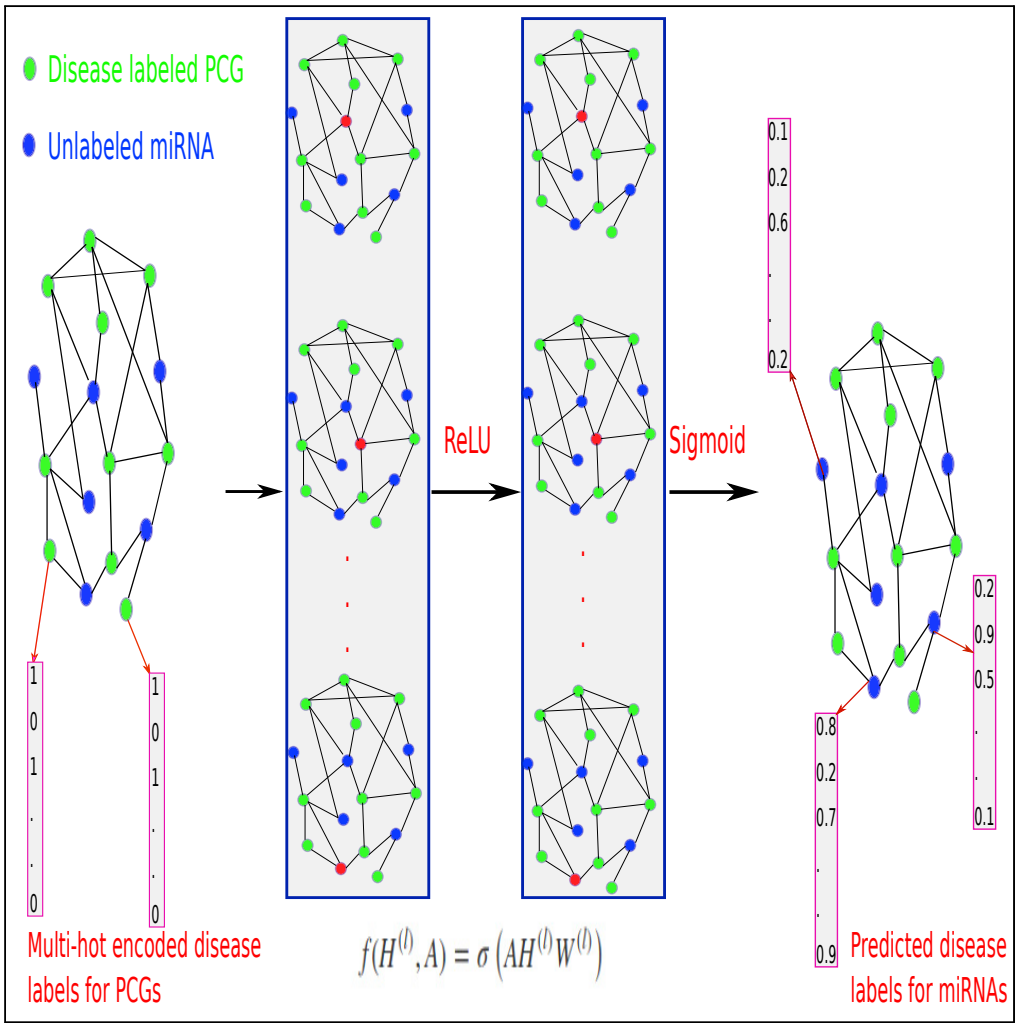
iScience

CellPress

Article

# Inferring Disease-Associated MicroRNAs Using Semi-supervised Multi-Label Graph Convolutional Networks




Xiaoyong Pan, Hong-Bin Shen

xypan172436@gmail.com (X.P.)
hbshen@sjtu.edu.cn (H.-B.S.)

## HIGHLIGHTS

Formulate disease-miRNA association prediction as node classification in a network

Semi-supervised graph convolutional network is used to infer miRNA-associated diseases

Introduce domain knowledge based on tissue expression profiles into model training

Evaluate DimiG using a strictly independent way; DimiG is superior to other methods

Pan & Shen, iScience *20*, 265–277
October 25, 2019 
https://doi.org/10.1016/j.isci.2019.09.013

iScience

CellPress

Article

# Inferring Disease-Associated MicroRNAs Using Semi-supervised Multi-Label Graph Convolutional Networks

Xiaoyong Pan[1,2,3,*] and Hong-Bin Shen[1,*]

## SUMMARY

**MicroRNAs (miRNAs) play crucial roles in biological processes involved in diseases. The associations between diseases and protein-coding genes (PCGs) have been well investigated, and miRNAs interact with PCGs to trigger them to be functional. We present a computational method, DimiG, to infer miRNA-associated diseases using a semi-supervised Graph Convolutional Network model (GCN). DimiG uses a multi-label framework to integrate PCG-PCG interactions, PCG-miRNA interactions, PCG-disease associations, and tissue expression profiles. DimiG is trained on disease-PCG associations and an interaction network using a GCN, which is further used to score associations between diseases and miRNAs. We evaluate DimiG on a benchmark set from verified disease-miRNA associations. Our results demonstrate that DimiG outperforms the best unsupervised method and is comparable to two supervised methods. Three case studies of prostate cancer, lung cancer, and inflammatory bowel disease further demonstrate the efficacy of DimiG, where top miRNAs predicted by DimiG are supported by literature.**

## INTRODUCTION

MicroRNAs (miRNAs) are a type of small non-coding RNAs with a size of about 22 nucleotides, and they interact with other RNAs to play important roles in transcriptional and post-transcriptional gene regulation (Bartel, 2004). It is estimated that over 60% of all human protein-coding genes (PCGs) are regulated by miRNAs (Friedman et al., 2009), and these miRNAs have been implicated in diseases. To date, the associations between diseases and PCGs are well investigated; many disease-PCG associations have been discovered and collected in public databases, e.g., DISEASES (Pletscher-Frankild et al., 2015), OUGene (Pan and Shen, 2016), and DisGeNET (Pinero et al., 2017). Compared with PCG's well-known important roles in diseases, the studies of effects of miRNAs are increasing. With increasing high-throughput sequencing data generated, more and more miRNAs are being discovered, and experimentally identifying their functions is costly and time consuming. Thus, it is imperative to develop computational methods to identify functional miRNA biomarkers associated with diseases, especially using rich information buried in disease-associated PCGs.

Some miRNAs are mainly expressed in certain tissues and show tissue specificity (Ludwig et al., 2016), which have certain tissue-specific expression patterns associated with diseases (Baker et al., 2017). They are expected to behave similarly to other disease-associated genes like PCGs or long non-coding RNAs (lncRNAs). Thus, several existing computational methods have used tissue expression data to infer gene-disease associations. For instance, GeneTIER makes use of disease-tissue associations to prioritize disease candidate genes (Antanaviciute et al., 2015). NetWAS identifies disease-associated genes by combining tissue-specific interaction networks and genome-wide association studies (Greene et al., 2015). Especially, some methods use tissue expression profiles with machine learning models to infer disease-associated lncRNAs. For example, DislncRF trains machine learning models on tissue expression profiles of disease-associated PCGs and further applies the trained models to infer disease-associated lncRNAs (Pan et al., 2019). All the above-mentioned studies demonstrated that tissue expression profiles indeed can facilitate the detection of disease-gene associations.

On the other hand, interaction networks contain rich clues for linking miRNAs to diseases. Many computational methods have been developed under the context of gene-gene networks (Chen et al., 2019). For example, Jiang el al. integrate miRNA and disease similarity network and miRNA-disease association to prioritize disease candidate miRNAs using a network-based approach (Jiang et al., 2010); midp applies random walk on the interaction network to infer disease-associated miRNAs (Xuan et al., 2015). Similarly,

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, 200240 Shanghai, China

[2]Department of Medical informatics, Erasmus Medical Center, 3015 CE Rotterdam, the Netherlands

[3]Lead contact

*Correspondence: xypan172436@gmail.com (X.P.), hbshen@sjtu.edu.cn (H.-B.S.)

https://doi.org/10.1016/j.isci.2019.09.013

RWRMDA implements random walk on the miRNA functional similarity network to link miRNAs to diseases (Chen et al., 2012); the MDHGI integrates the predicted association score based on sparse learning method to infer disease-associated miRNAs (Chen et al., 2018c). More closely related studies are as follows: DRMDA applies stacked autoencoder to learn deep representation for predicting miRNA-disease association (Chen et al., 2018a), LRSSLMDA and PBMDA use Laplacian regularized sparse subspace learning and path-based computational model for miRNA-disease association prediction (Chen and Huang, 2017; You et al., 2017), BNPMDA uses Bipartite Network Projection based on the known miRNA-disease associations (Chen et al., 2018b), and KBMF-MDI employs kernelized Bayesian matrix factorization to score miRNA-disease associations by integrating disease and miRNA similarity (Lan et al., 2018). Similarly, DLRMC infers disease-associated miRNAs using dual Laplacian regularized matrix completion (Tang et al., 2019). FamCluRank applies non-negative matrix factorization on the heterogeneous network with node attributes to predict disease-associated miRNAs (Xuan et al., 2018b). Especially deep learning has been utilized to extract deep representation for disease-miRNA association prediction (Xuan et al., 2018a).

A common hypothesis for the above methods is they assume that similar miRNAs can be associated with the same disease and similar diseases would be associated with the same miRNA. Thus, they commonly train and evaluate the models with representations of miRNAs and diseases as inputs on verified disease-miRNA associations through cross-validation approach.

However, as pointed out in Lehtinen et al. (2015), in the context of gene function prediction, cross-validation may be problematic because some gene-function associations are not independent in the benchmark set. There exists the same issue for disease-miRNA associations due to the following: (1) miRNAs from the same family may be associated with the same disease, (2) disease-associated miRNAs from miRNA-target assay may be derived from the targets that these miRNAs interact with, and (3) the associated miRNAs of child diseases are related to the miRNAs of parent diseases in disease ontology. When training and evaluating the models using cross-validation, randomly dividing the disease-miRNA associations may cause dependent associations to be separated into the training and test sets, potentially leading to an overestimated predictive performance. Cross-validating miRNA-disease associations may not actually reflect the method's ability to predict new miRNA-disease associations, but rather which information is dissipated in the benchmark set. In addition, as reported in Park and Marcotte (2012), there may exist flaws in cross-validation for computational pair-input prediction. One disease or one miRNA may be associated with multiple miRNAs or diseases, so randomly dividing disease-miRNA pairs into training and test sets will make some pairs in the test set share either the miRNA or the disease with the pairs in the training set, which causes the trained models to not generalize well to unseen disease-miRNA associations.

Thus, during cross-validation, complicated steps are required to make sure that dependent samples are divided into the same training set or the same test set and that pairs in the training and test sets do not share the miRNA or disease. It is almost impossible to construct a completely independent test set. An alternative strategy is that we do not use disease-miRNA associations for model training. For instance, instead of using miRNA-disease associations, the miRPD approach combines PCG-disease associations and miRNA-PCG network to score miRNAs and diseases (Mork et al., 2014). This has triggered us to further investigate disease-miRNA associations based on an interaction network. To date, there exist many high-confidence disease-PCG associations, and one miRNA may share the same disease with its PCG targets; we will be capable of transferring PCG-associated diseases to miRNAs on an interaction network under a new semi-supervised framework.

Recently, deep learning has achieved remarkable results in computational biology (Angermueller et al., 2016; Ching et al., 2018), especially convolutional neural networks (CNNs) (Lecun et al., 1998). CNNs can capture local correlation buried in data and mainly consist of convolutional layers, pooling layers, and fully connected layers. Many studies have demonstrated that the CNN networks are powerful in learning the hidden patterns from complicated biological data. For example, DeepBind (Alipanahi et al., 2015) and DeepSEA (Zhou and Troyanskaya, 2015) apply CNNs to predict preference of DNA/RNA-binding proteins and the impact of non-coding variants, respectively. iDeep (Pan and Shen, 2017) and iDeepE (Pan and Shen, 2018) further improve the performance of predicting RNA-binding protein (RBP)-binding sites and motifs using hybrid CNNs. The iDeepS (Pan et al., 2018) identifies binding sequence and structure preferences of RBPs simultaneously using CNNs and long short-term memory network.

Although the CNN has shown its power, it cannot handle structured datasets, like gene-gene networks. To analyze these types of network data, graph convolutional networks (GCNs) have been developed (Defferrard et al., 2016; Hamilton et al., 2017; Kipf and Welling, 2017). Under the framework of spectral graph convolutions, it encodes both local graph structure and features of nodes. The GCNs have been used on the graph data to predict polypharmacy side effects, where the graph is a multimodal graph constructed from protein-protein interactions, drug-protein interactions, and the polypharmacy side effects (Zitnik et al., 2018). The GCN is a graph-based semi-supervised learning method that does not require labels for all nodes. This setting is especially powerful for inferring miRNA-associated diseases, because many miRNAs are not well investigated about their associations with diseases and many disease-PCG associations are available. Compared with traditional semi-supervised methods (Jia et al., 2016; Wan and Wang, 2019; Zhang et al., 2018; Zoidi et al., 2018), GCNs can capture the structural information within the node's local network, similar to CNNs in images. In addition, one PCG or miRNA can be associated with multiple diseases. Thus, we can formulate the prediction of disease-miRNA associations as a multi-label classification problem.

In this study, we present a new semi-supervised multi-label learning method, DimiG, based on GCNs to integrate multiple networks of PCG-PCG interactions, PCG-miRNA interactions, PCG-disease associations, and tissue expression profiles to infer miRNA-associated diseases. The DimiG does not require the disease-miRNA associations, and it is trained on the graph consisting of PCG-PCG and miRNA-PCG interactions, where only PCGs have labeled diseases. Then DimiG is further used to score associations between diseases and miRNAs.

This study has made the following four major contributions for understanding disease-miRNA associations. (1) We further demonstrate that cross-validation performance of methods trained on known disease-miRNA associations could be overestimated and may not be able to reflect the method's actual ability to predict new disease-miRNA associations. We have proposed a network-based knowledge transfer approach for this problem. Considering that an miRNA may share the same disease with its PCG targets and there exist many high-confidence disease-PCG associations, we will be able to transfer the PCG-associated diseases to miRNAs in an interaction network framework. (2) We have formulated disease-miRNA association prediction as a semi-supervised multi-label node classification in a graph, which can help learn the complex networks composed of unlabeled miRNAs and labeled PCGs and the multi-label associations. This is a new prediction protocol for this problem. (3) We use semi-supervised GCN to learn patterns from PCG-associated diseases on an interaction network, which are further used to score diseases and miRNAs. This GCN-based approach combines the advantages of deep learning for representation learning and network-based methods. (4) We have further incorporated the domain knowledge into our model construction. Considering that miRNAs are often expressed in a tissue-specific way, we integrate the expression profiles across tissues into our GCN framework. Our results demonstrate that informative signals in more tissues can be captured for aiding the inference of disease-associated miRNAs.

## RESULTS

In this study, we first evaluate the prediction performance of DimiG on four tissue expression data with different number of tissues. Then we compare DimiG with other baseline methods that do not use disease-miRNA associations for model training on the independent test set, and further compare with BNPMDA and DRMDA trained using disease-miRNA associations on the unseen disease-miRNA set. Last, we present three case studies for prostate cancer, lung cancer, and inflammatory bowel disease (IBD).

### DimiG Pipeline

DimiG integrates gene network, expression profiles, and PCG-disease associations (Figure S1) using semi-supervised multi-label GCN. DimiG only requires PCGs with associated diseases (Figure S2) as labels during the model training, and then it propagates the node embedding to those miRNAs and further infers their associated labels. DimiG consists of two layers of GCNs, which require a node feature matrix, an adjacency matrix, and a label matrix. Each node (gene) is represented as a vector of expression profiles across tissues from GTEx (Lonsdale et al., 2013). The adjacency matrix is derived from the PCG-PCG and PCG-miRNA interactions. Only PCGs have assigned labels, which are a multi-hot vector corresponding to the presence of 248 associated diseases. We train GCN models on labeled PCGs and the interaction network, and the trained GCN model is used to score associations between diseases and unlabeled miRNAs. In the end, DimiG outputs a 1,034 × 248 score matrix, where 1,034 is the number of miRNAs and 248 is the number of diseases. More details are shown in Figure 1.

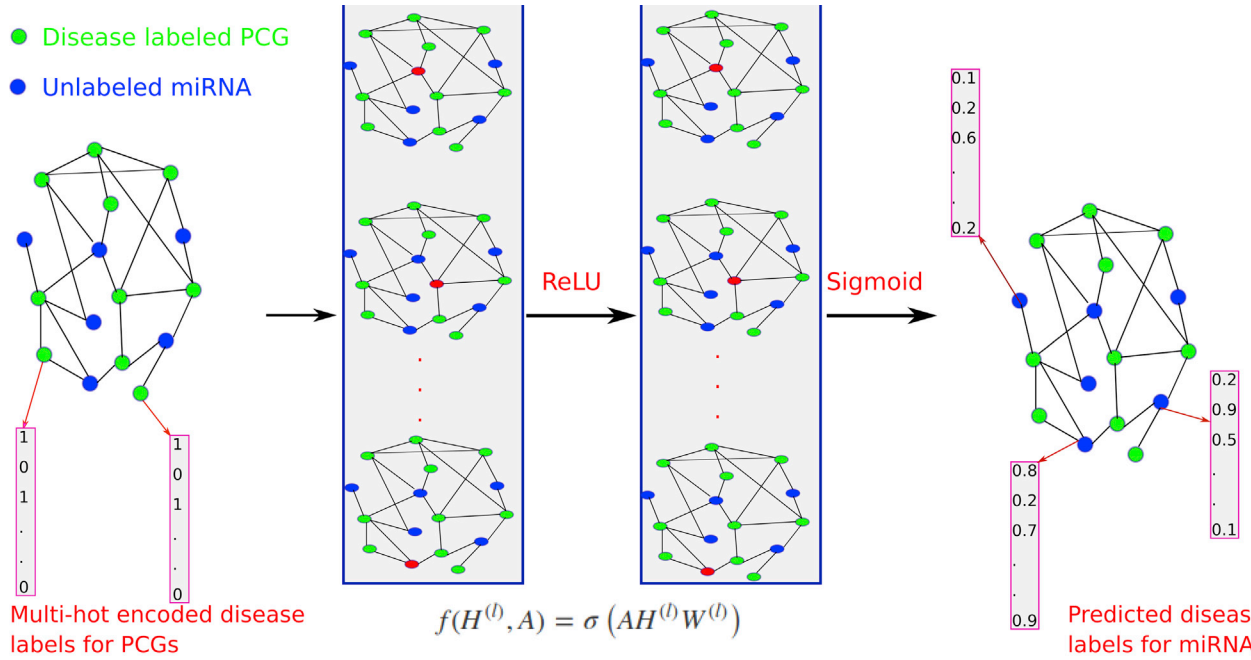



**Figure 1. The Flowchart of DimiG with Two-Layer GCN**

Each node (gene) is represented as a vector of expression values across tissues with its sum across tissues from GTEx, and the network is constructed from PCG-PCG and PCG-miRNA interactions. When doing forward propagation, the embedding of the red node in each network is the weighted sum of the embedding of its neighbors, where all nodes in the network are updated simultaneously. The label is a multi-hot vector indicating the presence of diseases. In the end, DimiG can infer the probability between diseases and unlabeled miRNAs.

## The Performance Comparison of DimiG on Four Tissue Expression Datasets

We first checked the impact of the number of epochs on training DimiG. As shown in Figure 2A, the training and validation loss converge to the same as the number of epochs approaches 50. Thus, in this study, we use 50 epochs for our below-mentioned experiments. It should be noted that the training loss is larger than validation loss during the first 30 epochs, which is also observed for the citation network data in the GCN article (Kipf and Welling, 2017). One possible reason is regularization (e.g., dropout), which is used during training, but not during validation. Another potential reason is that the features of the genes have certain discriminative power and they behave similarly to other genes in the training set. After several epochs, the learned node features are propagated well from interacting neighbor nodes for both training and validation genes.

Grid search approach is used to select the best parameters for DimiG, where we search the learning rate with values of [0.001, 0.005, 0.0001, 0.0005], number of neurons in hidden layer with values of [248, 496, 744, 992, 1984], weight_decay with values of [0.001, 0.005, 0.0001] and Dropout with values of [0.5, 0.7, 0.8]. We yield the best area under receiver operating characteristic (ROC) curve (AUC) when learning rate = 0.0001, number of neurons = 744, weight_decay = 0.005, and Dropout = 0.8, which are finally used in our DimiG model.

As shown Figure 2B, DimiG yields the AUCs of 0.748, 0.729, 0.707, and 0.707 on GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. It yields the best performance on GTEx, which covers 53 tissues and thus contains more complete tissues. Compared with GTEx, DimiG achieves a lower performance on GSE43520. It is presumably because GSE43520 only covers four tissues and many tissues associated with certain diseases are missing.

When expression profiles from fewer tissues represent node features, DimiG easily suffers from noises that may be caused by sequencing errors. Of more interest, the area of the ROC curve with low false-positive rate for GTEx is much bigger than those for other three datasets; this region is especially important for evaluating predictive models. The results indicate that expression profiles

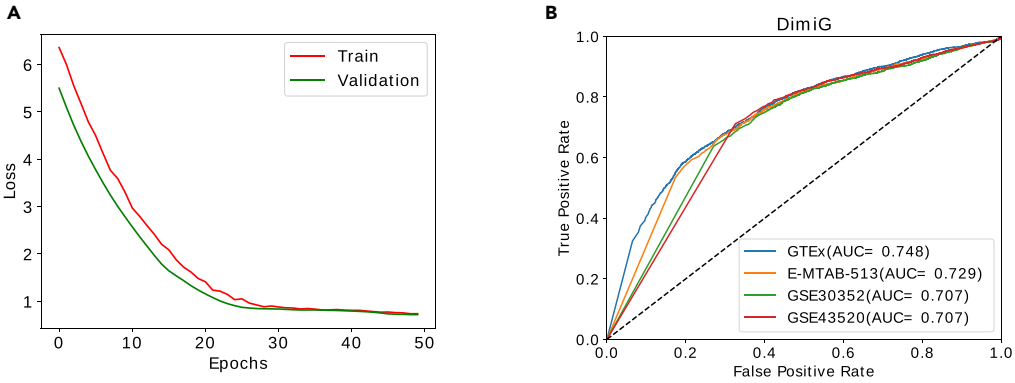



**Figure 2. The Training and Performance of DimiG**

(A) The training and validation loss change with the number of epochs on GTEx dataset.

(B) The ROC curve of DimiG using expression profiles from different datasets, where GTEx, E-MTAB-513, GSE30352, and GSE43520 cover 53, 16, 6, and 4 human tissues, respectively. We train DimiG for each expression dataset separately.

across more tissues as node features can improve the prediction performance for disease-miRNA associations.

## Comparing DimiG with Other Baseline Methods that Do Not Use Disease-miRNA Associations for Model Training

In DimiG, tissue expression profiles from GTEx are used as node features. Its variant DimiG-I just uses the one-hot encoding as the node features. As shown in Figure 3A, the proposed final DimiG yields an AUC of 0.748, which is an increase of 6% over AUC 0.706 of DimiG-I. These results demonstrate that expression profiles across tissues are very informative for inferring disease-miRNA associations. Another variant DimiG-C combines the expression profiles and one-hot encoding of nodes as node features and yields an AUC of 0.728, which is better than DimiG-I, but still worse than DimiG. These results demonstrate that simply concatenating the expression profile and one-hot encoding of nodes not only introduces computational burden but also may decrease the prediction performance. One possible reason is that the one-hot encoding of nodes has no correlations with the expression profiles, making the GCN unable to encode the node features well.

We also compare DimiG with another published method miRPD, which provides association scores between diseases and miRNAs derived from three different sources of miRNA-PCG interactions. As shown in Figure 3A, DimiG outperforms the AUC 0.673 of miRPD-T by 11.1%. Of the three miRPD-based methods, miRPD -T yields the best AUC 0.673, and it is based on the miRNA-PCGs predicted by TargetScan. The miRPD-C and miRPD-M yield similar AUC, which is much worse than miRPD-T. One potential reason is that miRPD-C infers association scores only based on a small number of verified miRNA-PCG. miRPD-M uses a similar number of miRNA-PCG interactions, but it may suffer higher false-positives than TargetScan. To make a fair comparison, we rerun the miRPD method using the same miRNA-PCG interactions and disease-PCG associations as DimiG. As shown in Figure 3, miRPD-I-sum and miRPD-I-max yield an AUC of 0.682 and 0.716, respectively, and both are superior to miRPD-T, miRPD-C, and miRPD-M, but they still perform worse than DimiG.

In addition, our results show that the coexpression-based coding-non-coding co-expression (CNC) method yields the performance AUC = 0.518, suggesting the co-expressed PCGs with miRNAs are not enough for identifying high-confidence disease-miRNA associations. It is because CNC is only based on expression profiles and interaction information of miRNAs is completely ignored.

We also calculate the area under precision-recall curve (AUPRC) of different methods. As shown in Figure 3B, DimiG yields the best AUPRC of 0.765, which is also better than its two variants DimiG-I and DimiG-C. In addition, compared with state-of-the-art method miPRD, the miRBD-T yields an AUPRC of 0.708, which is ~9% worse than the AUPRC 0.765 of DimiG. However, miRPD-I-sum and miRPD-I-max yield much lower AUPRCs than other methods.

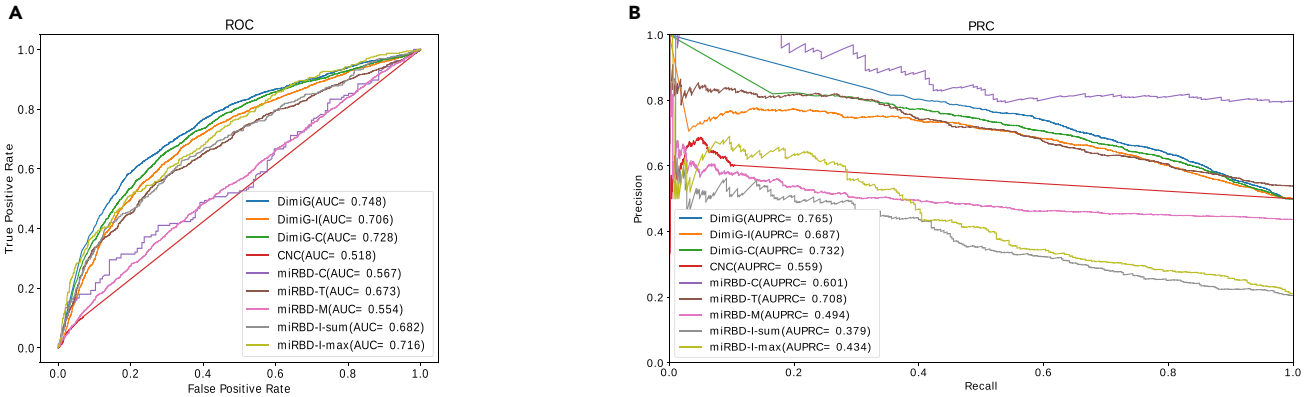



**Figure 3. The Performance of DimiG Using Expression Profiles as Node Features from GTEx and Baseline Methods**

(A) ROC curve.

(B) Precision-recall curve. AUC is the area under ROC curve, and AUPRC is the area under precision-recall curve.

All the above results show that DimiG is are able to achieve better performance for inferring disease-miRNA associations, which is not a surprise because GCNs can better integrate interaction network data and tissue expression profiles, can operate on graphs similarly to CNNs on images, and can take the features and connectivity of nearby nodes into account.

## Performance Comparison of DimiG on the Unseen Disease-miRNA Set

We also compare DimiG with other two state-of-the-art supervised methods BNPMDA and DRMDA on the unseen disease-miRNA set; both BNPMDA and DRMDA use disease-miRNA associations for model training. As shown in Figure 4, on this unseen disease-miRNA set, DimiG yields an AUC of 0.710 and an AUPRC of 0.724, which are better than the performance of BNPMDA with an AUC of 0.686 and an AUPRC of 0.698 and the performance of DRMDA with an AUC of 0.708 and an AUC of 0.715. The results indicate that DimiG outperforms the state-of-the-art supervised methods on inferring new disease-miRNA associations. The AUC 0.687 of BNPMDA on this unseen disease-miRNA set is lower than the reported 5-fold cross-validation AUC of 0.898. Similarly, the AUC 0.708 of DRMDA is also lower than the reported 5-fold cross-validation AUC of 0.916. It should be noted that there still exist some possible dependent disease-miRNA pairs with the training set derived from HMDD v2.0. The results demonstrate that the methods trained on disease-miRNA associations may yield biased cross-validation performance, which could not reflect the methods' actual ability to predict unseen miRNA-disease associations and generalize well to new miRNA-disease associations, as observed in Park and Marcotte (2012). As DimiG does not use any disease-miRNA associations, the performance of DimiG (an AUC of 0.710) on this unseen disease-miRNA set is consistent with the performance (an AUC of 0.748) on the full independent test set constructed from HMDD v3.0. The results indicate that the reported performance reflects DimiG's ability to infer new disease-miRNA associations.

## Case Studies

We present three case studies of miRNAs associated with prostate cancer, lung cancer, and IBD to demonstrate the applicability of DimiG for inferring disease-associated miRNAs. The top predicted candidates for these three diseases are checked with verified associations from literature and public databases, HMDD v3.0 and dbDEMC v2.0. In addition, we evaluate the disease-specific prediction performance of DimiG.

### *Prostate Cancer*

We first investigate the prediction of prostate cancer-associated miRNAs from DimiG. Of the 1,034 miRNAs, top 30 miRNA candidates predicted by DimiG are given in Table 1. Eighteen miRNAs are supported by the literature or database dbDEMC v2.0 and five miRNAs are used as biomarker for detecting prostate cancer. For example, a meta-analysis shows that the first miRNA miR-939 and the eighth miRNA miR-661 are downregulated and the ninth miRNA miR-637 is upregulated in recurrent prostate cancer (Pashaei et al., 2017). This study also finds that prostate cancer-associated CTNNB1 (Anastas and Moon, 2013)

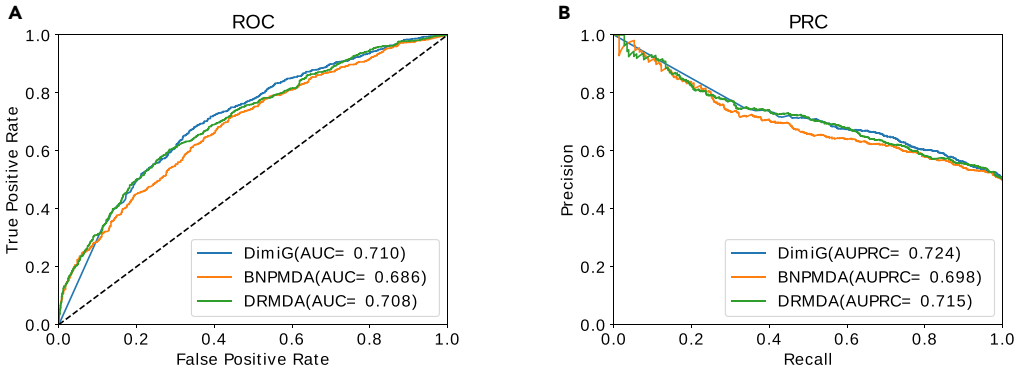



**Figure 4. The Performance of DimiG, BNPMDA, and DRMDA on the Unseen Disease-miRNA Dataset**

Here, DimiG uses expression profiles as node features from GTEx; BNPMDA and DRMDA are trained on known disease-miRNA associations.

(A) ROC curve.

(B) Precision-recall curve.

is one hub gene for its interacting targets in a gene network. That is to say, the predicted miRNA genes by DimiG in Table 1 are well consistent with existing knowledge.

In another study, the second miRNA miR-93 is frequently overexpressed in prostate cancer and downregulates capicua levels (Choi et al., 2015). In dbDEMC v2.0, three miRNAs miR-874, miR-766, and miR-625 are differentially expressed in prostate cancer. Of them, miR-625 and miR-874 share the same gene target HMGA1 in prediction channel of RAIN database with miR-765. The remaining three miRNAs in the top 10 candidates either regulate prostate cancer-associated genes or activate prostate cancer-associated pathway. We have also noted that of the top 10 miRNAs, only miR-92a and miR-765 are recorded in HMDD v3.0, and the others are not. These results indicate that DimiG can infer novel disease-miRNA associations currently not in the curated databases. Of the remaining 20 miRNAs, eight miRNAs are supported to be associated with prostate cancer by literature and five miRNAs are used as biomarkers for detecting prostate cancer in a filed patent. The results indicate that DimiG is powerful in identifying disease-associated miRNAs.

As shown in Figure 5A, of the 176 prostate cancer-associated miRNAs in HMDD v3.0, 21 are in the top 50 predicted candidates by DimiG. Six of the 15 prostate cancer-associated miRNAs in dbDEMC v2.0 belong to the top 50 candidates predicted by DimiG (Figure S3A). On the prostate cancer-specific dataset constructed from HMDD v3.0, DimiG yields an AUC of 0.724, 0.697, 0.675, and 0.664 on GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively (Figure 5B). Similarly, on the dataset constructed from dbDEMC v2.0, DimiG achieves an AUC of 0.844, 0.836, 0.729, and 0.684 on GTEx, E-MTAB-513, GSE30352, and GSE43520 for prostate cancer, respectively (Figure S3B). The performance is better than that on the dataset constructed from HMDD v3.0 because the extracted disease-miRNA associations are experimentally verified using low-throughput methods and are more reliable. We can observe that more tissues can provide informative clues for predicting prostate cancer-associated miRNAs; even some tissues may be considered not relevant to prostate cancer. The results further demonstrate the power of DimiG.

We further investigate the predicted miRNA candidates using verified prostate cancer-associated miRNAs from miRCancer and PhenomiR. Of the top 50 predicted miRNAs by DimiG, 10 miRNAs are supported by miRCancer (Figure S4A) and 16 miRNAs are supported by PhenomiR (Figure S5A). On the prostate cancer-specific set derived from miRCancer, as shown in Figure S4B, DimiG yields an AUC of 0.755, 0.743, 0.690, and 0.695 using GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. As demonstrated in Figure S5B, for the prostate cancer-specific set collected from PhenomiR, DimiG achieves an AUC of 0.723, 0.713, 0.645, and 0.653 using GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. DimiG achieves similar results on the two databases as HMDD v3.0.

### *Lung Cancer*

We investigate the prediction ability of DimiG for lung cancer-associated miRNAs. There are 172 such miRNAs recorded in the HMDD v3.0 for lung cancer. As shown in Figure 5C, of the 172 miRNAs, 12 miRNAs

| Rank | MiRNA Name | Reason | Support Evidence |
|---|---|---|---|
| 1 | miR-939 | Downregulated | PMID:28651018 |
| 2 | miR-93 | Overexpressed in prostate cancer, and downregulates capicua levels to promote prostate cancer progression | PMID:26124181 |
| 3 | miR-92a | Regulates tumor suppressor PTEN | PMID:24098737 |
| 4 | miR-874 | Upregulated | dbDEMC v2.0 |
| 5 | miR-766 | Downregulated | dbDEMC v2.0 |
| 6 | miR-765 | Mimics the expression of oncogenic HMGA1 in prostate cancer | PMID:24837491 |
| 7 | miR-744 | Promotes prostate cancer progression by activating Wnt/β-catenin pathway | PMID:28107193 |
| 8 | miR-661 | Downregulated | PMID:28651018 |
| 9 | miR-637 | Upregulated | PMID:28651018 |
| 10 | miR-625 | Upregulated | dbDEMC v2.0 |
| 11 | miR-608 | Target overexpressed FLOT1 in prostate cancer | PMID:28549468 |
| 12 | miR-5196 | – | – |
| 13 | miR-5193 | Downregulates TRIM11 in prostate cancer | PMID:30608062 |
| 14 | miR-504 | Represses overexpressed FOXP1 in prostate cancer | PMID:23022474 |
| 15 | miR-5011 | – | – |
| 16 | miR-491 | Interact with PDGFRA to reduce prostate cancer cell migration | PMID:29312807 |
| 17 | miR-486 | A prostate cancer driver | PMID:29069829 |
| 18 | miR-4739 | Biomarker for prostate cancer | Patent: WO2015190584A1 |
| 19 | miR-4731 | – | – |
| 20 | miR-4728 | Target MST4 involved in prostate cancer progression | PMID:25950472 |
| 21 | miR-4726 | Biomarker for prostate cancer | Patent: WO2015190584A1 |
| 22 | miR-4725 | Biomarker for prostate cancer | Patent: WO2015190584A1 |
| 23 | miR-4723 | Regulate Abl kinases in prostate cancer | PMID:24223753 |
| 24 | miR-4716 | – | – |
| 25 | miR-4667 | Biomarker for prostate cancer | Patent: WO2015190584A1 |
| 26 | miR-4644 | – | – |
| 27 | miR-455 | Target eIF4E as tumor suppressor in prostate cancer | PMID:28350134 |
| 28 | miR-4505 | Biomarker for prostate cancer | Patent: WO2015190584A1 |
| 29 | miR-4498 | – | – |
| 30 | miR-4447 | – | – |

**Table 1. The Top 30 Candidate Prostate Cancer-Related miRNAs Predicted by DimiG and Their Support Evidences in Literature**

– Means no support evidence.

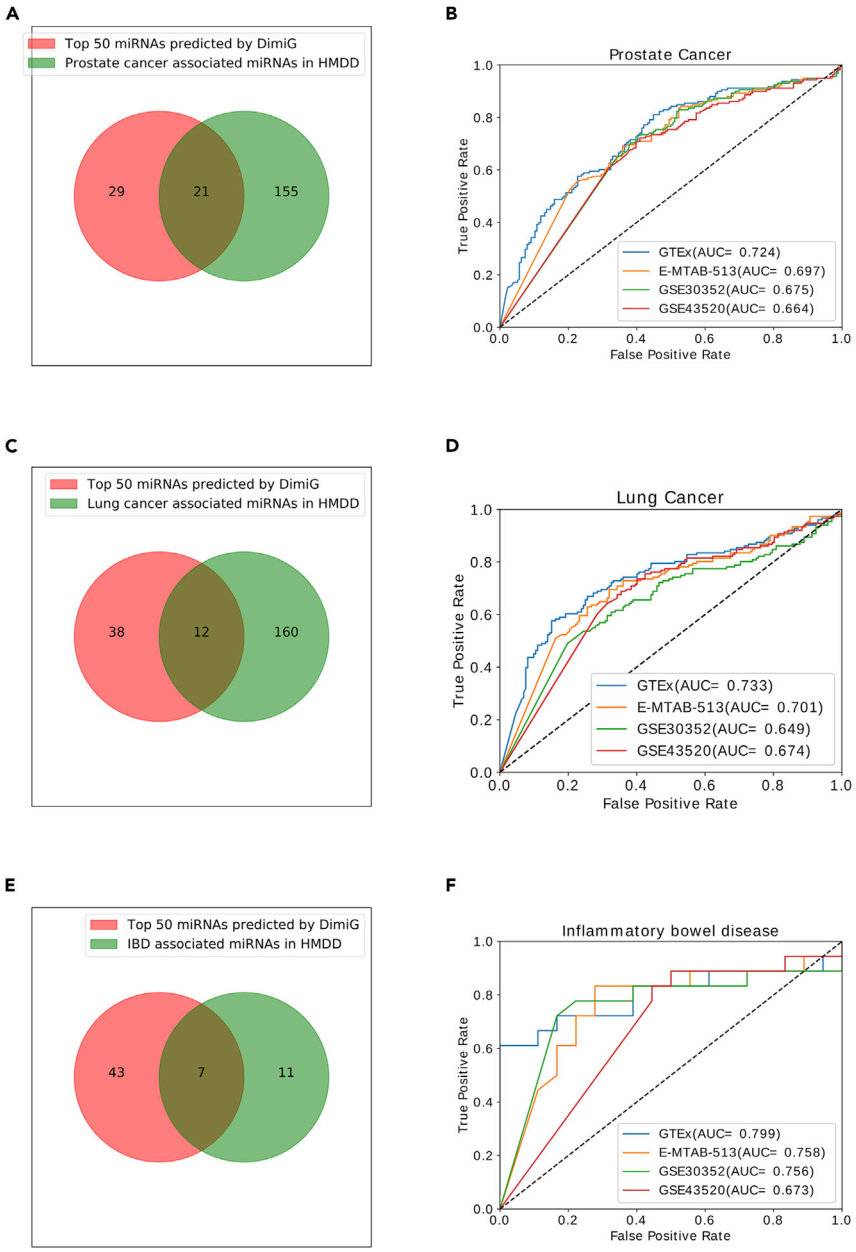



**Figure 5. Venn Diagram and ROC Curve for Predicting Associated miRNAs for Prostate Cancer, Lung Cancer, and IBD Using DimiG**

(A) The overlap between the top 50 predicted miRNAs by DimiG and prostate cancer-associated miRNAs in HMDD v3.0.
(B) ROC curve for predicting prostate cancer associated miRNAs using four tissue expression datasets.
(C) and (D) (C) Venn diagram and (D) ROC for lung cancer.
(E) and (F) (E) Venn diagram and (F) ROC for inflammatory bowel disease (IBD)

are in the top 50 lung cancer-associated miRNAs predicted by DimiG. Then we evaluate the prediction performance on the dataset consisting of 172 lung cancer-associated miRNAs and 172 miRNAs not associated with lung cancer in HMDD v3.0. As shown in Figure 5D, DimiG yields an AUC of 0.733, 0.701, 0.649, and 0.674 for GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively.

We further report the prediction ability on another lung cancer-specific dataset derived from dbDEMC v2.0. In this dataset, there are 16 miRNAs associated with lung cancer and another 16 miRNAs not

associated with lung cancer. Of the 16 miRNAs associated with lung cancer, one is in the top 50 predicted miRNAs by DimiG (Figure S3C). As shown in Figure S3D, DimiG achieves an AUC of 0.925, 0.808, 0.806, and 0.869 for GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. The results also indicate that informative clues can be captured in more tissues for predicting lung cancer-associated miRNAs.

According to miRCancer database, 60 miRNAs are associated with lung cancer. Of the 60 miRNAs, three are in the top 50 miRNAs predicted by DimiG (Figure S4C). On the lung cancer-specific set derived from miRCancer (Figure S4D), DimiG yields an AUC of 0.778, 0.740, 0.775, and 0.705 using GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. Based on PhenomiR, 190 miRNAs are associated with lung cancer. Of them, 10 miRNAs are in the top 50 candidates predicted by DimiG (Figure S5C). In addition, as shown in Figure S5D, DimiG obtains an AUC of 0.781, 0.751, 0.757, and 0.754 using the four expression datasets GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. All the above results show that DimiG achieves promising performance for lung cancer.

### *Inflammatory Bowel Disease*

Beside cancer, we also investigate DimiG's prediction ability for another non-cancer disease IBD. In this study, we predicted associated diseases for 1,034 miRNAs, of which 18 miRNAs are deposited for IBD. Seven of the 18 miRNAs are in the top 50 miRNAs predicted by DimiG (Figure 5E). As shown in Figure 5F, DimiG yields an AUC of 0.799, 0.758, 0.756, and 0.673 for GTEx, E-MTAB-513, GSE30352, and GSE43520, respectively. The results show that we can use DimiG for other non-cancer diseases.

## DISCUSSION

In this study, we present a semi-supervised multi-label GCN framework to integrate heterogeneous networks of tissue expression profiles, miRNA-PCG interactions, PCG-PCG interactions, and disease-PCG interaction to infer disease-associated miRNAs. The whole pipeline is under the context of interaction network, where disease-miRNA associations are completely not involved in model training. CNN cannot directly process the non-Euclidean domain data, like network data. However, GCN can handle these types of data; they are specially designed to extract abstract features from network data. To prove that, we use T-distributed Stochastic Neighbor Embedding (T-SNE) to map the learned node features in the last hidden layer by the GCN and the original node features from expression profiles in GTEx to 2D space. As shown in Figure S6, the learned node features by GCN have a better shape than original node features. We demonstrate that cross-validating the methods trained on disease-miRNA associations yields an overestimated performance. Our results demonstrate that DimiG outperforms other state-of-the-art methods, which do not require disease-miRNA association information, and two methods trained on disease-miRNA associations.

In this study, we need to set several cutoff values (Figure S1) for PCG-PCG interactions from STRING database, PCG-miRNA interactions from RAIN database, and disease-PCG associations from DISEASES database. All the three databases are developed by the same group, they use similar data quality control, and the confidence values are all scored using the similar pipeline from multiple channels, including experiments, knowledge, text mining, and prediction. Thus the constructed graph should follow GCN's assumption that it is a simple one-modal graph, in which all nodes are of the same type (all nodes are genes) and all edges have the same semantic meaning (Kipf and Welling, 2017).

As shown in Figure S7, for the low cutoff values of STRING with 300 and RAIN with 0.10, DimiG yields much lower performance with AUC 0.653, it is because lower cutoff may introduce more false-positive interactions. For cutoff value of STRING greater than 400 and cutoff value of RAIN greater than 0.15, DimiG yields similar AUCs. DimiG yields the best AUC 0.754 at cutoff values 400 and 0.20 for STRING and RAIN, respectively, but these higher cutoff values will lead to fewer miRNAs. Thus, we use cutoff value 400 of STRING and 0.15 of RAIN in this study, and DimiG yields an AUC 0f 0.748 and is capable of finding more miRNAs, which is just a little lower than 0.753 with cutoff value 400 and 0.20 for STRING and RAIN databases, respectively, as the trade-off. We also evaluate the impact of confidence thresholds 1.5 and 2.5 for disease-PCG associations on the performance of DimiG. As shown in Figure S8, DimiG yields an AUC of 0.719 and 0.742 for thresholds 1.5 and 2.5, respectively; both are lower than 0.748 when using threshold 2. The possible reason is that lower confidence threshold 1.5 introduces more false-positives for model training and higher threshold 2.5 makes the number of training samples much fewer, which are both not good for training machine learning model. In this study, we use PCG-PCG interactions from STRING v10 instead of STRING v11, because both RAIN and DISEASES

databases are based on STRING v10. Some gene identifiers are changed between STRING v10 and v11. We directly use PCG-PCG interactions from STRING v11 for DimiG, only 5,092 PCGs are kept for constructing the interaction network, and DimiG achieves a lower AUC of 0.718.

There are other computational models with reported cross-validation AUC over 0.8, in which disease-miRNA associations are involved in model training. We demonstrate that cross-validation could report overoptimistic performance of the methods and could not generalize well to unseen disease-miRNA associations. Disease-miRNA associations can be incorporated into model training; however, randomly dividing the disease-miRNA pairs into the training and test sets for cross-validation could be biased. To better evaluate the performance of one model, a strictly independent test set should be at least constructed; e.g., the model is trained on all data published up to a specific year and predictions are evaluated on data published after that, or the model is trained on data in the older version of database and evaluated on those new added disease-miRNA associations in the updated database. In addition, DimiG predicts associated miRNAs for 248 diseases in one model. Many previous methods formulate the disease-miRNA prediction as binary classification problems, and they require constructing negative disease-miRNA associations, which may introduce false-negatives into model training.

### Limitation of the Study

In this study, we used only expression profiles across tissues as node features. Some studies have revealed that functional domain information can assist identifying disease-associated miRNAs (Yang et al., 2018). In future work, we can combine the gene ontology (GO) information and expression values across tissues into the node features, or ensemble the two GCN models trained on each representation, which is expected to further improve the prediction performance of DimiG.

DimiG does not require disease-miRNA associations for model training, but it requires the miRNA-PCG interactions to construct the graph. Each miRNA must have at least one interacting PCG; all nodes, including miRNAs and PCGs, need be present during the training. Thus the trained node embedding can be propagated to miRNAs and further used for inferring miRNA-associated diseases. This precondition makes us to discard some miRNAs, and DimiG cannot infer associated diseases for these miRNAs without interacting PCGs. In addition, more and more novel miRNAs are being discovered, and their interactions with PCGs may not be readily available. Thus the trained models cannot be generalized to miRNAs not in the graph. Luckily, some recent GCN models have tried to solve this issue, which are trained on a set of nodes and generalized to any augmentation of the graph (Hamilton et al., 2017). This inductive GCN model can be applied for novel miRNAs not in the graph in our future study.

### Conclusion

In this study, we present a semi-supervised multi-label learning framework DimiG to integrate interaction data for inferring miRNA-associated diseases. DimiG does not use any disease-miRNA associations for model training. This new approach achieves promising performance and outperforms other baseline methods not trained on disease-miRNA associations with a large margin on our benchmark dataset. We observe that cross-validation performance of methods trained on known disease-miRNA associations could be overestimated and could not reflect their actual abilities for inferring new disease-miRNA associations. Our results demonstrate that the tissue expression profiles can provide informative signals for inferring disease-miRNA associations. We expect DimiG to be used to discover novel miRNA biomarkers for diseases and that the framework can be extended to other tasks based on network data, e.g., functional annotations of proteins.

## METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

## DATA AND CODE AVAILABILITY

All the data and code are available at https://github.com/xypan1232/DimiG or http://www.csbio.sjtu.edu.cn/bioinf/DimiG.

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.09.013.

## ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910500), the National Natural Science Foundation of China (No. 61903248, 61725302, 61671288, 91530321, 61603161), and the Science and Technology Commission of Shanghai Municipality (No. 16ZR1448700, 16JC1404300, 17JC1403500).

## AUTHOR CONTRIBUTIONS

Conceptualization, X.P. and H.-B.S.; Methodology: X.P.; Writing – Original Draft, X.P.; Writing – Review & Editing, X.P. and H.-B.S.; Funding Acquisition, H.-B.S.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 11, 2019
Revised: September 5, 2019
Accepted: September 11, 2019
Published: October 25, 2019

## REFERENCES

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Anastas, J.N., and Moon, R.T. (2013). WNT signalling pathways as therapeutic targets in cancer. Nat. Rev. Cancer *13*, 11–26.

Angermueller, C., Parnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. *12*, 878.

Antanaviciute, A., Daly, C., Crinnion, L.A., Markham, A.F., Watson, C.M., Bonthron, D.T., and Carr, I.M. (2015). GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. Bioinformatics *31*, 2728–2735.

Baker, M.A., Davis, S.J., Liu, P.Y., Pan, X.Q., Williams, A.M., Iczkowski, K.A., Gallagher, S.T., Bishop, K., Regner, K.R., Liu, Y., et al. (2017). Tissue-specific microRNA expression patterns in four types of kidney disease. J. Am. Soc. Nephrol. *28*, 2985–2992.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-Disease Association prediction. PLoS Comput. Biol. *13*, e1005912.

Chen, X., Gong, Y., Zhang, D.H., You, Z.H., and Li, Z.W. (2018a). DRMDA: deep representations-based miRNA-disease association prediction. J. Cell Mol. Med. *22*, 472–485.

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H., and Liu, H. (2018b). BNPMDA: bipartite network projection for MiRNA-disease association prediction. Bioinformatics *34*, 3178–3186.

Chen, X., Xie, D., Zhao, Q., and You, Z.H. (2019). MicroRNAs and complex diseases: from experimental results to computational models. Brief Bioinform. *20*, 515–539.

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Comput. Biol. *14*, e1006418.

Chen, X., Liu, M.X., and Yan, G.Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. Mol. Biosyst. *8*, 2792–2798.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface *15*, https://doi.org/10.1098/rsif.2017.0387.

Choi, N., Park, J., Lee, J.S., Yoe, J., Park, G.Y., Kim, E., Jeon, H., Cho, Y.M., Roh, T.Y., and Lee, Y. (2015). miR-93/miR-106b/miR-375-CIC-CRABP1: a novel regulatory axis in prostate cancer progression. Oncotarget *6*, 23533–23547.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. J. Univers. Comput. Sci. *19*, 3844–3852.

Friedman, R.C., Farh, K.K.H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. *19*, 92–105.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

Hamilton, W.l., Ying, R., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. 31st Conference on Neural Information Processing Systems (NIPS 2017).

Jia, L., Zhang, Z., Wang, L., Jiang, W.M., and Zhao, M.B. (2016). Adaptive neighborhood propagation by joint L2,1-norm regularized sparse coding for representation and classification. IEEE 16th International Conference on Data Mining (ICDM), 201–210, https://doi.org/10.1109/ICDM.2016.0031.

Jiang, Q.H., Hao, Y.Y., Wang, G.H., Juan, L.R., Zhang, T.J., Teng, M.X., Liu, Y.L., and Wang, Y.D. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst. Biol. *4* (*Suppl 1*), S2.

Kipf, T.N., and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations (ICLR-17).

Lan, W., Wang, J., Li, M., Liu, J., Wu, F.X., and Pan, Y. (2018). Predicting MicroRNA-disease associations based on improved MicroRNA and disease similarities. IEEE/ACM Trans. Comput. Biol. Bioinform. *15*, 1774–1782.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324.

Lehtinen, S., Lees, J., Bahler, J., Shawe-Taylor, J., and Orengo, C. (2015). Gene function prediction from functional association networks using kernel partial least squares regression. PLoS One *10*, e0134668.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., Rheinheimer, S., Meder, B., Stahler, C., Meese, E., et al. (2016). Distribution of miRNA expression across human tissues. Nucleic Acids Res. *44*, 3865–3877.

Mork, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J., and Jensen, L.J. (2014). Protein-driven inference of miRNA-disease associations. Bioinformatics *30*, 392–397.

Pan, X.Y., and Shen, H.B. (2016). OUGENE: a disease associated over-expressed and under-expressed gene database. Sci. Bull. *61*, 752–754.

Pan, X., and Shen, H.B. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics *18*, 136.

Pan, X., and Shen, H.B. (2018). Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics *34*, 3427–3436.

Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics *19*, 511.

Pan, X., Jensen, L.J., and Gorodkin, J. (2019). Inferring disease-associated long non-coding RNAs using genome-wide tissue expression profiles. Bioinformatics *35*, 1494–1502.

Park, Y., and Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. Nat. Methods *9*, 1134–1136.

Pashaei, E., Pashaei, E., Ahmady, M., Ozen, M., and Aydin, N. (2017). Meta-analysis of miRNA expression profiles for prostate cancer recurrence following radical prostatectomy. PLoS One *12*, e0179543.

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833–D839.

Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. Methods *74*, 83–89.

Tang, C., Zhou, H., Zheng, X., Zhang, Y., and Sha, X. (2019). Dual Laplacian regularized matrix completion for microRNA-disease associations prediction. RNA Biol. *16*, 601–611.

Wan, J.W., and Wang, Y. (2019). Cost-sensitive label propagation for semi-supervised face recognition. IEEE Trans. Inf. Foren. Sec. *14*, 1729–1743.

Xuan, P., Dong, Y., Guo, Y., Zhang, T., and Liu, Y. (2018a). Dual convolutional neural network based method for predicting disease-related miRNAs. Int. J. Mol. Sci. *19*, https://doi.org/10.3390/ijms19123732.

Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., Zhang, Z., and Ding, J. (2015). Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics *31*, 1805–1815.

Xuan, P., Shen, T., Wang, X., Zhang, T., and Zhang, W. (2018b). Inferring disease-associated microRNAs in heterogeneous networks with node attributes. IEEE/ACM Trans. Comput. Biol. Bioinform. https://doi.org/10.1109/TCBB.2018.2872574.

Yang, Y., Fu, X., Qu, W., Xiao, Y., and Shen, H.B. (2018). MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association. Bioinformatics *34*, 3547–3556.

You, Z.H., Huang, Z.A., Zhu, Z., Yan, G.Y., Li, Z.W., Wen, Z., and Chen, X. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.

Zhang, Z., Li, F.Z., Jia, L., Qin, J., Zhang, L., and Yan, S.C. (2018). Robust adaptive embedded label propagation with weight learning for inductive classification. IEEE Trans. Neur. Net. Lear. *29*, 3388–3403.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics *34*, i457–i466.

Zoidi, O., Tefas, A., Nikolaidis, N., and Pitas, I. (2018). Positive and negative label propagations. IEEE Trans. Circ. Syst. Vid. *28*, 342–355.

**ISCI, Volume *20***

## Supplemental Information

# Inferring Disease-Associated MicroRNAs Using Semi-supervised Multi-Label Graph Convolutional Networks

**Xiaoyong Pan and Hong-Bin Shen**

# Supplemental information

## Transparent Methods

We collect genome-wide tissue expression profiles from RNA-seq data, disease-associated PCGs, PCG-PCG interactions, miRNA-PCG interactions. These data are fed into a semi-supervised multi-label GCN using PCGs as training and validating. Then the trained GCN model is propagated into miRNAs to score their associations with diseases.

### 1 Data sources

One recent benchmark study (Huang et al., 2018) shows that among 21 widely used protein-protein interaction databases, STRING (Szklarczyk et al., 2015)is one of the three databases having the best performance for discovery of disease genes. Like the STRING database, miRNA-gene interaction database RAIN(Junge et al., 2017) and disease-gene association database DISEASES(Pletscher-Frankild et al., 2015) also integrate different sources of data, including text mining, knowledge, experiments and predictions using similar techniques and scoring schema, and they are developed by the same group and use the same gene identifiers. For disease-miRNA associations, the widely used HMDD database(Li et al., 2014) is used as the benchmark set for evaluating disease-miRNA association prediction.

#### 1.1 Tissue expression data

We download the GTEx tissue expression data (GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8) across 53 tissues (Lonsdale et al., 2013). It contains 19,732 PCGs and 2,833 miRNA genes. Of the 2,833 miRNAs, only 1,244 miRNAs are in miRBase v20 (Kozomara and Griffiths-Jones, 2011). As a comparison, we also collected three raw RNA-seq datasets, E-MTAB-513 across 16 tissues (Derrien et al., 2012), GSE43520 across four tissues (Necsulea et al., 2014) and GSE30352 across six tissues (Brawand et al., 2011). The raw reads are mapped to the same reference genome as GTEx using STAR version 2.5.0b (Dobin et al., 2013) and quantified using Cufflinks (Trapnell et al., 2014). For a tissue with multiple samples, we calculate the median expression values for this tissue. The expression values are log-transformed using $\log_2(1+x)$, which is further normalized across tissues by the fraction of expression of one gene in one tissue relative to the sum of its expression in all tissues. In addition, we add another feature of the sum of one gene's expression across all tissues.

#### 1.2 Disease-PCG associations

We download the integrated disease-PCG associations from DISEASES database (Pletscher-Frankild et al., 2015). DISEASES database has multiple channel of evidences, e.g. knowledge, experiments and text mining, to support the associations. Each disease-PCG association is assigned a confidence score. In this study, we only use associations with confidence scores greater or equal to 2. All gene names are mapped to Ensembl gene identifiers and diseases are represented in Disease Ontology ID (Schriml et al., 2012). In total, we obtain 86,430 associations between 4,161 diseases and 9,636 PCGs. Figure S2A shows the disease label distribution for PCGs.

#### 1.3 Disease-miRNA associations

We extract the disease-miRNA associations from database HMDD v3.0 (Li et al., 2014), which has 32,281 associations between 1,102 miRNA genes and 850 diseases. After removing some associations whose disease name have no DOID, we build a set with 24,320 unique disease-miRNA associations

between 1,007 miRNAs and 582 diseases. We further map the miRNA names into Ensembl gene identifiers. In the end, we obtain 6,829 associations between 548 miRNAs and 486 diseases. Figure S2B illustrates the disease label distribution for miRNAs.

**1.4 PCG-PCG interactions**

The human gene-gene interactions are extracted from widely used database STRING v10 (Szklarczyk et al., 2015), which houses millions of gene-gene interactions across multiple species. In STRING, each interaction is scored according to multiple evidences, including experiments, knowledge, prediction and text mining. The higher the score is, the more reliable this interaction is. We mapped all gene names to Ensembl gene ID and only select those interactions with confidence score greater than 400 (confidence values are between 1 and 1000), which is the medium confidence score in STRING database. In total, we obtain 1,481,757 interaction pairs of 18,883 PCGs.

**1.5 miRNA-PCG interactions**

We extract the human miRNA-PCG interactions from RAIN database (Junge et al., 2017), which integrates miRNA-gene interactions from text mining, experiments, knowledge and predictions. In total, RAIN scores 46,472 human miRNA-gene interactions according to different evidences. In this study, we only select those interactions with combined score greater than 0.15, which is used as default cut-off in the webserver. We mapped all miRNA names to Ensembl gene ID. In total, we obtained 173,662 interaction pairs between 17,686 PCGs and 1,725 miRNAs.

**2 Data processing**

We integrate the above tissue expression profiles, PCG-PCG interactions, miRNA-PCG interactions and PCG-disease associations to score diseases and miRNAs. When constructing the DimiG, we process the data as the following: 1) when constructing a graph, each PCG has at least one interacting PCG and each miRNA has at least one interacting PCG; 2) both the PCGs and miRNAs in the graph have expression profiles across tissues as node features; 3) the PCGs in the graph should be associated with at least one disease, since diseases are the labels of PCG nodes and they are used to calculate the training loss for model training. The whole processing is shown in Figure S1. Finally, we obtained 7,222 genes with 6,188 PCGs and 1,034 miRNAs, and 248 diseases as the labels.

Here, we use the experimentally verified miRNA-disease associations to evaluate the performance of DimiG. After removing those associations whose disease is not among the 248 diseases, we obtain 2,695 associations for 91 diseases. As negative control examples, for each disease, we randomly select the same number of miRNAs not associated with the disease in HMDD v3.0 as we have miRNAs associated with it. In the end, we obtain a dataset with 5,390 disease-miRNA pairs, where half of them are positives and the other half are negatives. This dataset is balanced not only overall but also for each disease, and is used as independent test set, whose association pairs are not involved in model training.

**3 Graph Convolutional Networks**

GCN (Kipf and Welling, 2017) is trained in an end-to-end way and learns the informative node embedding for the semi-supervised classification tasks on a graph. The GCN is different from previous network embedding methods, e.g. node2vec (Grover and Leskovec, 2016), which needs firstly to learn

node embedding from a graph, and then the learned embedding is fed into a supervised classifier for downstream classification tasks.

Given a graph G and additional node feature vectors *X*:

1. A node feature matrix *X*: a $N \times T$ feature matrix, where *N* is the number of nodes in the graph G, and T is number of input features. In DimiG, for each node (gene), its expression values across 53 tissues from GTEx and the sum of its expression values across the 53 tissues are used as features, thus each node is represented as a 54-dimensional vector.
2. An adjacency matrix *A*: a $N \times N$ matrix, showing the connections between nodes in the graph G. If a PCG *i* interact with a miRNA or a PCG *j*, then the value $A_{i,j}$ is 1, otherwise 0.
3. The multi-hot encoded label matrix Y for some nodes but not all nodes in the graph, one node can have multiple labels.

The GCN frames the classifying nodes in a graph as graph-based semi-supervised learning, where not all nodes in the graph need labels. To this end, a graph Laplacian regularization term is introduced into the loss function *L*:

$$L = L_0 + \varepsilon \sum_{ij} A_{ij} \left\| f(X_i) - f(X_j) \right\|^2 = L_0 + \varepsilon f(X)^T (D - A) f(X) \quad (1)$$

where $L_0$is the supervised loss, $f$ can be a neural network-based function, $\varepsilon$ is the factor, $X$ is the node features, $A$ is the adjacency matrix and $D$ is the degree matrix.

A multi-layer GCN is propagated with the following rule:

$$H^{l+1} = f(H^l, A) = ReLU(AH^lW^l) \quad (2)$$

where $H^l$ is the output of layer *l* and $W^l$ is a weight matrix.

The GCN performs spectral convolutions on the graph and perform neighborhood weighting. A first-order approximation of localized spectral filters in Fourier-domain can be used to approximate the propagation rule (Defferrard et al., 2016). Please refer to (Kipf and Welling, 2017) for more details on the GCN.

For semi-supervised classification, we minimize the binary cross entropy loss on labeled data:

$$L(w) = -\sum_{i=1}^{N} y_i log\ (\hat{y}_i) + (1-y_i) log(1 - \hat{y}_i) + \alpha \|w\|_2^2 \quad (3)$$

where $y_i$is the true label , $\hat{y}_i$ is the output probability of the last sigmoid layer and α is the weight factor for regularization.

**4 Multi-label classification**

Multi-label classification assigns multiple labels to one node, and it is categorized into two groups: (1) problem transformation methods and (2) algorithm adaptation methods. For problem transformation methods, they are transformed into multiple binary classification problems, where each binary classifier is trained separately for each label. For algorithm adaptation methods, they usually perform multi-label classification directly using one classifier, which outputs predicted probabilities for all the labels.

In DimiG's multi-label assignment, we use the 2nd approach and set up a sigmoid layer as the last output layer, whose number of neurons is the number of labels. Given a set of labels with *K* classes (248 diseases in this study), each node is assigned with several labels from it. Denote *x* as an input node, and *y* as a multi-hot encoded vector with 248 elements, which indicates absence of class *i* and presence of class *i*: *y* = [0, 1, 0, …, 1].

**6 Baseline methods**

We formulate disease-miRNA association prediction as node classification in a graph, where nodes of PCGs are assigned multiple diseases as labels, and nodes of miRNAs are used as the test set. The miRNA-disease associations are completely not involved in model training of DimiG. To demonstrate the

power of DimiG, we compare with two existing similar baseline methods that do not require disease information for miRNAs, and two variants of our DimiG model:

1. miRPD: it scores miRNA–disease associations by network analysis of miRNA–protein associations and protein–disease associations from text mining (Mork et al., 2014). We downloaded the inferred associations between diseases and miRNAs from their websites. In this study, we downloaded full dataset including croft_full.tsv.gz, miranda_full.tsv.gz and targetscan_full.tsv.gz. The three sets give the association scores between diseases and miRNAs, they are called miRBD-C, miRBD-M and miRBD-T, respectively. To make a fair comparison, we construct other two baseline methods miRBD-I-sum and miRBD-I-max, which use the same integrated miRNA-PCG associations from the RAIN database and disease-PCG associations from DISEASES database as DimiG. miRBD-I-sum and miRBD-I-max use sum and max to calculate the association scores between diseases and miRNAs, respectively.
2. CNC: it is a co-expression-based method under guilt-by-association framework. For each disease *d* and miRNA, CNC consists of the following steps: 1) Calculate Pearson's correlation coefficients (PCC) between D-associated PCGs and the miRNA. 2) Keep only those co-expression PCGs with absolute PCC > 0.3 and *P*-value < 0.01 (Liu and Zhao, 2016). 3) Calculate the mean value of up to K largest PCC value as the score for the miRNA with disease *d*.
3. Two variant methods of DimiG: instead of using expression profiles across tissues as node features, we used one-hot encoding of all nodes. Simply, we use the identity matrix as the node feature matrix *X*. This variant method is called DimiG-I. In addition, we also construct another variant method DimiG-C, which combines the expression profiles and one-hot encoding as node features.

After obtaining the associated scores between diseases and miRNAs, we extracted scores for those positive and negative disease-miRNA pairs in the independent test set, and these scores are pooled together to calculate the area under the receiver operating characteristic curve (AUC) and the area under precision-recall curve (AUPRC). For miRPD, only those disease-miRNA pairs with scores in the download files are kept.

We also compare DimiG with other two state-of-the-art supervised methods BNPMDA (Chen et al., 2018) and DRMDA, both use miRNA-disease associations for model training. BNPMDA trains a bipartite network recommendation method on known disease-miRNA associations by integrating miRNA and disease similarity. DRMDA further uses stacked autoencoder to learn deep representation for diseases and miRNAs, which are fed into a support vector machine for predicting disease-miRNA associations. Both methods are trained on disease-miRNA associations from the database HMDD v2.0.

As mentioned above, cross-validation may yield biased performance. Thus, we evaluate how BNPMDA and DRMDA generalize to unseen miRNA-disease associations in the training set. We download the prediction association scores for available miRNAs and diseases from the supplementary material of these two studies, and the disease names are mapped to the disease ontology ID. Of the 1034 miRNAs and 248 diseases in this study, 304 miRNAs and 55 diseases are in the score matrix predicted by BNPMDA and DRMDA. Instead, the DimiG does not use any disease-miRNA association information for model training, to make a fair and objective comparison, we evaluate the prediction performance of DimiG, BNPMDA and DRMDA on a dataset consisting of newly recorded verified miRNA-disease associations in HDMM v3.0 but not in HMDD v2.0. We call this dataset as an unseen disease-miRNA set, which is freely available at https://github.com/xypan1232/DimiG/tree/master/data. In this data set, there are 1954 disease-miRNA pairs consisting of 977 positives and 977 negatives, and it is also balanced not only overall but also for each disease. It should be noted that this data set may still have some possible dependent associations with the training set derived from HMDD v2.0.

**7 Experimental settings**

In this study, we modify the implemented GCN from https://github.com/tkipf/pygcn to support multi-label classification using PyTorch framework. DimiG consists of two-layer GCNs. The last layer is the sigmoid layer with number of neurons 248, which is the number of diseases as labels. We use Adam to optimize the Binary Cross Entropy between the targets(Diederik and Jimmy, 2015). We use grid search to select the best parameters for learning rate, weight_decay and dropout probability.

Of the 6,188 PCGs, we keep 80% of PCGs as the training set, 20% of PCGs as the validation set and all 1034 miRNAs as the test set, which is used to evaluate the performance of DimiG for predicting disease-associated miRNAs. For each PCG, the label is a 248-dimensional multi-hot vector indicating the presence of diseases associated with this PCG.

**8 Data construction for case studies**

In this study, we further evaluate the DimiG for predicting associated miRNAs for three types of diseases, i.e., prostate cancer, lung cancer and inflammatory bowel disease (IBD). For prostate cancer, of all the 1034 miRNAs, 176 miRNAs are associated with prostate cancer in HMDD v3.0, which are used as the positives. For negative control, we randomly select other 176 miRNAs not associated with prostate cancer in the HMDD v3.0 as negatives. The 176 positives and 176 negatives comprise the prostate cancer specific dataset. Then we pool the predicted scores by DimiG of these total 352 miRNAs for prostate cancer and calculate the AUC scores for performance evaluation.

We also use the same way to construct lung cancer and IBD specific datasets. Of the 1034 miRNAs, 172 and 18 are recorded as association with lung cancer and IBD in HMDD v3.0 as positives, respectively. We also randomly selected the same number of miRNAs as negatives. For each of the three diseases, we will also check the overlap between the top 50 predicted miRNAs and its associated miRNAs in HMDD v3.0.

In addition to the HMDD database, we also extract verified disease-miRNA associations from cancer database dbDEMC v2.0(Yang et al., 2017). We download disease-miRNA associations derived from low-throughput methods. Of the 1034 miRNAs, 15 and 16 miRNAs are recorded as association with prostate cancer and lung cancer in dbDEMC v2.0, respectively. These two numbers are much fewer than that of HMDD v3.0, it is because only disease-miRNA associations verified by low-throughput methods before 2017 are used for dbDEMC v2.0. Similarly, for each disease, we also randomly select the same number of miRNAs as negative controls.

We also construct the disease-specific dataset from PhenomiR v2.0 (Ruepp et al., 2010) and miRCancer (version miRCancer18February2019) (Xie et al., 2013) for prostate cancer and lung cancer. We do the same data processing as the dbDEMC for the two databases, respectively. Of the 1034 miRNAs, 86 and 60 miRNAs are associated with prostate cancer and lung cancer in miRCancer, respectively. We also select the same number of miRNAs not associated to prostate cancer and lung cancer in miRCancer as negative controls, respectively. For PhenomiR v2.0, we extract 157 and 190 miRNAs associated with prostate cancer and lung cancer, respectively. For each of the two cancers, the same number of miRNAs not associated with it in PhenomiR v2.0 is selected as negative samples.

## Reference

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M.*, et al.* (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343-+.

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.H., and Liu, H. (2018). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. Bioinformatics *34*, 3178-3186.
Defferrard , M., Bresson, X., and Vandergheynst, P. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. J Univers Comput Sci *19*, 3844-3852.
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.
Diederik, P.K., and Jimmy, B. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations (San Diego).
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
Grover, A., and Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 855-864
Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P., and Ideker, T. (2018). Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst *6*, 484-495 e485.
Junge, A., Refsgaard, J.C., Garde, C., Pan, X., Santos, A., Alkan, F., Anthon, C., von Mering, C., Workman, C.T., Jensen, L.J.*, et al.* (2017). RAIN: RNA-protein Association and Interaction Networks. Database (Oxford) *2017*.
Kipf, T.N., and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations (ICLR-17).
Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-D157.
Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res *42*, D1070-1074.
Liu, Y., and Zhao, M. (2016). lnCaNet: pan-cancer co-expression network for human lncRNA and cancer genes. Bioinformatics *32*, 1595-1597.
Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N.*, et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.
Mork, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J., and Jensen, L.J. (2014). Protein-driven inference of miRNA-disease associations. Bioinformatics *30*, 392-397.
Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635-+.
Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. Methods *74*, 83-89.
Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., and Theis, F.J. (2010). PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome biology *11*, R6.
Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res *40*, D940-946.
Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res *43*, D447-452.
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2014). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012). Nat Protoc *9*, 2513-2513.

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics *29*, 638-644.
Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., and Teschendorff, A.E. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res *45*, D812-D818.

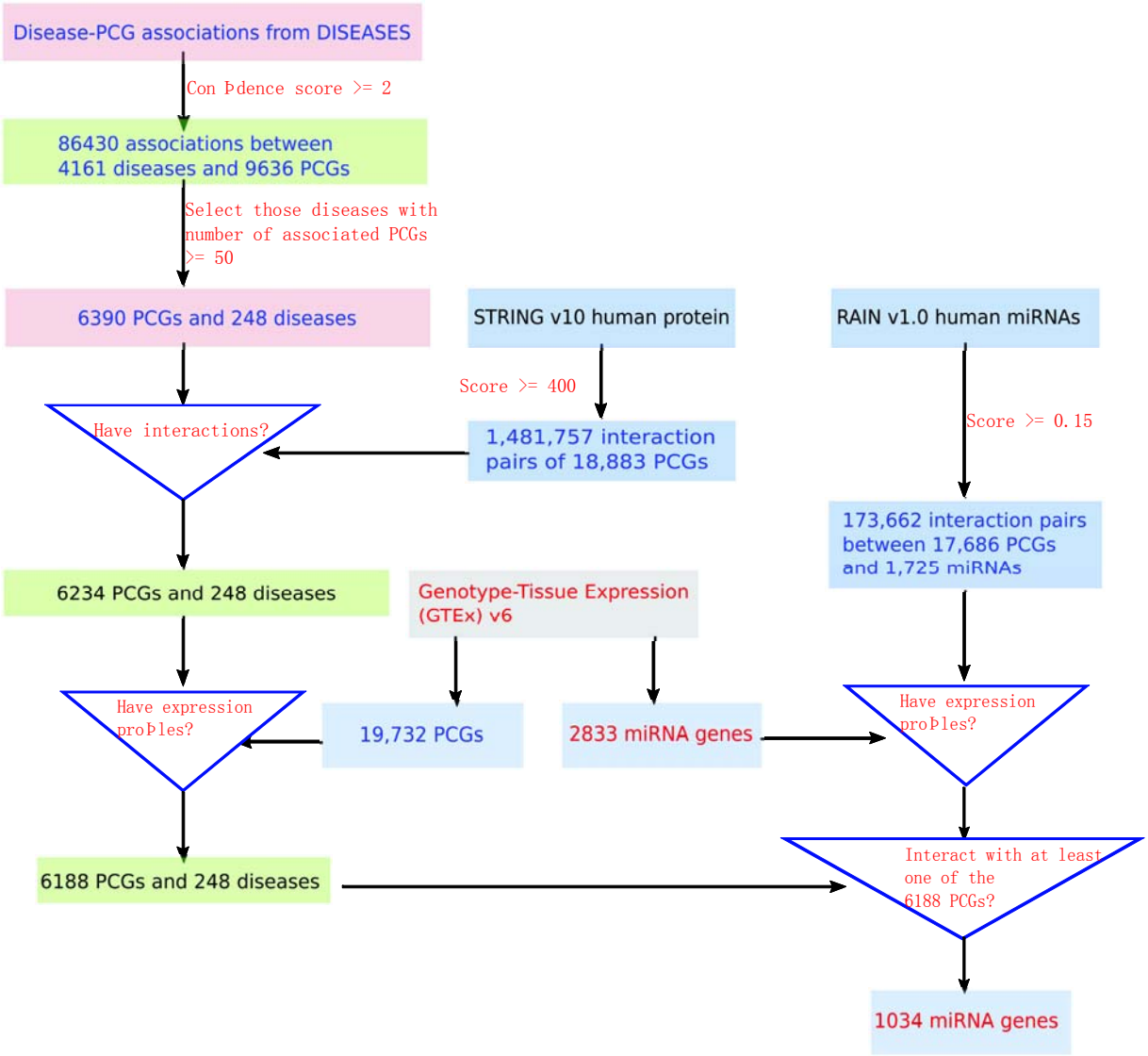



**Figure S1**.The flowchart of data processing for constructing the benchmark set for DimiG. All the gene names are mapped to Ensembl gene IDs and disease names are mapped to Disease Ontology ID. Related to Figure 1

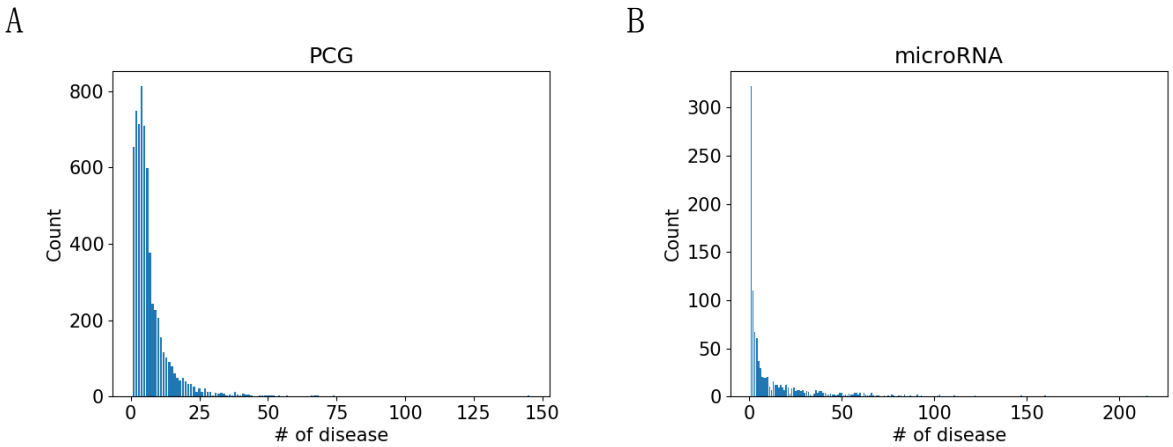



**Figure S2.** The number of disease distribution for protein coding genes and miRNAs. A) PCGs; B) miRNAs. Related to Figure 1

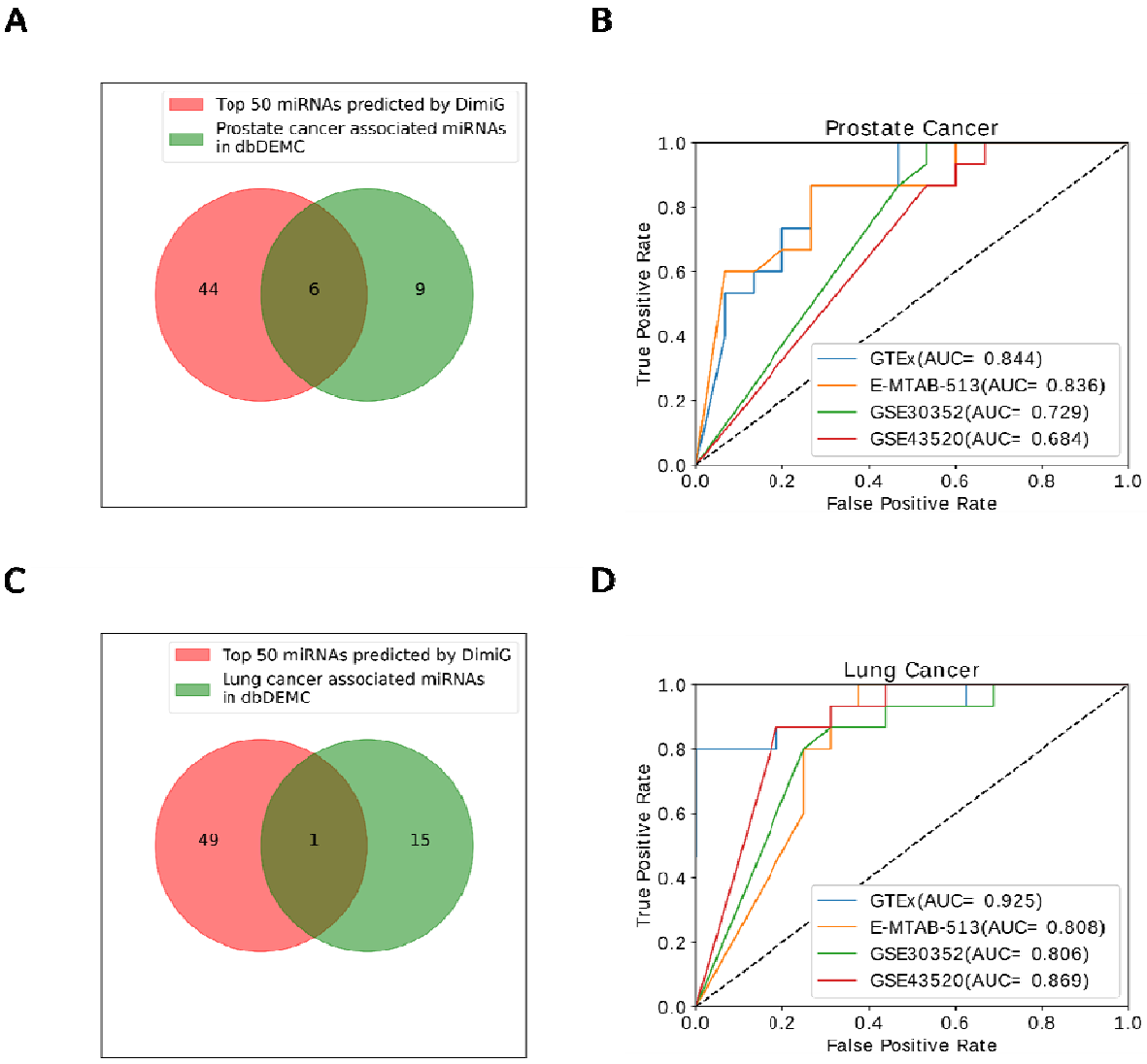
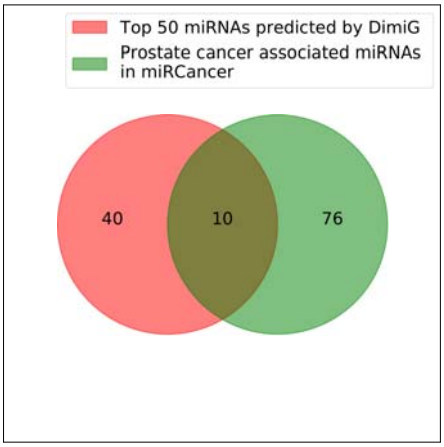
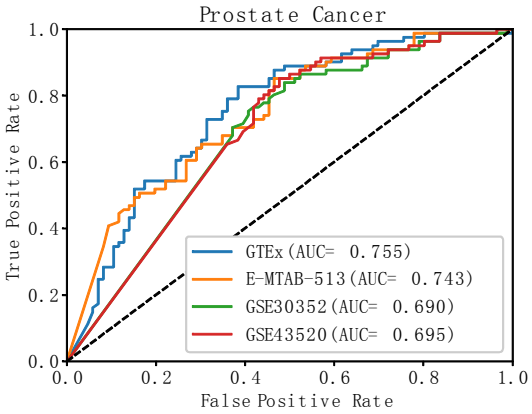



**Figure S3.** Venn diagram and ROC curve of predicting associated miRNAs for prostate cancer, lung cancer using DimiG compared to verified disease-miRNA associations from low throughput method in dbDEMC v2.0. A) and B), prostate cancer; C) and D), lung cancer. Related to Figure 5

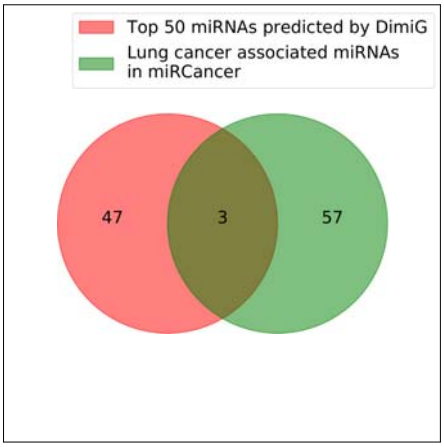
A




B




C




D

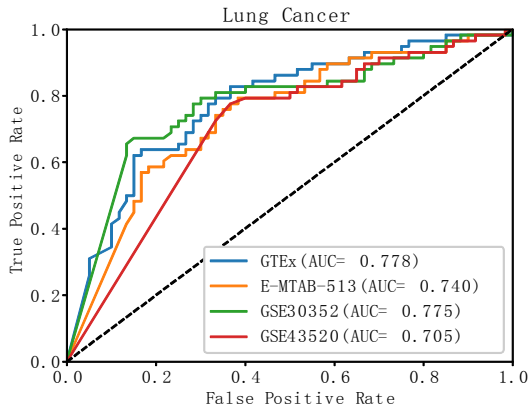
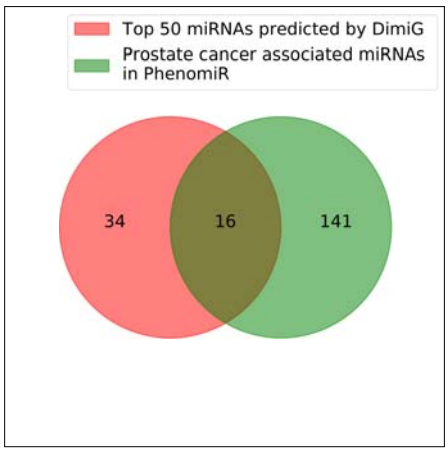
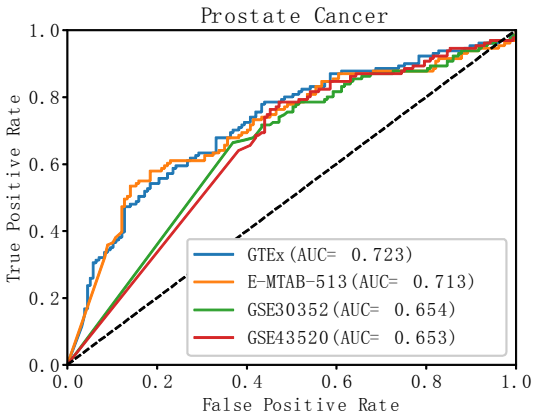



**Figure S4.** Venn diagram and ROC curve of predicting associated miRNAs for prostate cancer, lung cancer using DimiG compared to verified disease-miRNA associations from miRCancer. A) and B), prostate cancer; C) and D), lung cancer. Related to Figure 5

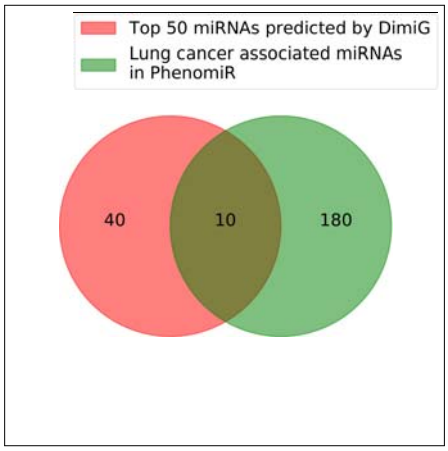
A




B




C




D

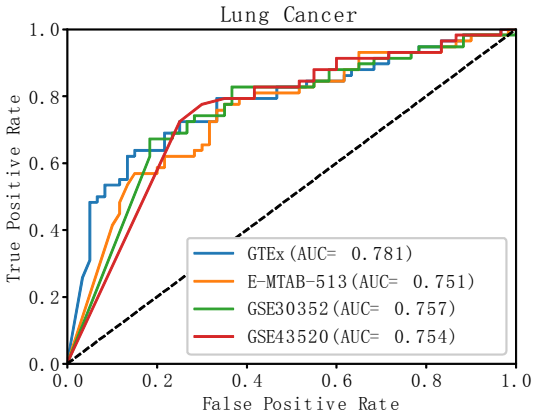



**Figure S5.** Venn diagram and ROC curve of predicting associated miRNAs for prostate cancer, lung cancer using DimiG compared to verified disease-miRNA associations from PhenomiR 2.0. A) and B), prostate cancer; C) and D), lung cancer. Related to Figure 5

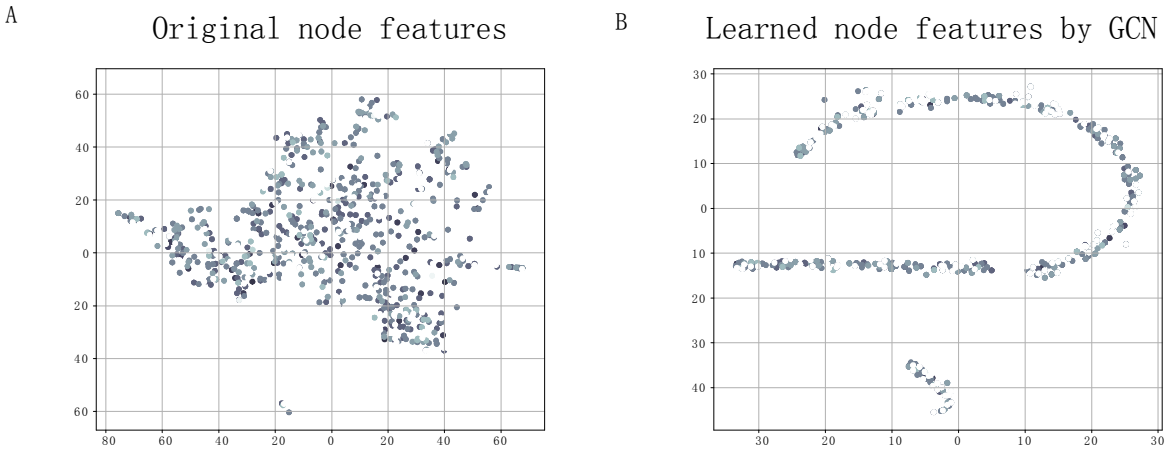



**Figure S6**. Scatter plot for the original node features from tissue expression profiles and the learned features in the last hidden layer of DimiG. T-SNE is used to map the original space to 2-D space. Related to Figure 3

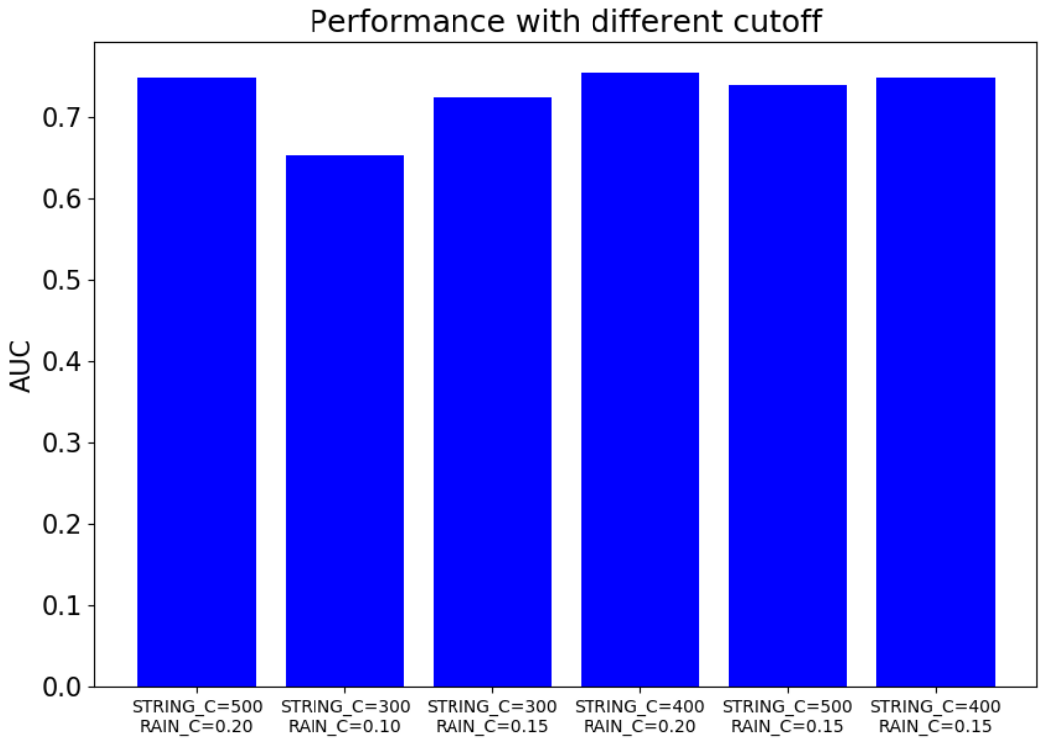



**Figure S7.**The impact of different interaction cutoff values on DimiG. For example, STRING_C =500 means using 500 as the cutoff value for interacting PCGs in STRING database, and RAIN_C= 0.2 is using 0.2 as the cutoff value for interacting PCG-miRNA in RAIN database. Related to Figure 2

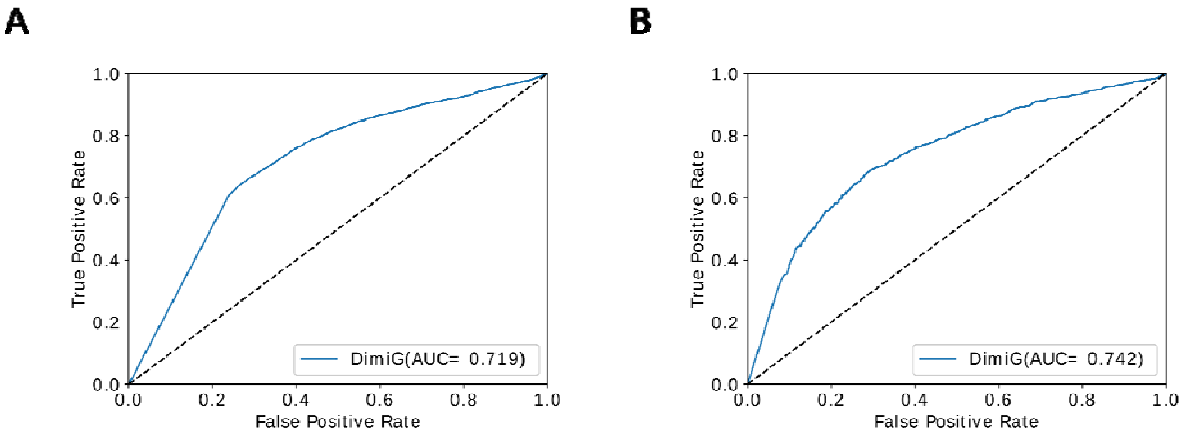



**Figure S8.** The performance of DimiG using confidence threshold 1.5 and 2.5 for disease-PCG associations in DISEASES database. Related to Figure 2