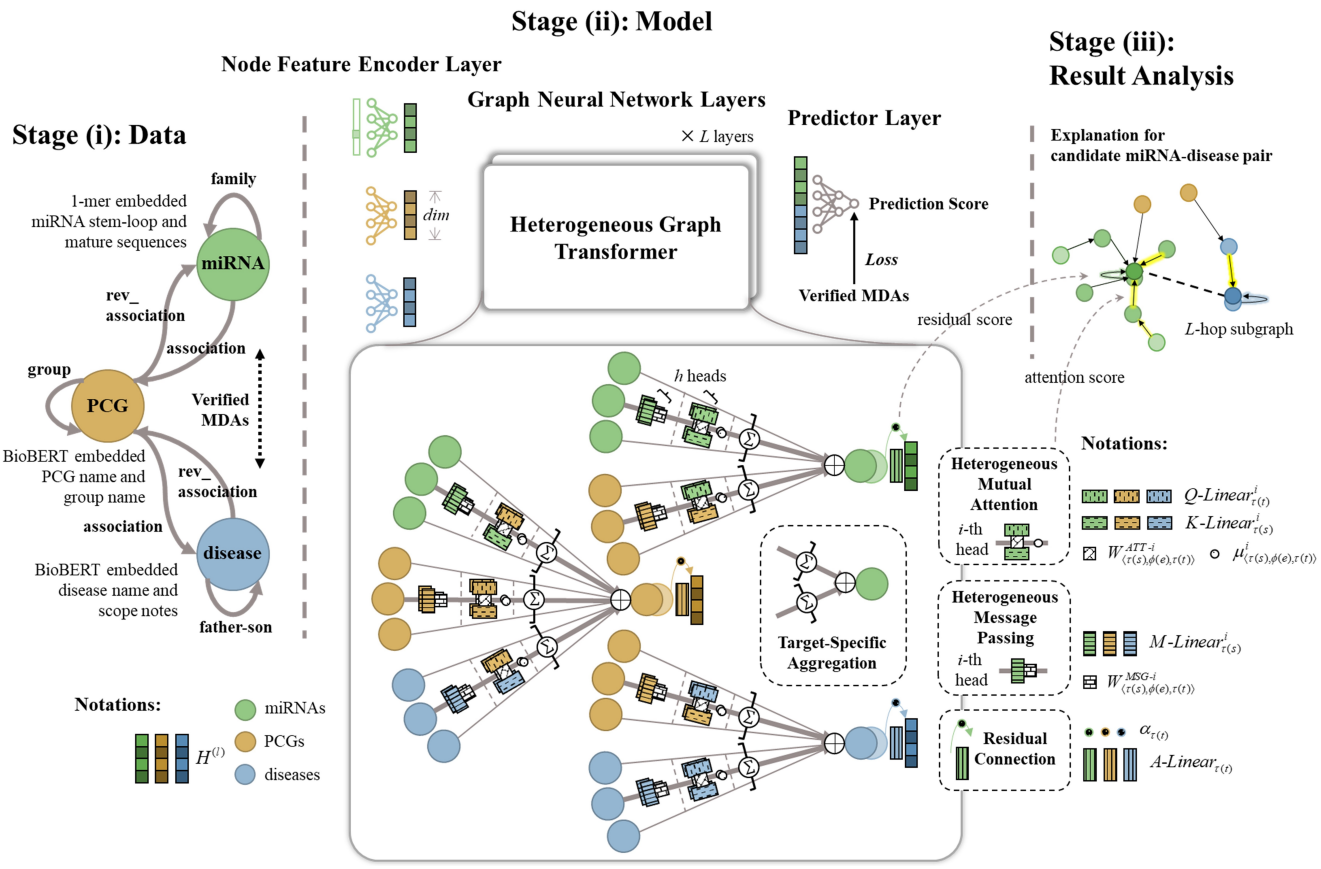


Figure 3: The flowchart of our MDA prediction method. Relevant data is integrated into a miRNA-PCG-disease graph with biomedical semantic node features. The prediction process consists of an encoder layer that learns node features, stacked GNN layers that learn graph structures, and a predictor layer for deriving classifications. The key module is a heterogeneous graph transformer (HGT), which prioritizes critical information by performing heterogeneous mutual attention, heterogeneous message passing, and target-specific aggregation. And we can extract the attention and residual scores to provide case-by-case explanations.

where there are three fundamental operators: the **ATT**ention operator estimates the importance of each source node s with respect to a target node t; The **MeSsaGe** operator represents the source node s; and the **Agg**regation operator aggregates the neighboring messages by using the attention weights as coefficients.

Concretely, to effectively utilize the available data, we employ the **Heterogeneous Graph Transformer (HGT)** [40] to extract crucial information. HGT leverages the Transformer architecture to learn heterogeneous mutual attention weights that facilitate source-to-target aggregation during message passing.

Each HGT layer begins with the **Heterogeneous Mutual Attention**. It calculates h heads of attention for each edge $e = (s, t)$, and the i-th head attention $\text{ATT}^i(s, e, t)$ can be formulated as follows:

$$
\begin{aligned}
Q^i(t) &= \text{Q-Linear}^i_{\tau(t)}\left(H^{(l-1)}[t]\right), \\
K^i(s) &= \text{K-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right), \\
\text{ATT}^i(s,e,t) &= \left(K^i(s) W^{ATT\text{-}i}_{\langle \tau(s),\phi(e),\tau(t)\rangle} Q^i(t)^T\right) \cdot \frac{\mu^i_{\langle \tau(s),\phi(e),\tau(t)\rangle}}{\sqrt{d}},
\end{aligned} \tag{3}
$$

where there are four main groups of learnable parameters: First, $\text{Q-Linear}^i_{\tau(t)}$ transforms the target node t of $\tau(t)$-type into a Query vector $Q^i(t) \in \mathbb{R}^{\frac{dim}{h}}$. Parallelly, $\text{K-Linear}^i_{\tau(s)}$ projects the source node s of $\tau(s)$-type into a Key vector $K^i(s) \in \mathbb{R}^{\frac{dim}{h}}$. Matrix $W^{ATT\text{-}i}_{\langle \tau(s),\phi(e),\tau(t)\rangle} \in \mathbb{R}^{\frac{dim}{h} \times \frac{dim}{h}}$ captures the distinct semantic of each meta-relation from s to t, and the similarity between Query and Key vectors is computed through two consecutive matrix multiplications. Especially, $\mu^i_{\langle \tau(s),\phi(e),\tau(t)\rangle}$ estimates the overall importance of each meta-relation and scales the attention score adaptively. It is initialized to 1, and the learned variations provide insights into the relative focus on different meta-relations.

Lastly, the attention vectors from the h heads are concatenated together, considering the grouping of attentions by meta-relations:

$$
\mathbf{ATT}_{\text{HGT}}(s,e,t) = \underset{\substack{s \in N(t) \\ \langle \tau(s),\phi(e),\tau(t)\rangle}}{\text{Softmax}} \left(\Big\Vert_{i \in [1,h]} \text{ATT}^i(s,e,t) \right). \tag{4}
$$

Here, all source neighbors of a target node t along the meta-relation $\langle \tau(s), \phi(e), \tau(t)\rangle$ are represented as $s \in N_{\langle \tau(s),\phi(e),\tau(t)\rangle}(t)$. Consequently, for each target node, the heterogeneous mutual

attention fulfills that $\sum_{\substack{s\in N(t)\\ \langle\tau(s),\phi(e),\tau(t)\rangle}} \mathbf{ATT}_{\mathbf{HGT}}(s,e,t) = 1_{h\times 1}$ through the application of a softmax function.

The HGT layer performs **Heterogeneous Message Passing** in parallel with the attention calculation, learning the information passing from the source nodes by each meta-relation. The encoding process of the i-th head message, denoted as $\mathrm{MSG}^i(s,e,t)$, can be formulated as:

$$\mathrm{MSG}^i(s,e,t) = \text{M-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) W^{MSG\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle}, \quad (5)$$

where $\text{M-Linear}^i_{\tau(s)}$ linearly transforms the $\tau(s)$-type source node s from $\mathbb{R}^{dim}$ to $\mathbb{R}^{\frac{dim}{h}}$, and a meta-relation-based matrix $W^{MSG\text{-}i}_{\langle\tau(s),\phi(e),\tau(t)\rangle} \in \mathbb{R}^{\frac{dim}{h}\times\frac{dim}{h}}$ is designed to incorporate the edge dependency during message passing.

Furthermore, all h heads of message vectors for edge $e = (s,t)$ are concatenated together to obtain $\mathbf{MSG}_{\mathbf{HGT}}(s,e,t)$:

$$\mathbf{MSG}_{\mathbf{HGT}}(s,e,t) = \Big\|_{i\in[1,h]} \mathrm{MSG}^i(s,e,t). \quad (6)$$

After calculating the heterogeneous mutual attention and messages, the HGT layer performs **Target-Specific Aggregation** to aggregate the messages from all source nodes to the target node, both within and across meta-relations. This aggregation is represented as follows:

$$\widetilde{H}^{(l)}[t] = \underset{\forall\langle\tau(s),\phi(e),\tau(t)\rangle}{\textbf{Meta-Agg}} \left(\sum_{\substack{s\in N(t)\\ \langle\tau(s),\phi(e),\tau(t)\rangle}} \mathbf{ATT}_{\mathbf{HGT}}(s,e,t)\cdot \mathbf{MSG}_{\mathbf{HGT}}(s,e,t) \right), \quad (7)$$

Within each meta-relation, the attention scores summing to 1, serve as weights to average the corresponding messages. The function **Meta-Agg** represents the aggregation operation across different meta-relations, where the default setting is Sum.

Eventually, the target node t's vector is mapped back to the specific feature distribution of the $\tau(t)$-type. The output of (l)-th layer is derived as follows:

$$\begin{aligned} H^{(l)}[t] =& \sigma_s\left(\alpha_{\tau(t)}\right)\cdot \text{A-Linear}_{\tau(t)}\left(\sigma_g\left(\widetilde{H}^{(l)}[t]\right)\right) + \\ & \left(1-\sigma_s\left(\alpha_{\tau(t)}\right)\right)\cdot H^{(l-1)}[t], \end{aligned} \quad (8)$$

where $\text{A-Linear}_{\tau(t)}$ performs a projection on the heterogeneously aggregated vector, followed by a GELU activation σ_g. And the term $\alpha_{\tau(t)}$ is a parameter controlled by a Sigmoid activation function σ_s, regulating the strength of the residual connections.

3.3. Predictor Layer

For each candidate miRNA-disease pair, the corresponding miRNA and disease vectors are concatenated together. The Predictor layer then generates the MDA prediction score using the following equation:

$$y = \sigma_s\left(\text{P}_2\text{-Linear}\left(\text{P}_1\text{-Linear}\left(H^{(L)}[m] \,\|\, H^{(L)}[d]\right)\right)\right). \quad (9)$$

Here, two consecutive linear transformations are applied to the concatenated feature vector from $\mathbb{R}^{2dim}$ to $\mathbb{R}^{dim}$ and to $\mathbb{R}^1$. The resulting vector is passed through a sigmoid activation function σ_s, which generates the MDA prediction score, denoted as y. A higher value of y signifies a greater likelihood of association between the miRNA and disease.

We adopt binary cross entropy as the loss function for optimizing our model, which is formulated as:

$$Loss = -\sum \left[\hat{y}\cdot log(y) + (1-\hat{y})\cdot log(1-y)\right], \quad (10)$$

where $\hat{y}$ denotes the supervision label, $\hat{y} = 1$ if the miRNA-disease pair is verified associated, otherwise $\hat{y} = 0$.

4. Results

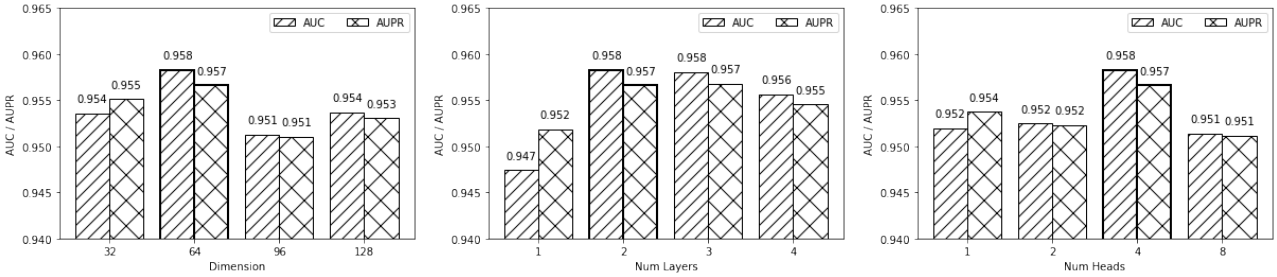
In this section, we conduct a comprehensive evaluation of our method with the aim of investigating the following research questions:

- **RQ1**. How does our EGPMDA compare to state-of-the-art baseline methods in terms of overall performance?
- **RQ2**. How is the generalizability of our method? Does it mitigate the neglect of entities that have fewer or no existing MDAs?
- **RQ3**. What is the impact of each component of the miRNA-PCG-disease graph on the performance of our method?
- **RQ4**. Does our method exhibit explainability? Can we derive the prediction basis for each candidate miRNA-disease pair?

4.1. Experimental Settings

Training, Validation, and Test Sets. As mentioned in Section 2, the verified MDAs are split into training, validation, and test sets based on the earliest publication time of literature evidence. The positive samples consist of verified MDAs. For the training and validation sets, an equal amount of miRNA-disease pairs without verified associations are randomly selected as negative samples. For the test set, in addition to maintaining a balanced ratio between positive and negative samples, we also increase the amount of the random negative samples to 100 times. Particularly, as depicted in Figure 1, the test set is split into sparse (*L-M*, *M-L* and *M-M*) and almost-blank (*0-M*, *L-0*, *M-0* and *L-L*) subsets by the known degree of miRNAs and diseases.

Implementation Details. The implementation of EGPMDA is based on PyTorch and PyTorch Geometric, the code is available at https://github.com/EchoChou990919/EGPMDA, and 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 employ an early stopping with a maximum of 50 epochs. During hyperparameter selection, the model is trained on



(a) Dimension of Hidden Layers (b) Number of Layers (c) Number of Heads

Figure 4: Hyperparameter selection. We set hyperparameters by comparing AUC and AUPR values on the validation set. The selected ones are framed by bold lines.

the training set and evaluated on the validation set. The weights that achieved the highest validation accuracy are preserved, with the patience parameter set to 5. After fixing all hyperparameters, the model is trained on the union of the training and validation sets to utilize more supervision information. The training process is stopped when the loss does not decrease for two consecutive epochs. All the computational experiments are repeated five times with different random initializations.

Evaluation Metrics. In binary classification tasks with prediction scores, the threshold for determining positive or negative labels is crucial. AUC and AUPR are threshold-insensitive metrics, denoting the areas under the receiver operating characteristic curve and precision-recall curve, respectively. Threshold-sensitive metrics include Accuracy (Acc), Precision (P), Recall (R), and F1-score (F1). When models are trained with balanced supervision and make 0-1 symmetric predictions, the default threshold is set to 0.5. Moreover, alternative metrics like $\text{Recall}_{@N}$ ($\text{R}_{@N}$) can be calculated by considering the top N-ranked samples as predicted positives and the remaining samples as predicted negatives.

In evaluating MDA predictions, successful detections are defined as miRNA-disease pairs that are predicted as associated and have verified association evidence. Precision is the ratio of successful detections to all miRNA-disease pairs predicted as associated. Recall represents the ratio of successful detections to all actual verified MDAs. Given that MDA prediction aims to identify potential associations that have not yet been verified, we prioritize Recall over Precision. The goal is to maximize the successful detections, even if it means predicting slightly more miRNA-disease pairs as potentially associated.

Therefore, in the scenario where there is a balance between positive and negative test samples, we utilize AUC, AUPR, Accuracy, Precision, Recall, and F1-score as evaluation metrics. In contrast, when the number of negative samples is increased, we calculate AUC, AUPR, $\text{Recall}_{@5\%}$, and $\text{Recall}_{@10\%}$ as evaluation metrics. Additionally, for analyzing generalizability, we compare the Recall on test subsets and visualize the distribution of top-ranked predictions.

Hyperparameter Selection. There are three main hyperparameters in EGPMDA: the dimension of hidden layers dim, the number of GNN layers L, and the number of multi-heads h. To determine the optimal values for these hyperparameters, we performed a grid search for $dim \in \{32, 64, 96, 128\}$, $L \in \{1, 2, 3, 4\}$ and $h \in \{1, 2, 4, 8\}$. In the histograms presented in Figure 4, the x-axis represents one unfixed hyperparameter, and the y-axis displays the corresponding average AUC and AUPR on the validation set. Eventually, we set $dim = 64$, $L = 2$ and $h = 4$.

4.2. Comparison to Baseline Methods

To assess the effectiveness of our proposed method, we compare it with five baseline methods:

- NIMCGCN (2020) [8] utilizes GCNs to learn latent representations from the miRNA and disease similarity networks and generate an MDA prediction matrix by neural inductive matrix completion.
- MMGCN (2021) [16] employs a multi-view GCN encoder with multi-channel attention to encode miRNAs and diseases from several similarity views, and derives predictions by multiplying the encoded representations.
- DFELMDA (2022) [6] integrates two branches of deep autoencoders that operate on similarity matrices, and obtains predictions using the deep random forest.
- MINIMDA (2022) [9] obtains representations of miRNAs and diseases from association and similarity networks. It utilizes GCNs to fuse mixed high-order neighborhood information and makes predictions through the multilayer perceptron.
- AGAEMD (2022) [10] designs a node-level attention graph encoder on a heterogeneous network that incorporates similarities and associations, and generates predictions through an inner product decoder.

Each baseline method utilizes and combines various types of similarity measures, which are recalculated specifically for our dataset. We carefully prevent information leakage by exclusively using training and validation MDAs for similarity calculation, GNN message propagation, and supervised training processes. It is notable that our reproductions already represent an extension of the original studies,

Table 2
Comparisons with Baseline Methods

	Positive & Negative Samples Balanced						Imbalanced			
Method	AUC	AUPR	Acc	P	R	F1	AUC	AUPR	$R_{@5\%}$	$R_{@10\%}$
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	0.948	0.950	0.847	0.957	0.728	0.827	-	-	-	-
MINIMDA	0.951	0.951	0.842	**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**

Table 3
Comparisons with Baseline Methods: Recall 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%**$_{(\uparrow\mathbf{131})}$
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%**$_{(\uparrow\mathbf{147})}$	92.7%
EGPMDA	**26.7%**$_{(\uparrow\mathbf{1})}$	0.0%	**17.0%**$_{(\uparrow\mathbf{152})}$	**1.5%**$_{(\uparrow 2)}$	**66.9%**$_{(\uparrow\mathbf{179})}$	67.5%	93.5%

as we have broadened their prediction scopes to encompass all human miRNAs and diseases based on our dataset.

The comparison between EGPMDA and the baseline methods is presented in Table 2. NIMCGCN and MMGCN exhibit relatively lower performance across all evaluation metrics. As they require the entire adjacency matrix for supervision, the abundance of negative label information results in an overmuch number of unassociated predictions. On a competitive level, our EGPMDA outperforms DFELMDA, MINIMDA, and AGAEMD. In the balanced case, EGPMDA achieves the highest values in terms of AUC, AUPR, Accuracy, Recall, and F1-score. Notably, EGPMDA surpasses the second-best method, AGAEMD, by 0.029 in Recall, indicating approximately 320 (0.029×11039) additional successful MDA detections. In the imbalanced case, our method achieves the highest values in AUC, AUPR, and $\text{Recall}_{@10\%}$, while ranking second to MINIMDA in $\text{Recall}_{@5\%}$.

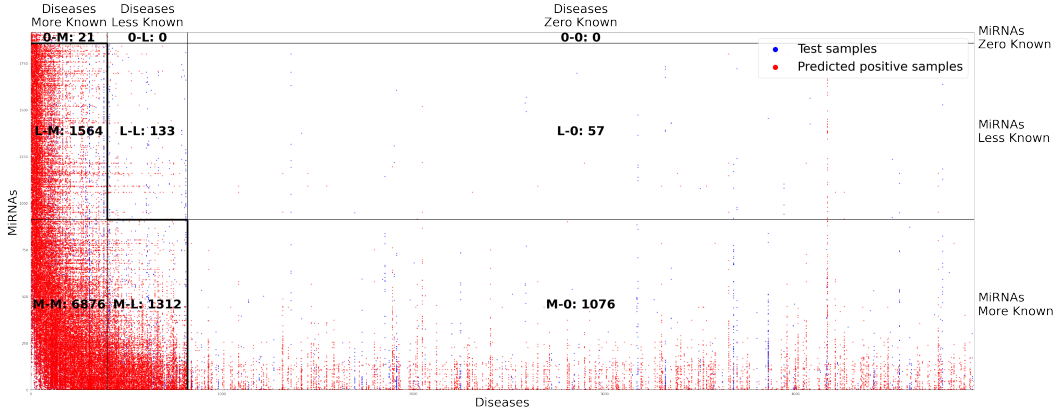
Further analysis of the generalizability is presented in Table 3, which showcases the average Recall across different subsets categorized as almost-blank (*0-M*, *L-0*, *M-0* and *L-L*) and sparse (*L-M*, *M-L* and *M-M*). DFELMDA achieves the highest Recall on the *M-M* subset, while AGAEMD performs the best on the *M-L* subset. However, both methods struggle to make successful MDA detections in almost-blank regions. In contrast, our EGPMDA not only performs well in sparse regions, achieving the highest Recall on the *L-M* subset and the second highest on the *M-L* and *M-M* subsets, but it also displays excellent performance in almost-blank subsets. Notably, EGPMDA demonstrates the most effective MDA predictions on three out of the four almost-blank subsets. For instance, in the *M-0* subset, which consists of 1076 MDA samples with disease ends that were never seen during model training, EGPMDA successfully detects 17.0% of these samples. This performance significantly surpasses the second-best method, MINIMDA, which only achieves a detection rate of 2.9%, equivalent to approximately 152 samples more.

Rethinking the imbalanced case, how does EGPMDA perform better overall but be secondary to MINIMDA on $\text{Recall}_{@5\%}$? It can be attributed to their distinct prediction patterns. As illustrated in Figure 5(a) and 5(b), the visualization showcases the Top@5% predictions from one of the five repeats. It is evident that MINIMDA's predictions tend to concentrate on sparse regions, while EGPMDA's predictions are more evenly distributed, with a relatively higher presence in almost-blank regions. When an equal number of predictions are considered positive (the top 5%), if a method "bravely" distributes its predictions in the almost-blank regions, it reduces the opportunity for successful detection in sparse subsets, which hold the majority of positive labels. Consequently, EGPMDA achieves better performance but slightly lower $\text{Recall}_{@5\%}$ compared to MINIMDA. These visualizations serve to reaffirm the generalizability of our EGPMDA method, which demonstrates a broader distribution of predictions, including almost-blank regions, albeit at a small trade-off in terms of $\text{Recall}_{@5\%}$.

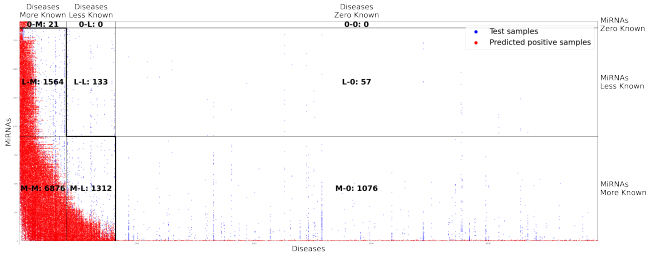
In conclusion, we can answer **RQ1** and **RQ2**: EGPMDA surpasses state-of-the-art baseline methods in terms of basic performance and demonstrates good generalizability. It effectively addresses the issue of neglecting unknown miRNAs and diseases.

4.3. Ablation study

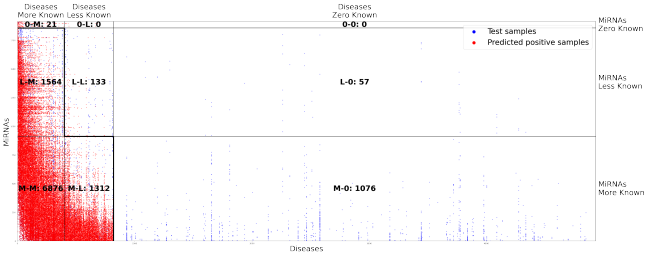
To assess the impact of different components in the miRNA-PCG-disease graph, we incrementally consider the following five conditions:



(a) EGPMDA, $\text{Recall}_{@5\%} = 0.757$, Recall on *M-0* = 17.0%



(b) MINIMDA, $\text{Recall}_{@5\%} = 0.777$, Recall on *M-0* = 2.9%



(c) EGPMDA while incorporating existing MDAs as an input feature, $\text{Recall}_{@5\%} = 0.771$, Recall on *M-0* = 0.0%

Figure 5: Top@5% predictions from EGPMDA, MINIMDA, and EGPMDA under condition 4 of the ablation study. In addition to comparing an evaluation metric, we visualize what miRNA-disease pairs can be predicted as associated. (a) EGPMDA demonstrates greater generalizability by successfully detecting miRNA-disease associations even when miRNAs and diseases have fewer or no verified associations. (b)(c) Although methods achieve high $\text{Recall}_{@5\%}$, they only provide accurate predictions for the sparse subsets (*L-M*, *M-L* and *M-M*), but fail to detect new associations on the almost blank subsets (*0-M*, *L-0*, *M-0* and *L-L*).

0 Utilization of supervision information only: we only utilize the supervision information, and the node features of miRNAs and diseases are randomly initialized. Since there is no graph structure, the GNN layers are skipped.

1 Incorporation of the biomedical semantic node ***Fea***tures: we further incorporate embeddings that represent miRNA sequences and disease description texts as node features. Similar to the previous condition, the GNN layers are excluded as there is no graph structure.

2 Inclusion of ***Intra***-class edges: we go a step further to utilize the graph structure that represents miRNA families and disease father-son relations.

3 Integration of information related to ***PCG***s: we incorporate node features transformed from PCG name texts, as well as the graph structure that represents PCG groups, miRNA-PCG associations, and disease-PCG associations.

4 Utilization of existing ***MDA***s: Finally, we leverage the graph structure that represents miRNA-disease associations, i.e. incorporating existing MDAs as an input feature.

It is worth noting that we primarily focus on the condition **3**, in which we also analyze the impact of the number of GNN layers ***L***.

Table 4 provides a summary of the results from our ablation study. Starting from condition 0 to **3**, the incorporation of different portions of the miRNA-PCG-disease graph leads to improved performance across most metrics. Moreover, stacking two GNN layers is found to be optimal as it allows for a 2-hop subgraph "reception field" that effectively captures information about PCGs without suffering from over-smoothing. Moving from condition **3** to 4, the performance remains competitive and even achieves higher balanced Precision, and imbalanced AUPR and $\text{Recall}_{@5\%}$.

However, as highlighted in Table 5, the incorporation of miRNA-disease edges can adversely affect the generalizability of the model. Specifically, the model's ability to successfully detect miRNA-disease associations for unknown miRNAs and diseases is significantly compromised. This limitation is evident in Figure 5(c), which displays the Top@5% predictions in condition 4. It is obvious that all the top-ranked predictions are concentrated in the undesired sparse regions, indicating a lack of effective predictions for unknown miRNAs and diseases.

The importance of meta-relations is captured by an explainable attention parameter $\mu^i_{\langle\tau(s),\phi(e),\tau(t)\rangle}$ (refer to Equation 3). Figure 6 illustrates the most significant 2-hop

Table 4
Ablation Study

	Ablation Types					Positive & Negative Samples Balanced						Imbalanced			
	Fea	*Intra*	*PCG*	*MDA*	*L*	AUC	AUPR	Acc	P	R	F1	AUC	AUPR	$R_{@5\%}$	$R_{@10\%}$
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

Table 5
Ablation Study: Recall on Almost-blank or Sparse Subsets

Ablation Types					Almost-blank Subsets				Sparse Subsets		
Attr	*Intra*	*PCG*	*MDA*	*L*	*0-M*$_{(21)}$	*L-0*$_{(57)}$	*M-0*$_{(1076)}$	*L-L*$_{(133)}$	*L-M*$_{(1564)}$	*M-L*$_{(1312)}$	*M-M*$_{(6878)}$
✓	✓	✓	×	1	**34.3**%	0.0%	10.7%	1.2%	66.5%	63.9%	93.4%
✓	✓	✓	×	2	26.7%	0.0%	**17.0**%	**1.5**%	**66.9**%	**67.5**%	93.5%
✓	✓	✓	×	3	13.3%	0.0%	12.0%	0.3%	68.3%	63.7%	**93.7**%
✓	✓	✓	✓	2	10.5%	0.0%	0.0%	0.2%	48.8%	65.4%	91.7%

paths for EGPMDA in the general condition **3**, which include < miRNA, family, miRNA, family, miRNA >, < PCG, rev_association, miRNA, family, miRNA >, and < disease, father-son, disease, father-son, disease >. These paths represent the natural attributes of miRNAs and diseases, and play a crucial role in the model's decision-making process. When incorporating existing MDAs into the message-passing framework, the model exhibits a heightened emphasis on < miRNA, association, disease, rev_association, miRNA > and < miRNA, family, miRNA, association, disease >, which are directly related to MDAs. Consequently, the model achieves precise predictions for entities with established associations, thereby maintaining a high level of evaluation metrics. However, this reliance on existing MDAs also results in a disregard for entities lacking such associations.

Based on our analysis, we can provide answers to **RQ3** and partially address **RQ4**: Each component of the miRNA-PCG-disease graph contributes to the prediction performance, but it is advisable to exclude existing miRNA-disease associations (MDAs) during GNN message passing to enhance generalizability. We can identify and explain the meta-relations that hold greater significance in general.

4.4. Case study

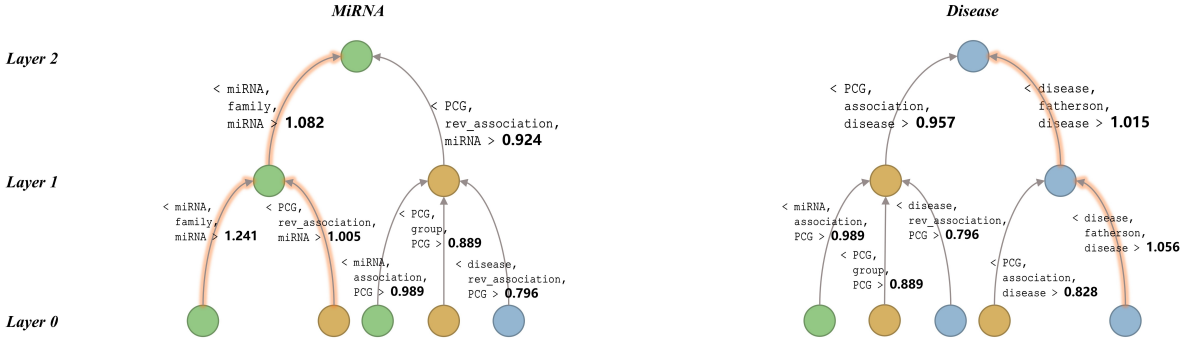
We conducted model retraining using all verified MDAs present in the dataset. Subsequently, predictions were made for all miRNA-disease pairs. The trained model and the corresponding prediction results can be accessed at https://github.com/EchoChou990919/EGPMDA. Additionally, we performed case studies on two diseases that do not have any existing MDA records in our dataset.

Table 6
The miRNAs associated with LADA

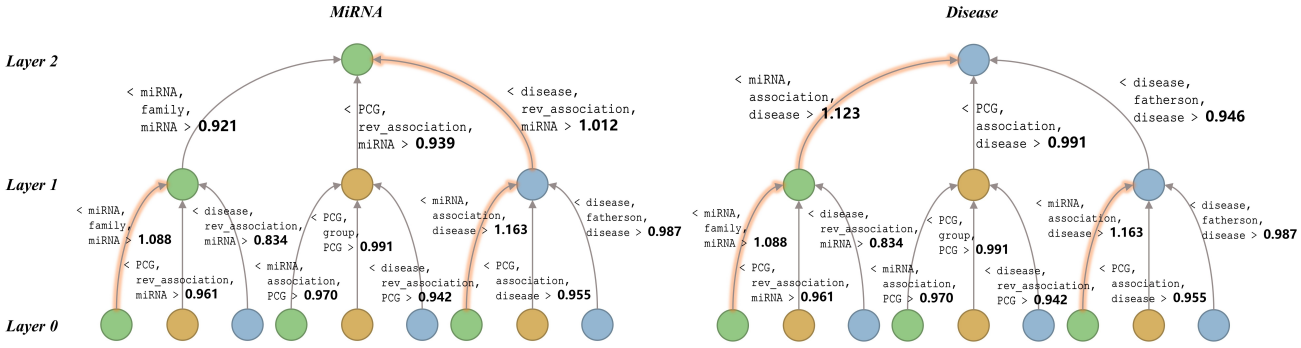
PMID	Mature Name	Accession	Pred. Score
36746199	hsa-miR-146a-5p	MI0000477	0.986
	hsa-miR-21-5p	MI0000077	0.965
	hsa-miR-223-3p	MI0000300	0.935
27558530	miR-34a	MI0000268	0.836
	miR-24.1	MI0000080	0.596
	miR-30d	MI0000255	0.703
32815005	hsa-miR-143-3p	MI0000459	0.584
31383887	hsa-miR-517b-3p	MI0003165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

Latent autoimmune diabetes in adults (LADA, MeSH ID: D000071698) is a heterogeneous disease characterized by a less intensive autoimmune process and a broad clinical phenotype compared to classical type 1 diabetes mellitus (T1DM) [41]. We've searched for literature and collected the following specific descriptions of the associations between miRNAs and LADA.

- Study [42]: "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 [43]: "People with LADA were best distinguished based on the levels of miR-34a, miR-24, and miR-21." and



(a) EGPMDA



(b) EGPMDA while including existing MDAs into message passing

Figure 6: Hierarchy of the learned meta-relations. For each meta-relation and each layer, we calculate the average attention value $\mu^i_{\langle\tau(s),\phi(e),\tau(t)\rangle}$ of four heads and five repeats. Important meta-relations with an attention score larger than 1 are highlighted.

"miRNAs like miR-34a, miR-30d, and miR-24 could be useful to classify subjects with LADA."

- Study [44]: "microRNA-143-3p contributes to inflammatory reactions by targeting FOSL2 in PBMCs from patients with autoimmune diabetes mellitus (T1DM and LADA)."
- Study [45]: "The qRT-PCR results further suggest the capability of circulating miRNAs, at least hsa-miR-517b-3p, as the LADA biomarker."

Table 6 shows the prediction results, and seven of the eight are successfully detected.

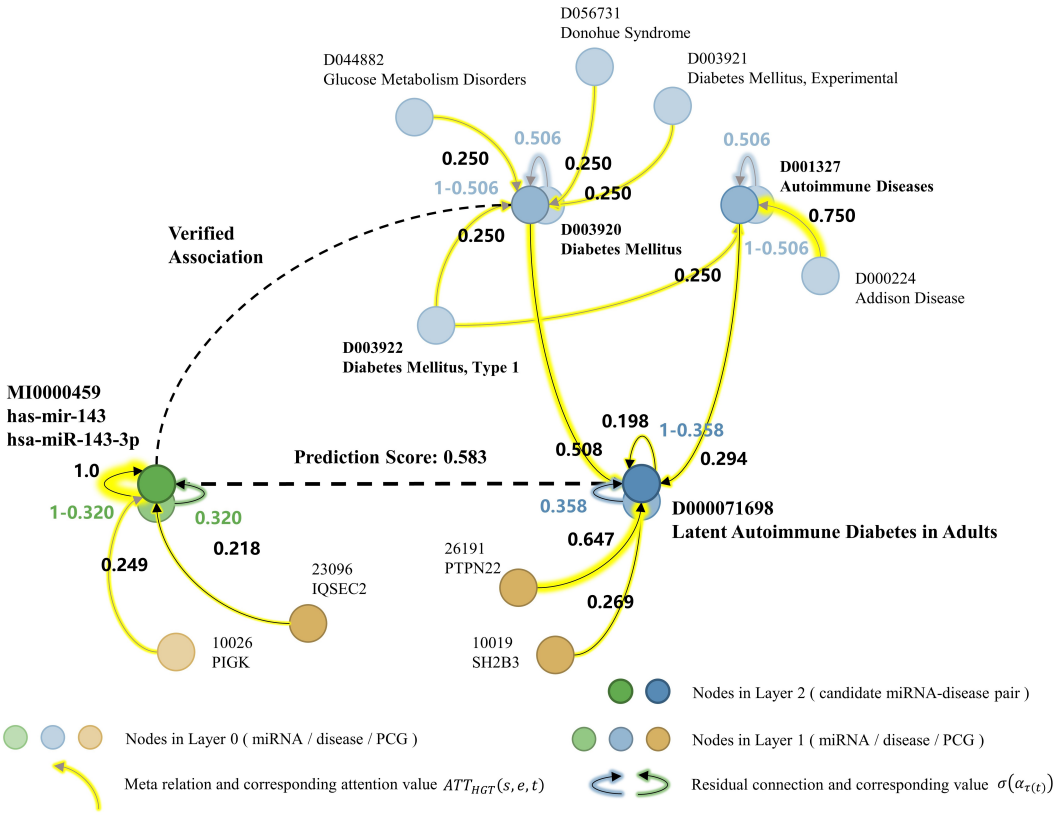
Teratozoospermia (MeSH ID: D000072660) is a type of male infertility, it is characterized by the presence of spermatozoa with abnormal morphology over 85% in sperm [46]. There are miRNAs associated with Teratozoospermia, and we've obtained the following descriptions:

- Study [47]: "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 [48]: "miR-182-5p, miR-192-5p, and miR-493-5p constitute a regulatory network with CRISP3 in seminal plasma fluid of teratozoospermia patients."
- Study [49]: "miR-34b and miR-34c were significantly associated with intracytoplasmic sperm injection (ICSI) clinical outcomes in male factor infertility, especially teratozoospermia."
- study [50]: "miR-582-5p expression significantly increased in teratozoospermia patients."

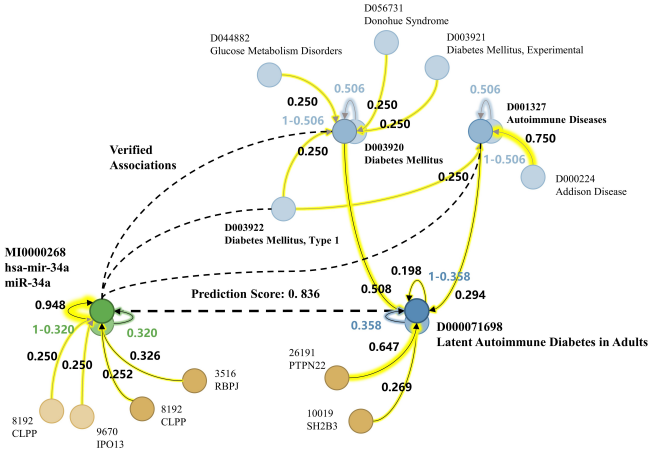
Table 7
The miRNAs associated with Teratozoospermia

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	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*

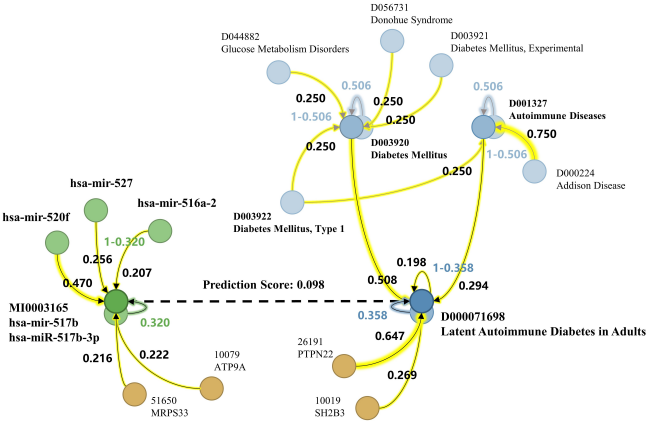
Table 7 shows the prediction results of these MDAs, and twelve of the fourteen are successfully detected.



(a) "hsa-miR-143-3p - LADA", Prediction Score = 0.583



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



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

Figure 7: Explanations of "hsa-miR-143-3p - LADA", "miR-34a - LADA" and "hsa-miR-517b-3p - LADA". Average residual and attention scores are mapped to the 2-hop subgraph surrounding each miRNA-disease pair. A higher prediction score is often obtained when there are similar miRNA-disease pairs that have already been learned to be associated.

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 $\mathbf{ATT}_{\mathbf{HGT}}(s,e,t)$ for meta-relations and $\sigma_s\left(\alpha_{\tau(t)}\right)$ for nodes in each HGT layer. We have developed a Jupyter Notebook[5], available in our open-source repository. It allows users to set any miRNA-disease pair of interest, and then delve into the explanation of the prediction process, ultimately uncovering significant messages (pathways) associated with the selected miRNA-disease pair.

Figure 7(a) visualizes a specific case involving the miRNA "hsa-miR-143-3p" and the disease "LADA". It illustrates that the semantics of "LADA" predominantly stem from "Diabetes Mellitus", "Diabetes Mellitus, Type 1", "Autoimmune Diseases", and the disease itself. Here the attention mechanism of HGT captures the closer relations in

[5] https://github.com/EchoChou990919/EGPMDA/blob/main/analysis_and_case_study.ipynb

reality correctly. Given the presence of a verified association "hsa-miR-143-3p - Diabetes Mellitus" within the dataset, our method suggests a potential association between "hsa-miR-143-3p" and LADA.

Furthermore, Figure 7(b) and 7(c) provides explanations for the predictions of "miR-34a - LADA" and "hsa-miR-517b-3p - LADA." The explanations shed light on the factors contributing to higher (0.836) or lower (0.098) prediction scores. Therefore, our method aligns with the conventional assumption that similar diseases are likely to be associated with similar miRNAs. However, the notion of "similarity" is not determined through heuristic similarity calculations; instead, it emerges from the message passing and aggregation process within the GNNs.

For **RQ4**, we can conclude that our method is explainable. It allows us to provide explanations for predictions on the instance level.

5. Conclusion

The identification of miRNA-disease associations (MDAs) holds significant value in disease diagnosis and treatment. Computational prediction methods have emerged as valuable tools for assisting biological experiments. It's crucial to generalize the effective predictions to entities with fewer or no existing MDAs and provide the basis of predictions. In the data stage, we construct a miRNA-PCG-disease graph that encompasses all authoritatively recorded human miRNAs and diseases. And verified MDAs are split reasonably to facilitate better evaluation. In the model stage, we propose EGPMDA, an end-to-end MDA prediction model. It comprises a node feature encoder layer, heterogeneous graph transformer-based graph neural network layers, and a predictor layer stacked sequentially. In the result analysis stage, computational experiments demonstrate that EGPMDA surpasses state-of-the-art methods in terms of both basic metrics and generalizability. Additionally, case studies showcase that our method can reliably detect potential MDAs for diseases without MDA records. Furthermore, we are able to explain the overall contribution of input features and the prediction basis for individual instances.

Despite the achievements, there is still room for improvement. Firstly, the heterogeneous graph can be expanded continuously by incorporating additional biological entities and associations, such as protein interactions. Secondly, obtaining trustworthy negative samples, rather than randomly selecting miRNA-disease pairs, would enhance the training of models. Moreover, future studies can enhance the human-computer interaction experience by integrating data, prediction results, and explanations into a visual analytic system, thereby providing more comprehensive benefits.

6. Acknowledgment

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 micrornas 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] Jihwan Ha. Smap: Similarity-based matrix factorization framework for inferring mirna-disease association. *Knowledge-Based Systems*, page 110295, 2023.

[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] Ping Li, Prayag Tiwari, Junhai Xu, Yuqing Qian, Chengwei Ai, Yijie Ding, and Fei Guo. Sparse regularized joint projection model for identifying associations of non-coding rnas and human diseases. *Knowledge-Based Systems*, 258:110044, 2022.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[30] 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.

[31] Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.

[32] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019.

[33] 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.

[34] 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.

[35] 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.

[36] Feng Xu, Yifan Wang, Yunchao Ling, Chenfen Zhou, Haizhou Wang, Andrew E Teschendorff, Yi Zhao, Haitao Zhao, Yungang He, Guoqing Zhang, et al. dbdemc 3.0: functional exploration of differentially expressed mirnas in cancers of human and model organisms. *Genomics, Proteomics & Bioinformatics*, 20(3):446–454, 2022.

[37] 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.

[38] 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.

[39] Jiaxuan You, Zhitao Ying, and Jure Leskovec. Design space for graph neural networks. *Advances in Neural Information Processing Systems*, 33:17009–17021, 2020.

[40] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[41] Silvia Pieralice and Paolo Pozzilli. Latent autoimmune diabetes in adults: a review on clinical implications and management. *Diabetes & Metabolism Journal*, 42(6):451, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.