MuCoMiD

A Multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction

Dong, Thi Ngan; Mucke, Stefanie; Khosla, Megha

DOI
10.1109/TCBB.2022.3176456

Publication date
2022

Document Version
Final published version

Published in
IEEE/ACM Transactions on Computational Biology and Bioinformatics

MuCoMiD: A **Mu**ltitask Graph **Co**nvolutional Learning Framework for **mi**RNA-**D**isease Association Prediction

Ngan Dong, Stefanie Mücke, and Megha Khosla

Abstract—Growing evidence from recent studies implies that microRNAs or miRNAs could serve as biomarkers in various complex human diseases. Since wet-lab experiments for detecting miRNAs associated with a disease are expensive and time-consuming, machine learning techniques for miRNA-disease association prediction have attracted much attention in recent years. A big challenge in building reliable machine learning models is that of data scarcity. In particular, existing approaches trained on the available small datasets, even when combined with precalculated handcrafted input features, often suffer from bad generalization and data leakage problems. We overcome the limitations of existing works by proposing a novel multitask graph convolution-based approach, which we refer to as MuCoMiD. MuCoMiD allows automatic feature extraction while incorporating knowledge from five heterogeneous biological information sources (associations between miRNAs/diseases and protein-coding genes (PCGs), interactions between protein-coding genes, miRNA family information, and disease ontology) in a multitask setting which is a novel perspective and has not been studied before. To effectively test the generalization capability of our model, we conduct large-scale experiments on the standard benchmark datasets as well as on our proposed large independent testing sets and case studies. MuCoMiD obtains significantly higher Average Precision (AP) scores than all benchmarked models on three large independent testing sets, especially those with many new miRNAs, as well as in the detection of *false positives*. Thanks to its capability of learning directly from raw input information, MuCoMiD is easier to maintain and update than handcrafted feature-based methods, which would require recomputation of features every time there is a change in the original information sources (e.g., disease ontology, miRNA/disease-PCG associations, etc.). We share our code for reproducibility and future research at https://git.l3s.uni-hannover.de/dong/cmtt.

Index Terms—Data integration, disease, graph representation learning, MiRNA, multitask

1 Introduction

Beginning in the early 2000 s, the biological dogma that proteins are responsible for most functions in a cell began to shift, and new classes of non-coding, regulatory RNAs became points of interest. The highly conserved class of microRNAs (miRNAs) with an approximate length of 22 nucleotides were first considered "junk" DNAs without any function. However, they later emerged as regulators in cell development, maturation, differentiation, and apoptosis, as well as cell signaling, cellular interactions, and homeostasis [1], [2], [3].

MicroRNAs fulfill their diverse functions by regulating gene expression of protein-coding genes (PCGs) after transcription. The transcribed messenger-RNA (mRNA) of PCGs can be directly bound by miRNAs, which leads to cleavage or destabilization of the mRNA and represses the translation into proteins [4]. Each miRNA can have hundreds of target mRNAs and each mRNA can be regulated by more than one miRNAs, resulting in complex networks that are yet to be fully understood. The mutation of miRNAs or changes in their expression can have widespread consequences that can be hard to predict.

Consequently, several associations between miRNAs and diseases have been confirmed using biological experiments leading to the belief that miRNAs could be potential biomarkers in certain diseases such as cancers or immune-related diseases [5], [6], [7], [8], [9], [10], [11]. Besides unveiling a deeper understanding of the molecular pathogenesis of diseases, identifying potential associations between miRNAs and diseases can also help in the treatment and discovery of possible drug targets.

Owing to the significance of the problem and the time-consuming nature of biological experiments, recent years have seen an upsurge in machine learning approaches [12], [13], [14], [15], [16] for *predicting miRNA-diseaseassociations*. Successfully predicting miRNA-diseaseassociations can lead to information prioritization in biological wet-lab experimentation, which will result in considerable time and cost savings. From a machine learning perspective, the problem of predicting miRNA-diseaseassociations can be formulated as a *link prediction problem* in a bipartite graph, where miRNAand

- *Ngan Dong is with the L3S Research Center, Leibniz University of Hannover, 30167 Hannover, Germany. E-mail: dong@l3s.de.*
- *Stefanie Mücke is with TRAIN Omics, Translational Alliance in Lower Saxony, 37081 Hannover, Germany. E-mail: trainomics@twincore.de.*
- *Megha Khosla is with the Delft University of Technology (TU Delft), 2628, CD, Delft, Netherlands. E-mail: m.khosla@tudelft.nl.*

Manuscript received 19 November 2021; revised 12 May 2022; accepted 16 May 2022. Date of publication 20 May 2022; date of current version 8 December 2022. This work was supported in part by present project funded by Volkswagenstiftung and in part by the Ministry for Science and Culture of Lower Saxony, Germany under Grant 11-76251-99-3/19 (ZN3434), in part by the Federal Ministry of Education and Research (BMBF), Germany under LeibnizKILabor project under Grant 01DD20003, and in part by the strategic funds of the Translational Alliance in Lower Saxony.
(Corresponding author: Ngan Dong.)
Digital Object Identifier no. 10.1109/TCBB.2022.3176456

disease form the two sets of nodes. However, as is typical to many biomedical applications, a major challenge for building generalizable and eventually well-performing models for the miRNA-disease association prediction problem is *data scarcity*. For example, the total number of miRNAand disease nodes (without any preprocessing and removal of duplicates) in the standard HMDD v2.0 [17] database are 578 and 383, respectively, with a total of 6,447 associations (links). These networks are, therefore, not only small with respect to the node set but are also sparsely connected. Even worse, the known association set is biased towards some well-studied diseases ($\sim$10% of the most well-studied diseases account for $\sim$60% of the known associations, and $\sim$20% of the most well-studied diseases account for $\sim$80 of the known associations).

Data Scarcity often leads to *biased* and *non-generalizable* models. Earlier attempts to address data scarcity rely on creating additional secondary features based on the initial feature set by computing intra-node similarities, e.g., miRNA functional similarity. So much so, these secondary features that encode miRNA functional similarity are pre-computed and are even deposited in well-known databases like MISIM [18]. The first problem associated with using such similarity-based input features is that as these features are derived from the training association data, the predictions become even more biased towards well-annotated diseases with many known associated miRNAs [19]. Moreover, errors in the training associations will be further exaggerated in the derived secondary features. Second, most secondary feature-based approaches cannot work effectively for new miRNAs or new diseases, i.e., instances for which no prior known associated disease (or miRNA) is available.

More worrying is the problem of *data leakage* when using pre-computed miRNA functional similarities indiscriminately from available databases. Dong and Khosla [20] find that most of the existing works that employ pre-computed similarities in model building ignore the actual train/test split giving rise to *data leakage*. In other words, some of the associations which are to be tested by the models are already presented in the association networks that are used to compute the similarity features. Finally, the small dataset size for the miRNA-diseaseprediction task prohibits the utility of flexible and more expressive modern representation learning approaches.

1.1 Present Work

We overcome the above limitations by avoiding the creation of secondary features altogether. Instead, our approach attempts to address the data scarcity and bias issues by integrating knowledge from multiple heterogeneous sources of information available for miRNA and diseases in addition to the miRNA-disease associations. Combining multiple data sources allows us to *compensate* for missing or unreliable information in any single data type and leads to more reliable predictions. Our key contribution is to model the integration of heterogeneous knowledge sources into a common representation space that can be trained end-to-end using modern representation learning machinery.

To this end, we propose a *Mu*ltitask graph *Co*nvolutational neural network for *mi*RNA-*D*isease association prediction, which we refer to as MuCoMiD for brevity. MuCoMiDallows automatic feature extraction while incorporating knowledge from five heterogeneous biological information sources (associations between miRNA/diseases and protein-coding genes (PCGs), interactions between PCGs, miRNA family information, and disease ontology) in a *multitask setting*. Instead of pre-calculated secondary features for miRNA and disease as in previous works, we employ graph convolution operation (with ReLU activation) over the corresponding biological networks to automatically learn informative representations for miRNA and disease at training time.

The added side tasks serve as regularizers and help us to incorporate domain knowledge. For example, a miRNA m regulates a set of proteins p that are responsible for some biological functions. Moreover, disruptions in the biological functions of p lead to certain disease condition d. Then m has some influence over disease d via p. The additional tasks of predicting miRNA-PCG association and disease-PCG association help us encode such influences by embedding m and d closer in the representational space. Besides, we employ an *adaptive loss balancing* technique to fine-tune the multitask loss gradients. This allows us to utilize the full power of multitask learning without resorting to exhaustive hyperparameter search.

We conduct an extensive evaluation of our approach in comparison to earlier works on existing benchmark datasets retrieved from the HMDD v2.0 [17] and HMDD v3.0 [21] databases. In addition to standard benchmark datasets, we also construct and test on new and large independent testing sets. We finally present case studies for specific diseases to showcase the utility of our approach in (i) differentiating between "true positives" and "false positives" and (ii) predicting the associations of novel diseases for which no prior miRN Aassociation information is available in the training set.

Our Contributions. To summarize, we make the following contributions:

- We model the miRNA-disease association prediction problem as a *multitask learning problem*, incorporating heterogeneous domain knowledge, which is a novel perspective and has not been studied before for the current problem.
- We construct *four new large testing sets* for testing in the transductive (when miRNA and disease nodes in the testing set are also present during training) and the inductive settings (when the testing set contains many new miRNAor disease nodes).
- We conduct large-scale experiments, ablation studies, and case studies for specific diseases to showcase the superiority of our approach.
- We release all the code and data used in this work for reproducibility and future research at https://git.l3s.uni-hannover.de/dong/cmtt.

2 Problem Statement and Related Work

The miRNA-disease association data can be represented using a bipartite graph $\mathcal{G}_{md} = (M, D, E)$ where M is the set of nodes representing miRNAs and D denotes the set of disease nodes. Each edge $e = (m, d) \in E$ denotes the association between the miRNAnode m and disease node d. We are then interested in the following problem statement.

Problem Statement. *Given the bipartite graph $\mathcal{G}_{md} = (M, D, E)$, we are interested in (i) predicting missing links among the given nodes (transductive setting) and (ii) predicting new links for unseen novel miRNAor disease nodes (inductive setting).*

2.1 Related Work

Existing computational approaches for miRNA-diseaseassociation prediction can be broadly grouped into three classes: *scoring-based*, *network topology* and *machine learning* based methods.

Assuming that the miRNA pairs linked to common diseases are functionally more related, scoring-based methods [18], [22] proposed scoring systems to prioritize miRNA-diseaseassociations. A more sophisticated scoring scheme integrating information from miRNAand disease similarity networks was proposed in [23]. Network-based approaches [24], [25], [26] construct miRNAs and/or disease similarity networks and aim at efficiently transferring known miRNA-disease association labels between similar miRNAs and/or similar diseases in the network. Chen *et al.* [24] employ random walks with restarts over the miRNA functional similarity network and prioritize candidate miRNA-disease associations using the final stable random walk transition probability.

More closely related to our work is the third category of machine learning-based methods. Approaches in this category mainly rely on using secondary or handcrafted features to construct similarity networks from which latent node features are extracted using graph-based representation learning techniques. EPMDA [12] extracts edge perturbation-based features from the miRNA-disease heterogeneous network and then trains a Multilayer Perceptron regression model to prioritize miRNA-disease associations. NNMDA [27] combines information from five different miRNA similarities and two disease similarities to build a heterogeneous network for feature learning and association prediction. [28] incorporates information from multiple domains, for example, miRNA-lncRNA and miRNA-PCG association, miRNA-drug association, disease-lncRNA, disease-PCG association, disease-drug association, to build a heterogeneous information network for feature extraction. The graph-based features, along with the miRNAk-mer feature (calculated from the miRNA sequence) and disease semantic similarity, are concatenated to form the input to a Random Forest classifier for association prediction. Non-graph based approaches like DBMDA [13] extracts latent features from input handcrafted features consisting of miRNA functional, disease semantic, and miRNAsequence similarities. Those latent features are then fed to a classifier to perform the miRNA-disease association prediction task.

Another line of works include NIMGCN [16], MMGCN [29], and DimiG [15] which propose end-to-end learning approaches in which graph convolution networks (GCNs) [30] are employed for extracting latent features of miRNA and disease nodes.

The reliance on handcrafted features based on existing association data limits the applicability of existing techniques. First, as most handcrafted-based methods rely on precalculated similarity features deposited in public databases, these methods are hard to update. When more data corresponding to the raw information sources become available (i.e., recently discovered miRNA-disease associations, disease ontology updates, more PCG associations information, etc.), new input features need to be generated before retraining the model. Second, besides the limitations concerning model generalization and data leakage (as already discussed in Section 1), a majority of these approaches cannot work effectively for new miRNAor disease nodes that have not been observed in the training data.

3 Proposed Approach

Given the input bipartite miRNA-disease network $\mathcal{G}_{md}$, we treat the miRNA-disease association problem as a binary classification where the label for an input pair node (m, d) is 1 if there is a known association between them and 0 otherwise.

To overcome the challenges of data scarcity, we propose MUCOMIDin which we focus on *effectively integrating heterogeneous biological information while learning to predict missing or new miRNA-disease associations*. Instead of relying on secondary or handcrafted features, we exploit additional knowledge sources to learn the input representation automatically at training time. Also, recent works indicate that PCGs are the most important links between miRNAs and their associated diseases [31] since changes in miRNAs lead to differently regulated PCGs, which in turn can cause diseases. Therefore, in addition to the miRNA-disease association prediction task (formulated as a binary classification problem), we add miRNA-PCG and disease-PCG association confidence score prediction (formulated as regression tasks) as two additional side tasks to incorporate additional domain knowledge and prevent overfitting.

In summary, MUCOMIDemploys different ways of integrating domain knowledge at different stages of the model building process. In particular, information from three different biological networks, namely, miRNA family, PCG-PCG interaction, and disease ontology, is directly used to learn the node representations during training. Besides, the miRNA-PCG and disease-PCG associations are employed to build additional regularization objective functions. Incorporating various information sources helps compensate for the lack of information in a single data source. It also helps in mitigating the data scarcity problem.

In the following, we describe the three modules of MUCOMID(also depicted in Fig. 1): (i) *input graph construction*, (ii) *graph convolution based feature extraction*, and (iii) *multitask optimization/learning*.

3.1 Input Graph Construction

We start by describing the construction or retrieval of various biological networks that we leverage as additional sources of information and the corresponding rationale.

miRNA family, $\mathcal{G}_m$. A miRNA family is the group of miRNAs that share a common ancestor in the phylogenetic tree. MiRNAs that belong to the same family usually have highly similar sequence secondary structures and tend to execute similar biological functions [32]. Similar miRNAs would tend to participate in the mechanisms of similar diseases. We retrieve the miRNA family information from the mirBase database [33]. The miRNAnetwork $\mathcal{G}_m$ is an unweighted undirected graph in which there is a connection between node A and node B if A and B belong to the same family.

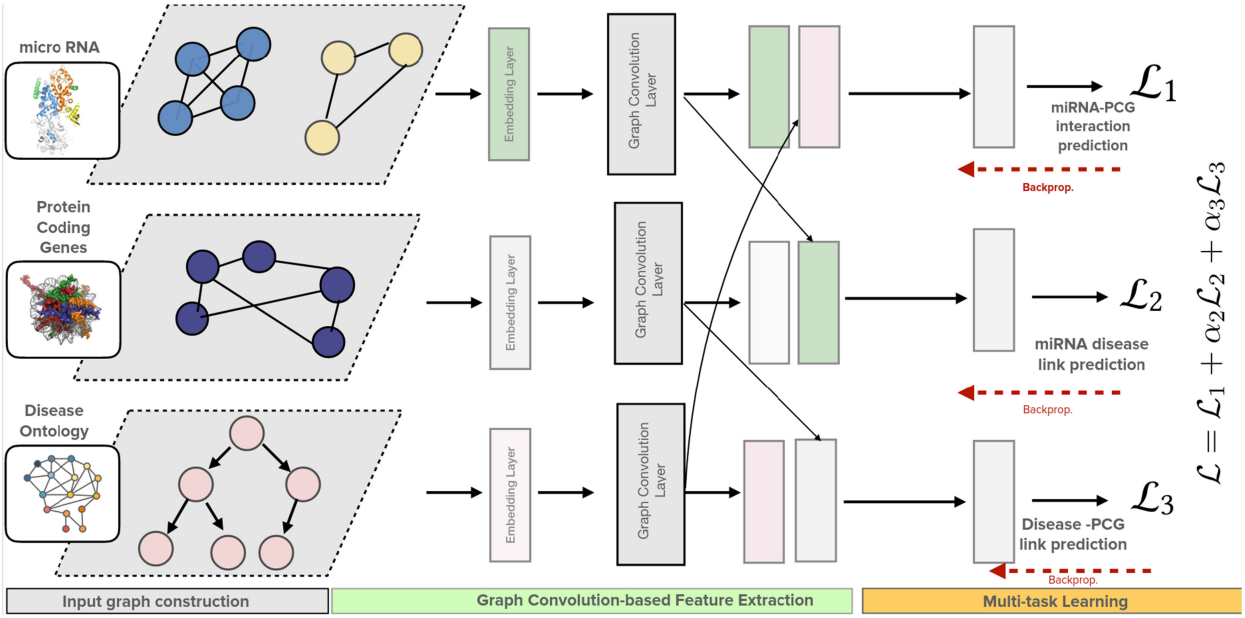

Fig. 1. A schematic diagram of MuCoMiD. MuCoMiDconsists of three main modules: (i) input graph construction in which we build networks corresponding to the available side information from miRNAfamily, PCG-PCG interactions, and disease ontology (ii) the second module takes the constructed networks as input and generates the nodes' representation according to their local neighbors (iii) finally, one classifier for miRNA-diseaseassociation prediction, two regressors for miRNA-PCG and disease-PCG association confidence score prediction are added. The second and third modules get trained jointly using a multitask loss. The multitask loss is a weighted sum of the three individual task losses and is optimized using a *dynamic loss balancing* technique. We use different colors to differentiate between embedding layers corresponding to different types of nodes. Green is used for miRNA, gray is for PCG, and pink is for disease.

Fig. 2 presents an illustration of the miRNA family network generated from our data.

Disease Ontology, $\mathcal{G}_d$. The disease ontology [34] represents the disease etiology classes. A directed connection between two diseases exists if there exists a *is-a* relationship between them. Similar diseases can be expected to associate with similar miRNAs. The disease ontology network $\mathcal{G}_d$ is an unweighted directed network in which there is a directed connection from A to B if B is a parent of A. $\mathcal{G}_d$ can be visualized as a directed tree which contains only directed connection between children and parents nodes. Each tree layer represents one layer of abstraction. The uppermost layer represents the most general disease category. An illustration of the disease ontology is given in Fig. 3.

PCG-PCG Interaction, $\mathcal{G}_p$. PCGs interact with PCGs to carry out biological functions. Therefore, given the fact that protein-coding gene p_1 activates the expression of protein-coding gene p_2, if the miRNA m can regulate p_1 then there should be some relation between m and p_2. In other words, information from the protein-protein interaction network will bring additional insights into the indirect relationship between miRNAs/diseases and the rest of the PCGs with which a direct interaction is not known. We download the PCG interaction data from the STRING v10.5 database [35]. As a preprocessing step, we retain only the PCG nodes that have at least one known association with miRNAs or diseases. We then divide the PCG-PCG interaction confidence scores by 1,000 to convert them to the [0,1] range and further filter out any PCG-PCG interaction with a confidence score smaller than θ_p. The results reported in Section 5 correspond to $\theta_p = 0.3$ as it leads to the highest AP score on the Novel-Diseasetest set. The PCG

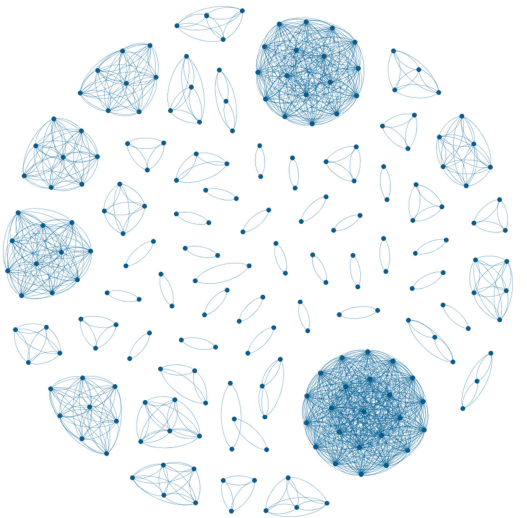

Fig. 2. An illustration of the miRNA family network. Each miRNA family forms a cluster in the network.

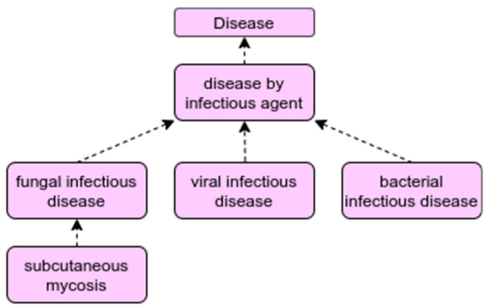

Fig. 3. An illustration of the disease ontology. Directed edges denote '*is_a*' relations.

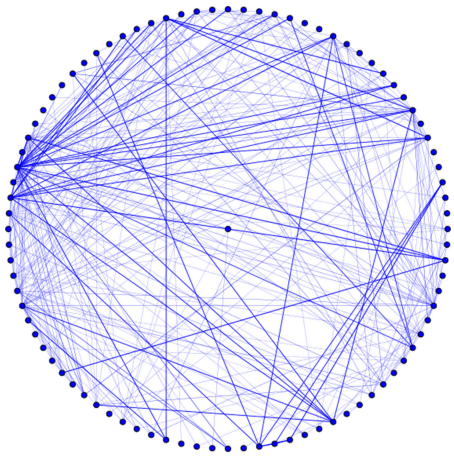

Fig. 4. An illustration of PCG-PCG interactions. The edge color intensity reflects the connection confidence score.

network $\mathcal{G}_p$ is an undirected weighted network in which the edge weights are the normalized PCG-PCG interactions' confidence scores. An example of the PCG-PCG interaction network is presented in Fig. 4 where the edge color intensity represents the interaction confidence score. The higher the score, the darker the color is.

3.2 Graph Convolution-Based Feature Extraction

Having constructed the relevant networks, we next extract informative node representations using the node neighborhood information. As we have no input node features, we use three embedding layers to encode the feature representation for miRNA, disease, and PCG nodes. As illustrated in Fig. 1, the green boxes represent the embedding layer and hidden representation for miRNA, gray is used for PCG, and pink is employed for disease. Those embedding layers are initialized randomly and will get updated during the model training process.

An embedding layer is essentially a look-up table where the ith row corresponds to the learned representation of the ith node. The node embedding is then passed as an input feature to the graph convolutional layer. A graph convolutional layer is essentially a linear layer that transforms the node feature as an aggregation of representations of its 1-hop neighbors. In particular, for the input adjacency matrix $\mathbf{A}$ and the node embedding matrix $\mathbf{X}$, we obtain transformed the node feature matrix $\mathbf{X}'$ as follows:

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}, \tag{1}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{D}}$ is the degree matrix of $\hat{\mathbf{A}}$, and $\mathbf{W}$ is the trainable weight matrix of the graph convolutional layer. We pass the transformed representation through a ReLU activation to obtain the final representation as follows:

$$\mathbf{X}'' = max(0, \mathbf{X}') \tag{2}$$

As the graph semantics are different for each network, no parameter sharing is employed at this stage. We use three separate graph convolutional layers to extract the representations for miRNA, PCG, and disease nodes. These learned representations will be fed as input to the multitask optimization/learning module explained in the next section.

3.3 Multitask optimization/learning

To effectively utilize information from the miRNA-PCG and disease-PCG associations, we design a multitask objective to train our model. In particular, for all input miRNA-disease, miRNA-PCG, and disease-PCG pairs, we model the pairwise representations as the elementwise products of the corresponding node features. For example, for an miRNA-disease input pair (m, d) denoted by nodes m and d, we obtain the corresponding feature vector representation as:

$$\mathbf{x}_{md} = \mathbf{X''}_m \odot \mathbf{X''}_d$$

where $\mathbf{X''}_m$ and $\mathbf{X''}_d$ correspond to the output representations of the graph convolution-based feature extraction for nodes m and d, respectively.

Using the pairwise representations, we then predict the existence of associations between miRNA-disease pairs and the confidence scores of associations for miRNA-PCG and disease-PCG pairs. In summary, we train our model with a multitask loss function calculated from these three supervised tasks and use an adaptive loss balancing technique to dynamically combine the three individual loss components at training time. Details about individual task loss and our optimization strategy are presented in the following sections.

3.3.1 MiRNA-Disease Binary Classification Task Loss ($\mathcal{L}_1$)

We compute the probability of observing an association between an miRNA-disease input pair (m, d) as:

$$y_{md} = \sigma\left(\mathbf{w}_{MD}^T \mathbf{x}_{md}\right) \tag{3}$$

where $\mathbf{w}_{MD}$ is a learnable weight matrix and $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function. We use binary cross entropy to calculate the training loss for the miRNA-disease classification module as follows:

$$\mathcal{L}_1 = \sum_{m,d} -z_{md}\log y_{md} - (1 - z_{md})\log\left(1 - y_{md}\right) \tag{4}$$

where z_{md} denote the target label known for the corresponding training pair.

3.3.2 MiRNA-PCG Regression Task Loss ($\mathcal{L}_2$)

For an input miRNA-PCG pair (m, p), we compute the association confidence score as:

$$y_{mp} = \sigma\left(\mathbf{w}_{MP}^T \mathbf{x}_{mp}\right) \tag{5}$$

where $\mathbf{w}_{MP}$ is a learnable weight matrix and $\sigma(x)$ is the sigmoid function. We use the sum of squared error to calculate the training loss for the miRNA-PCG regression module as follows:

$$\mathcal{L}_2 = \sum_{m,p} \left(y_{mp} - z_{mp}\right)^2, \tag{6}$$

where z_{mp} denotes the target confidence score.

3.3.3 Disease-PCG Regression Task Loss ($\mathcal{L}_3$)

We adapt the formula presented in Eq. (5) to compute the association confidence score y_{dp} for a disease-PCG input pair (d, p). $\mathcal{L}_3$ is then calculated using the sum of squared error as in 6:

$$\mathcal{L}_3 = \sum_{d,p} (y_{dp} - z_{dp})^2, \tag{7}$$

where z_{dp} denotes the target confidence score.

3.3.4 Multitask Optimization

We define the final loss for our model as the linear combination of three losses [36] as follows:

$$\mathcal{L} = \mathcal{L}_1 + \alpha_2 \mathcal{L}_2 + \alpha_3 \mathcal{L}_3, \tag{8}$$

where α_2 and α_3 are the loss weights for the two side tasks. Generally, multitask networks are difficult to train. Finding the optimal combination of individual task losses is challenging and problem-specific. A task that is too dominant during training will overwhelm the update signals and prevent the network parameters from converging to robust shared features that are useful across all tasks.

We follow the strategy presented in [36] and update α_2 and α_3 so that the difference between the two side tasks' contribution at each time step t is minimized. More specifically, at each time step t, the values for α_2, and α_3 are computed dynamically as follows:

$$\alpha_2(t) = \frac{\mathcal{L}_3(t-1)}{\mathcal{L}_1(t-1) + \mathcal{L}_2(t-1) + \mathcal{L}_3(t-1) + 10^{-10}} \tag{9}$$

$$\alpha_3(t) = \frac{\mathcal{L}_2(t-1)}{\mathcal{L}_1(t-1) + \mathcal{L}_2(t-1) + \mathcal{L}_3(t-1) + 10^{-10}} \tag{10}$$

We use an Adam optimizer with a learning rate of 10^{-3} to train the multitask model.

4 Experimental Setup

4.1 MiRNA-Disease Association Datasets

We retrieve the set of miRNA-disease associations from the HMDD v2.0 database [17] and the HMDD v3.0 database [21]. As pre-processing steps, we retain only the associations for the miRNAs and the diseases for which the PCG association information is available. The filtered data for the HMDD v2.0 database, which from now on is denoted as HMDD2, contains 2,303 known associations between 368 miRNAs and 124 diseases. The filtered data for the HMDD v3.0 database, which from now on is referred to as HMDD3, includes 8,747 known associations between 710 miRNAs and 311 diseases. Statistics about the data is presented in Table 2.

4.2 MiRNA-PCG Association

We obtain the miRNA-PCG associations from the RAIN database [37]. We include only the associations with the PCGs that are associated with at least one Reactome pathway [38] as these would be biologically more significant. We then normalize the association confidence scores retrieved from the database and filter out any miRNA-PCG association with a confidence score smaller than a cut-off threshold θ_m. The results presented in Section 5 correspond to $\theta_m = 0.5$ as it results in the highest AP score for the NOVEL-DISEASE testing set. In the end, the normalized confidence scores of the retained miRNA-PCG associations are used as the target values for the miRNA-PCG association confidence score prediction side task. Statistics about the data are presented in Table 1.

TABLE 1
Statistics for Datasets With Side Information

NETWORK	$\lvert E\rvert$	$\lvert V_m\rvert$	$\lvert V_d\rvert$	$\lvert V_P\rvert$
MIRNA-PCG	2,878	714	-	9,236
DISEASE-PCG	29,713	-	312	9,236
MIRNAFAMILY ($\mathcal{G}_m$)	1,354	217	-	-
DISEASE ONTOLOGY ($\mathcal{G}_d$)	90	-	128	-
PCG-PCG ($\mathcal{G}_p$)	1,407,590	-	-	9,236

$\lvert E\rvert$ is the Number of connections/associations. $\lvert V_m\rvert$, $\lvert V_d\rvert$, and $\lvert V_P\rvert$ are the Number of miRNAs, Diseases, and PCGs, Respectively.

TABLE 2
The miRNA-Disease Association Data Statistics Where $\lvert E\rvert$, $\lvert V_{miRNA}\rvert$, $\lvert V_{disease}\rvert$ Refer to the Number of associations/links, miRNAs and Diseases Respectively

DATASET	$\lvert E\rvert$	$\lvert V_{miRNA}\rvert$	$\lvert V_{disease}\rvert$
HMDD2	2,303	368	124
HMDD3	8,747	710	311
HELD-OUT1	2,669	324	110
HELD-OUT2	6,641	692	303
NOVEL-MIRNA	3,575	577	115
NOVEL-DISEASE	5,308	346	295

4.3 Disease-PCG Association

We obtain the disease-PCG associations from the DISEASES database [39]. Here also, we retain only the associations (i) with the PCGs that are associated with at least one Reactome pathway (ii) and have the normalized confidence scores greater than or equal to a confidence cut-off threshold θ_d. The results presented in Section 5 correspond to $\theta_d = 0.3$ as it results in the highest AP score in the NOVEL-DISEASEtesting set. In the end, the normalized confidence scores of the retained disease-PCG associations are used as the target values for the disease-PCG association confidence score prediction side task.

Table 1 provides statistics of the three biological networks as described in Section 3.1 and the two additional datasets described in Sections 4.2 and 4.3. Details about the number of miRNA-PCG, disease-PCG associations, and PCG-PCG interactions with different confidence cut-off thresholds can be found in Table 1 in the Supplementary file.

4.4 Our New Testing Sets

For small-size datasets like HMDD2 and HMDD3, 5-fold CV evaluation is limited as the size of the training and testing sets become much smaller. While one can use HMDD2for training and HMDD3for testing, such evaluation is limited as there are many overlapping associations in these two datasets. We, therefore, carefully construct the following four independent tests using the HMDD3 dataset. HMDD2*is used as*

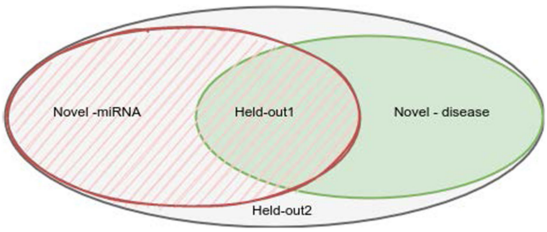

Fig. 5. An illustration of the four large independent testing sets relations where HELD-OUT1= NOVEL-MIRNA∩ NOVEL-DISEASE, HELD-OUT2= NOVEL-MIRNA∪ NOVEL-DISEASE= HMDD3\ HMDD2.

the training set for evaluation with the new testing sets. Let **M2** and **D2** be the set of all miRNAs and diseases in HMDD2, respectively. The construction of the four independent testing sets is described below.

HELD-OUT1 for Transductive Testing. The HELD-OUT1testing set contains only the associations that are present in HMDD3but not in HMDD2. We further remove any associations involving any miRNA that is not in *M2* and any disease that is not in *D2*. By doing that, we ensure that all nodes in the testing set are partly observed during training. Finally, HELD-OUT1-contains 2,669 known associations between 324 miRNAs and 110 diseases. We randomly generate the same number of negative samples from the set of unknown miRNA-diseasepairs.

HELD-OUT2 for Inductive Testing. We construct the HELD-OUT2testing set by including all miRNAand disease nodes and their known associations that are present in HMDD3but not in HMDD2. Note that different from HELD-OUT1, HELD-OUT2might also contain the associations corresponding to the miRNAand disease nodes that are not present in the training set HMDD2. HELD-OUT2consists of 6,641 known associations between 692 miRNAs and 303 diseases. We randomly generate the same number of negative samples from the set of unknown miRNA-diseasepairs.

NOVEL-MIRNA. From the set of known associations in the HELD-OUT2testing set, we remove any associations with the diseases that are not in *D2* to construct the NOVEL-MIRNA-testing set. NOVEL-MIRNA testing set consists of 3,575 associations between 577 miRNAs and 115 diseases. Regarding the node set, NOVEL-MIRNA contains data for 253 new miRNAs that are not observed in the training set HMDD2. We randomly generate the same number of negative samples from the set of unknown miRNA-diseasepairs.

NOVEL-DISEASE. Similarly, for constructing the NOVEL-DISEASEtesting set, we remove any associations with the miRNAs that are not in *M2* from the HELD-OUT2 testing set. NOVEL-DISEASEcontains 5,308 associations between 346 miRNAs and 295 diseases. Regarding the node set, NOVEL-DISEASE contains data for 185 new diseases that are not observed in the training set HMDD2. We randomly generate the same number of negative samples from the set of unknown miRNA-diseasepairs.

A schematic Venn diagram of the four large independent testing sets is presented in Fig. 5. The corresponding statistics are presented in Table 2. There are some inconsistencies between the number of miRNAs in HELD-OUT1, NOVEL-DISEASE, and *M2* or between the number of diseases in HELD-OUT2, NOVEL-MIRNA, and *D2* because for some miRNAs and diseases, all their known associations are already presented in the training set (HMDD2). Therefore, they do not appear in the known association set of the corresponding testing sets.

4.5 Benchmarked Models

We compare our model with seven recently proposed methods: EPMDA [12], NEMII [40], NIMGCN [16], DBMDA [13], DIMIG 2.0 [13], MMGCN [29], and GCSENET [41]. More details about our benchmarked models are given in Section 1 in the Supplementary file. Among the state-of-the-art methods, EPMDArelies on the network topology for feature extraction. NEMII learns structural embedding from the miRNA-disease network. The model also exploits the miRNA family and disease ontology information to enrich its input features. NIMGCN, DIMIG 2.0, MMGCN, and GCSENETutilize GCNs for feature learning. While DBMDAemploys autoencoders for feature transformation. DIMIG 2.0, MMGCN, and GCSENETintegrate similar information sources as those used by MUCOMID. Nevertheless, DIMIG 2.0 utilizes the disease-PCG associations to construct the model training objective and the miRNA-PCG associations to build the input network for feature learning. MMGCN exploits miRNA-PCG and disease-PCG associations to calculate the input similarity matrices. GCSENETuses miRNA-PCG and disease-PCG associations to construct the learning objective for its feature extractor.

4.6 Testing Setup and Evaluation

As in previous works, we perform 5-fold CV for testing on the HMDD2 and HMDD3datasets. We run 5-fold CV with 5 random initializations. In other words, for each dataset, we run each model $5 \times 5 = 25$ times and report the average performance with the standard deviation.

To test on our new testing sets, we train all models on the HMDD2 dataset. We run the experiments 5 times with random initializations and report the average performance scores along with the standard deviation.

Evaluation Metrics. We report the Area under the Receiver Operating Characteristic (AUC) and the Average Precision (AP) as our evaluation criteria. The reason for reporting AP instead of the Area under the Precision-Recall curve (AUPR) score is because AP provides a better performance estimate than AUPR as discussed in our previous work [20]. AP is calculated as the discrete sum of the changes in the recall at different thresholds instead of linear interpolation as that of AUPR, which can be too optimistic in cases where the number of thresholds (unique prediction values) is limited[1]. For our case studies, we report the number of "true" positives found at the top K highest prediction.

4.7 Hyperparameter Settings

4.7.1 MUCOMID

In all experiments, we fix the number of training epochs to 200, the embedding size and the hidden dimension both to 32. We employ Adam optimizer with a learning rate of 10^{-3} for training.

4.7.2 Benchmarked Models

For EPMDA, DBMDAand NIMGCN, we use the code and setup released in [20]. For NEMII, MMGCN, and GCSENET, we use the same code and setup as published by the authors. We emphasize that we substantially strengthened these methods

1. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

TABLE 3
Experimental Results for Benchmarked Models in the 5-Fold CV and Transductive Testing Setup

Method	HMDD2		HMDD3		HELD-OUT1	
	AUC	AP	AUC	AP	AUC	AP
EPMDA([12])	0.744 ± 0.019	$0.783 \pm .017$	0.520 ± 0.011	0.594 ± 0.010	0.427 ± 0.038	0.49 ± 0.02
NEMII([40])	0.837 ± 0.013	$\mathbf{0.844 \pm 0.017}$	0.898 ± 0.005	0.893 ± 0.006	$\mathbf{0.705 \pm 0.0}$	0.642 ± 0.0
NIMGCN([16])	0.785 ± 0.018	0.803 ± 0.015	0.795 ± 0.021	0.800 ± 0.018	0.623 ± 0.003	0.601 ± 0.006
DBMDA([13])	0.553 ± 0.019	0.537 ± 0.015	0.749 ± 0.010	0.696 ± 0.009	0.578 ± 0.0	0.548 ± 0.0
DIMIG 2.0([15])	0.493 ± 0.018	0.485 ± 0.012	0.516 ± 0.006	0.508 ± 0.005	0.429 ± 0.002	0.471 ± 0.003
MMGCN [29]	0.783 ± 0.014	0.780 ± 0.017	0.911 ± 0.04	$\mathbf{0.913 \pm 0.006}$	0.682 ± 0.004	0.627 ± 0.006
GCSENET [41]	0.593 ± 0.095	0.568 ± 0.072	0.613 ± 0.098	0.575 ± 0.079	0.552 ± 0.045	0.538 ± 0.031
MUCOMID(OURS)	0.839 ± 0.012	$\mathbf{0.837 \pm 0.015}$	$\mathbf{0.916 \pm 0.005}$	0.912 ± 0.006	0.684 ± 0.003	$\mathbf{0.68 \pm 0.005}$
Improvement over SOTA	0.2%	-0.8%	0.5%	-0.1%	-2.9%	5.9%

in our work since, in the original models, the authors simply take pre-calculated miRNA functional similarities from public databases (MISIM). Besides, the disease semantic similarity is originally proposed for the MESH ontology. Nevertheless, such information is not available for our disease set, and also using those pre-calculated miRNA functional similarities would lead to data leakage [20]. We instead calculate the disease semantic similarity from the disease ontology [34] and the miRNA functional similarity from the training data associations. This further points to the limited applicability of existing methods when the original information sources are updated.

For DIMIG 2.0, we use the code and parameters shared by the author. To test the model performance on our data, we compare DIMIG 2.0 with the input features as tissue expression profiles and DIMIG 2.0 with the one-hot vectors on the subset of our testing datasets that have miRNAexpression profiles available. The two models acquire similar performance. This implies that the use of tissue expression profiles as node features does not affect the model performance. We can, therefore, test the model on our data for which tissue expression information is unavailable by using one-hot encoding for input node features. The results reported in Section 5 correspond to the model with one-hot vectors as input features on our testing datasets without removing miRNAs that do not have expression profiles.

5 RESULTS

5.1 Results on Small Testing Sets

Following previous works, we perform 5-fold CV experiments on the HMDD2 and HMDD3 datasets. The results are shown in Table 3. The testing set size for this scenario is considerably small and contains only $1/5$th of the total associations. Such a train-test scenario allows us to quantify how well the models learn but is limited in testing the generalization power of the models.

MUCOMID gains a comparable performance compared with state-of-the-art approaches. There are only minor differences in MUCOMID's and the benchmarked methods' performance on the two datasets. Nevertheless, among the compared models, NEMII performs the best on the HMDD2dataset while MMGCN achieves the highest scores on the larger set HMDD3. In other words, no single benchmarked method claims its superior in both datasets. For such a reason, we claim that MUCOMID is better than all benchmarked models in the 5-fold CV testing setup.

The performance of EPMDAdrops considerably for the HMDD3dataset. EPMDAlearns edge features in an unsupervised manner corresponding to its contribution to a cycle of a particular length. Usually, the cycle length parameter is fixed to a small value due to an exponential increase in run time with an increase in cycle length. Moreover, the task signal is not used in learning the edge features. The loss of performance of EPMDAin HMDD3can be attributed to the limitation of finding the best cycle length hyperparameter applicable for HMDD3. This also limits the applicability of this model to a larger variety of datasets. NEMII, NIMGCN and MMGCN perform better than DBMDAdue to the higher representational capacity of the employed GCNs and the exploitation of additional graph structure information.

5.2 Results on Small Train but Large Testing Sets in Transductive Setting

Table 3 shows the results corresponding to the HELD-OUT1-testing set with a positive:negative sample rate of 1:1. Table 2 in the Supplementary file provides additional results regarding the larger negative sample rates. Recall that HMDD2 is used as the training set. In this scenario, the testing set size is much larger than the training set size allowing us to compare the generalization capability of the models. In general, NEMII acquires the highest AUC score, followed by MUCOMID with 2.9% lower. Regarding the AP score, MUCOMID attains the highest with a gain of at least 5.9%. This result is consistent with the experimental results for different positive: negative test sample rates presented in Table 2 in the Supplementary file. The overall drop in performance of all models in this scenario as compared to the small testing set size cases points to the hardness of this particular testing set.

Among the benchmarked models, EPMDAand DIMIG 2.0perform the worst, suggesting that auto-encoders or handcrafted topology-based features alone without parameter optimization are not promising approaches for the current problem. NEMII and MMGCN, which exploit multiple sources of information, gain the highest performance scores. This further emphasizes the importance of information integration.

5.3 Results on Inductive Setting Testing Sets

Table 4 shows the results corresponding to the inductive setting testing sets with the positive:negative sample rate of

TABLE 4
Experimental Results on the Large Inductive Setting Testing Sets With Many New miRNAs and Diseases

Method	Novel-miRNA		Novel-Disease		Held-out2	
	AUC	AP	AUC	AP	AUC	AP
Epmda	0.44 ± 0.028	0.529 ± 0.012	0.5 ± 0.0	0.5 ± 0.0	0.417 ± 0.071	0.513 ± 0.04
NEMII	0.68 ± 0.0	0.652 ± 0.0	0.709 ± 0.0	**0.68 ± 0.0**	0.66 ± 0.0	0.681 ± 0.0
DimiG 2.0	0.452 ± 0.001	0.480 ± 0.001	0.421 ± 0.001	0.467 ± 0.001	0.417 ± 0.003	0.465 ± 0.004
Nimgcn	0.533 ± 0.002	0.519 ± 0.005	0.672 ± 0.001	0.666 ± 0.003	0.534 ± 0.001	0.509 ± 0.002
Dbmda	0.537 ± 0.0	0.518 ± 0.0	0.569 ± 0.0	0.551 ± 0.0	0.553 ± 0.0	0.595 ± 0.0
MMGCN	0.556 ± 0.025	0.504	**0.711 ± 0.003**	0.678 ± 0.003	0.553 ± 0.03	0.493 ± 0.02
GCSENet	0.543 ± 0.027	0.518 ± 0.016	0.557 ± 0.053	0.536 ± 0.034	0.557 ± 0.031	0.517 ± 0.013
MuCoMiD	**0.701 ± 0.002**	**0.704 ± 0.002**	0.649 ± 0.005	0.658 ± 0.006	**0.667 ± 0.006**	**0.697 ± 0.007**
Improvement over SOTA	3.1%	8.0%	-8.7%	-3.2%	1.1%	2.3%

1:1. Table 2 in the Supplementary file provides additional results regarding the larger negative sample rates. Recall that Hmdd2 is used as the training set. Note that Held-out2is more than three times larger than the training data and contains new nodes that have not been seen in Hmdd2.

In general, MuCoMiD works effectively also for the inductive setting and outperforms all of its competitors on two out of the three testing sets.

5.3.1 Input Features for Benchmarked Models

We note that except DimiG 2.0, none of the benchmarked models can be directly used in inductive settings. These methods rely entirely or partially on the known miRNA-disease associations to construct their input features.

Specifically for Epmda, which fully relies on the Gaussian Interaction Profile kernel similarities extracted from the known associations, input features for new miRNAand diseases will be all zeros. NEMII, which concatenates the extracted features from the miRNA-disease association network with the miRNA family and the disease semantic similarity features, will have part of its input features for new miRNAs or new diseases as random values. Likewise, the miRNA functional similarity for new miRNAs in the MMGCN, Nimgcn, and Dbmda, GCSENet models will be all zeros. Therefore, a part of the miRNA-diseasepairs' final input representation in those models will be random values.

5.3.2 The Novel-miRNAand Held-out2Testing Sets

These two testing sets contain known associations for hundreds of new miRNAs, which are not observed in the training dataset.

MuCoMiD versus DimiG 2.0. Though DimiG 2.0can predict the association probabilities for new miRNAs and new diseases, the differences in the two models' architecture and learning objectives lead to a significant difference in their performance. Unlike MuCoMiD and other state-of-the-art models, DimiG 2.0is a semi-supervised method that uses only disease-PCG associations during training but not the known miRNA-disease associations. Also, DimiG 2.0 is formulated as a multi-label classification problem with large but very sparse label matrices. The high sparsity of the labels, along with the high class imbalance, leads to a degradation in learning.

MuCoMiD versus Other Methods. As discussed in Section 5.3.1, Epmda, NEMII, Nimgcn, MMGCN, Dbmda, and GCSENet have a part of the input features corresponding to new miRNAs to be zeros or random values. For such reasons, the performance of state-of-the-art methods drops significantly on those two testing sets. MuCoMiD significantly outperforms all of its competitors with a gain of up to 8% in AP score. Though the structural embeddings for new miRNAs and new diseases in the NEMII model are random, its performance still ranks the second-highest, suggesting that the miRNA family and disease semantic similarity features are quite informative for the current classification problem.

5.3.3 The Novel-DiseaseTesting Set

The Novel-Diseasetesting set contains known associations for 185 diseases that are not observed during training. For this testing set, MuCoMiD is outperformed by NEMII and MMGCNby small margins. We argue that the direct use of miRNA-disease association training data to compute miRNAsimilarity or structural embedding by Nimgcn, MMGCN, and NEMII leads to their better performance (on the Novel-Disease dataset) than MuCoMiD. These models are usually biased towards giving high scores to the well-known miRNAs (for which a lot of association information is already known in training data), leading to overall better scores.

5.4 Ablation Study

We conduct an ablation study to analyze the contribution of the additional tasks. The single task baseline (MuCoMiD-STT) employs a similar architecture as that of MuCoMiD but without the miRNA-PCG and disease-PCG association confidence score prediction side tasks. In other words, it also learns miRNA and disease representation from the miRNA family and the disease ontology networks, respectively. However, MuCoMiD-STT only has one classifier layer for the miRNA-diseaseassociation prediction task, instead of one classifier and two regressors as that of MuCoMiD.

Table 5 presents the results for MuCoMiDand MuCoMiD-STT on the both 5-fold CV and the independent testing setup. MuCoMiD performs comparably to its single task variant on the 5-fold CV testing setup while significantly supersedes its competitor on all of the large independent testing sets with *much less standard deviation values among runs*. These results are even more significant when considering the size of the testing data. The performance gain highlights the contribution of the two added side tasks. Since PCGs are the most important links between miRNAs and

TABLE 5
Ablation Study Results for Both the 5-Fold CV and the Independent Testing Setup

Method	MuCoMiD		Single task MuCoMiD	
	AUC	AP	AUC	AP
Hmdd2	0.839 ± 0.012	0.837 ± 0.015	$\mathbf{0.843 \pm 0.01}$	$\mathbf{0.843 \pm 0.016}$
Hmdd3	$\mathbf{0.916 \pm 0.005}$	0.912 ± 0.006	$\mathbf{0.916 \pm 0.004}$	$\mathbf{0.913 \pm 0.005}$
Held-out1	$\mathbf{0.684 \pm 0.003}$	$\mathbf{0.68 \pm 0.005}$	0.674 ± 0.094	0.663 ± 0.092
Novel-miRNA	$\mathbf{0.701 \pm 0.002}$	$\mathbf{0.704 \pm 0.002}$	0.683 ± 0.095	0.689 ± 0.096
Novel-disease	$\mathbf{0.649 \pm 0.005}$	$\mathbf{0.658 \pm 0.006}$	0.606 ± 0.084	0.618 ± 0.086
Held-out2	$\mathbf{0.667 \pm 0.006}$	$\mathbf{0.697 \pm 0.007}$	0.618 ± 0.086	0.66 ± 0.092

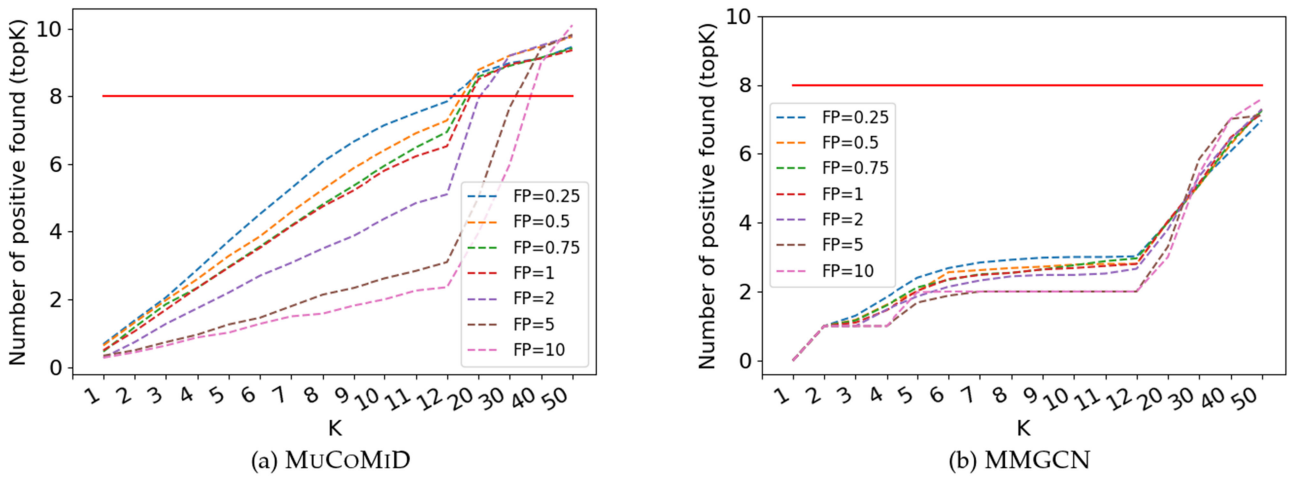

Fig. 6. The number of true positives found in the top K highest prediction of MuCoMiDand MMGCNfor the Parkinson disease case study. FP refers to the false positive:true positive sample rate that we add to the training data. The dense red line represents the number of true positives in the training data.

their associated diseases [31], miRNA-PCG and disease-PCG prediction tasks also bring additional insights into the miRNA-disease association prediction problem.

6 The Parkinson Disease Case Study

Parkinson disease (PD) is the second most common neurodegenerative disease worldwide [42]. Existing human association studies for Parkinson disease resulted in inconsistent findings with several "false positives" reported by [43].

In this case study, we aim to answer the question of "*How effectively can* MuCoMiD*help in identifying false positives?*". Towards that, we manually construct a "gold standard" dataset based on the data deposited in the HMDD databases and the data collected from [43]. We mark 12 miRNAs as "true positives" (those that are confirmed as true positives in the meta analysis [43]), 33 miRNAs as "false positives" (which are marked as positives in the HMDD databases but are confirmed as negatives in the meta-analysis) and 116 miRNAs as "true negatives" (those that are confirmed as negative by the meta-analysis). Note that among the 12 true positive miRNAs, only 8 are marked as positive in the HMDD databases.

Training and Testing Data Setup. We first construct the training dataset by including diseases other than Parkinson. Let $\mathbf{H} = \text{Hmdd2} \cup \text{Hmdd3}$. We remove any known associations for Parkinson disease from $\mathbf{H}$ to obtain the $\mathbf{H}'$ dataset. As for Parkinson alone, the false positive rate is nearly 3 fold. We expect such a high number of false positives also for other diseases. To mitigate the effect of high false-positive rates for other diseases, we construct $\mathbf{H}^*$ from $\mathbf{H}'$ such that each miRNA only associates with μ diseases where μ is the average number of diseases associated with a particular miRNA in $\mathbf{H}'$. If the number of diseases associated with miRNA m_1 is larger than μ, we randomly sample μ diseases from the set of known associated diseases. If the number of associated diseases for m_1 is smaller than μ, we sample with duplicates μ diseases from the set of known diseases. We follow the same strategy to obtain the negative samples set for $\mathbf{H}^*$.

Next, we create the training subset corresponding to the Parkinson's disease associations, which we refer to as $\mathbf{P}_{train}$. The number of positive samples in $\mathbf{P}_{train}$ consists of 8 miRNA-Parkinson associations with 8 "true positive" miRNAs that appear in $\mathbf{H}$ and $FP \times 8$ false positives Parkinson-miRNA associations. In our experiment, FP is varied over: {0.25, 0.5, 0.75, 1, 2, 5, 10}. The negative samples in $\mathbf{P}_{train}$ consist of every possible combination between Parkinson and any miRNAs that are not "true positive" or "false positive".

The Parkinson's training data is the union of $\mathbf{H}^*$ and $\mathbf{P}_{train}$. The Parkinson's testing data consists of all pairwise combinations between Parkinson and 12 "true positive" and 116 "true negative" miRNAs identified in [43].

Fig. 6 presents the number of true positives found in the top K predictions of MuCoMiDand the MMGCN model on the Parkinson testing set with varying false-positive rates.

Note that for each false positive rate, we run each model with 10 different sampled sets, and for each set, we run the model 5 times to make the comparison as fair as possible.

With the increase in the false positive rate, there is a decrease in the performance of both the models, which is intuitive. Remarkably, MuCoMiD significantly outperforms its competitor in differentiating between the "true positives" and the "false positives". "True positive" miRNAs consistently appear in the top predictions generated by MuCoMiD. While for MMGCN, even with only 25% "false positives" added to the training set, for K=8, there are only 2.92 "true positives" (on average) found. On the other hand, the average number of true positives (for K=8) found by MuCoMiDis 6.06. These results highlight the added benefit of our proposed model. We believe that the two added side tasks help inform the model and prevent it from overfitting the noisy training association data. At the same time, it further validates our concerns associated with secondary features-based methods. MMGCN, which extracts the input features from the training associations, cannot well differentiate the false positives from the true positives. We observe similar results in comparison with other baselines.

Due to space constraints, the 3 case studies on predicting the associations for novel diseases are provided in Section 3 of the Supplementary file.

7 Conclusion

We propose a multitask graph convolutional learning framework, MuCoMiD for the problem of predicting miRNA-disease associations. Our end-to-end learning approach allows automatic feature extraction while incorporating knowledge from five heterogeneous biological information sources. Incorporating multiple sources of information helps compensate for the lack of information in any single source and, at the same time, enables the model to generate predictions for any new miRNA or disease. Unlike previous works, our model can be employed in both transductive and inductive settings. To test the generalization power of models, we test them on both the existing benchmarked setup and on our constructed large independent testing sets. Large-scale experiments in several testing scenarios highlight the superiority of our approach. An ablation study is added to highlight the side tasks' contribution. We release all the code and data used in this study for reproducibility and future research at https://git.l3s.uni-hannover.de/dong/cmtt.

We believe that our design principles will be of independent interest to other biomedical applications where data scarcity is a major challenge. In particular, the use of multitask learning to integrate information from heterogeneous information sources to overcome the problems of data scarcity and unreliability of one single data type is a unique perspective and has not been studied for computational problems in biomedicine.

References

[1] J. S. Mattick and I. V. Makunin, "Small regulatory RNAs in mammals," *Hum. Mol. Genet.*, vol. 14, no. suppl_1, pp. R121–R132, 2005.

[2] V. N. Kim and J.-W. Nam, "Genomics of microRNA," *TRENDS Genet.*, vol. 22, no. 3, pp. 165–173, 2006.

[3] H. K. Saini, S. Griffiths-Jones, and A. J. Enright, "Genomic analysis of human microRNA transcripts," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 45, pp. 17719–17724, 2007.

[4] Y. Cai, X. Yu, S. Hu, and J. Yu, "A brief review on the mechanisms of mirna regulation," *Genomics, Proteomic. Bioinf.*, vol. 7, no. 4, pp. 147–154, 2009.

[5] M. W. de Ronde, J. M. Ruijter, P. D. Moerland, E. E. Creemers, and S.-J. Pinto-Sietsma, "Study design and qPCR data analysis guidelines for reliable circulating mirna biomarker experiments: A review," *Clin. Chem.*, vol. 64, no. 9, pp. 1308–1318, 2018.

[6] W. Usuba *et al.*, "Circulating miRNA panels for specific and early detection in bladder cancer," *Cancer Sci.*, vol. 110, no. 1, pp. 408–419, 2019.

[7] F. Jin *et al.*, "Serum microRNA profiles serve as novel biomarkers for autoimmune diseases," *Front. Immunol.*, vol. 9, 2018, Art. no. 2381.

[8] A. Keller *et al.*, "Toward the blood-borne mirnome of human diseases," *Nature Methods*, vol. 8, no. 10, pp. 841–843, 2011.

[9] R. Schickel, B. Boyerinas, S. Park, and M. Peter, "MicroRNAs: Key players in the immune system, differentiation, tumorigenesis and cell death," *Oncogene*, vol. 27, no. 45, pp. 5959–5974, 2008.

[10] W. Zhang, J. E. Dahlberg, and W. Tam, "MicroRNAs in tumorigenesis: A primer," *Amer. J. Pathol.*, vol. 171, no. 3, pp. 728–738, 2007.

[11] Y. Lin *et al.*, "Characterization of microRNA expression profiles and the discovery of novel microRNAs involved in cancer during human embryonic development," *PLoS One*, vol. 8, no. 8, 2013, Art. no. e69230.

[12] Y. Dong, Y. Sun, C. Qin, and W. Zhu, "EPMDA: Edge perturbation based method for miRNA-disease association prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2170–2175, Nov.–Dec. 2020.

[13] K. Zheng, Z.-H. You, L. Wang, Y. Zhou, L.-P. Li, and Z.-W. Li, "DBMDA: A unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease associations," *Mol. Ther. Nucleic Acids*, vol. 19, pp. 602–611, 2020.

[14] M. Liu, J. Yang, J. Wang, and L. Deng, "Predicting miRNA-disease associations using a hybrid feature representation in the heterogeneous network," *BMC Med. Genomic.*, vol. 13, no. 10, pp. 1–11, 2020.

[15] X. Pan and H.-B. Shen, "Scoring disease-microRNA associations by integrating disease hierarchy into graph convolutional networks," *Pattern Recognit.*, 2020, Art. no. 107385.

[16] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction," *Bioinformatics*, vol. 36, pp. 2538–2546, 2020.

[17] Y. Li *et al.*, "HMDD V2. 0: A database for experimentally supported human microRNA and disease associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1070–D1074, 2014.

[18] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[19] Z. Huang *et al.*, "Benchmark of computational methods for predicting microRNA-disease associations," *Genome Biol.*, vol. 20, no. 1, pp. 1–13, 2019.

[20] T. N. Dong and M. Khosla, "Towards a consistent evaluation of miRNA-disease association prediction models," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2020, pp. 1835–1842.

[21] Z. Huang *et al.*, "HMDD v3. 0: A database for experimentally supported human microRNA–disease associations," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1013–D1017, 2019.

[22] E. M. Small, R. J. Frost, and E. N. Olson, "MicroRNAs add a new dimension to cardiovascular disease," *Circulation*, vol. 121, no. 8, pp. 1022–1032, 2010.

[23] Z. Yang *et al.*, "dbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D812–D818, 2017.

[24] X. Chen, M.-X. Liu, and G.-Y. Yan, "Rwrmda: Predicting novel human microRNA–disease associations," *Mol. Biosyst.*, vol. 8, no. 10, pp. 2792–2798, 2012.

[25] G. Li, J. Luo, Q. Xiao, C. Liang, and P. Ding, "Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity," *J. Biomed. Informat.*, vol. 82, pp. 169–177, 2018.

[26] X. Chen, D.-H. Zhang, and Z.-H. You, "A heterogeneous label propagation approach to explore the potential associations between miRAN and disease," *J. Transl. Med.*, vol. 16, no. 1, 2018, Art. no. 348.

[27] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated microRNAs by using neural networks," *Mol. Ther. Nucleic Acids*, vol. 16, pp. 566–575, 2019.

[28] B.-Y. Ji, Z.-H. You, L. Cheng, J.-R. Zhou, D. Alghazzawi, and L.-P. Li, "Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.
[29] X. Tang, J. Luo, C. Shen, and Z. Lai, "Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction," *Brief. Bioinf.*, vol. 22, 2021, Art. no. bbab174.
[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[31] S. MΦrk, S. Pletscher-Frankild, A. Palleja Caro, J. Gorodkin, and L. J. Jensen, "Protein-driven inference of miRNA–disease associations," *Bioinformatics*, vol. 30, no. 3, pp. 392–397, 2014.
[32] B. Kaczkowski, E. Torarinsson, K. Reiche, J. H. Havgaard, P. F. Stadler, and J. Gorodkin, "Structural profiles of human miRNA families from pairwise clustering," *Bioinformatics*, vol. 25, no. 3, pp. 291–294, 2009.
[33] A. Kozomara and S. Griffiths-Jones, "miRBase: Integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D152–D157, 2010.
[34] L. M. Schriml *et al.*, "Disease ontology: A backbone for disease semantic integration," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D940–D946, 2012.
[35] D. Szklarczyk *et al.*, "String v10: Protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D447–D452, 2015.
[36] I. Leang, G. Sistu, F. Bürger, A. Bursuc, and S. Yogamani, "Dynamic task weighting methods for multi-task networks in autonomous driving systems," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–8.
[37] A. Junge *et al.*, "RAIN: RNA–protein association and interaction networks," *Database*, vol. 2017, Jan. 2017, doi: 10.1093/database/baw167.
[38] A. Fabregat *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, 2018.
[39] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "Diseases: Text mining and data integration of disease–gene associations," *Methods*, vol. 74, pp. 83–89, 2015.
[40] Y. Gong, Y. Niu, W. Zhang, and X. Li, "A network embedding-based multiple information integration method for the miRNA-disease association prediction," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–13, 2019.
[41] Z. Li, K. Jiang, S. Qin, Y. Zhong, and A. Elofsson, "GCSENet: A GCN, CNN and SENet ensemble model for microRNA-disease association prediction," *PLOS Comput. Biol.*, vol. 17, no. 6, 2021, Art. no. e1009048.
[42] M.-C. Kuo, S.-H. Liu, Y.-F. Hsu, and R.-M. Wu, "The role of noncoding RNAs in Parkinson's disease: Biomarkers and associations with pathogenic pathways," *J. Biomed. Sci.*, vol. 28, no. 1, pp. 1–28, 2021.
[43] J. Schulz *et al.*, "Meta-analyses identify differentially expressed microRNAs in Parkinson's disease," *Ann. Neurol.*, vol. 85, no. 6, pp. 835–851, 2019.

Ngan Dong is currently working toward the PhD degree with the L3S Research Center, Leibniz University of Hannover, Germany. She works as a research assistant with the PRESENt Project, which aims at integrating clinical, biological, and Big Data research to advance our understanding of norovirus gastroenteritis. Her current research interests include network analysis, feature selection, graph-based representation learning, multitask models, and data integration.

Stefanie Mücke received the PhD degree in biotechnology from the Gottfried Wilhelm Leibniz Universität Hannover, Germany, in 2020, working on host-pathogen interactions, omics analyses, genetic engineering, and biotechnology. Since 2020, she has been a scientific officer for the Translational Alliance in Lower Saxony, a network of ten research institutions in Hannover and Braunschweig (Germany). She supports cooperation projects between these institutions in the capacity of a project manager and scientific advisor.

Megha Khosla is an assistant professor with the Intelligent Systems Department, TU Delft, The Netherlands. Her main research interest include machine learning on graphs. She focuses on three key aspects of effectiveness, interpretability and privacy-preserving learning on graphs with applications to bio-medicine.