picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A consistent evaluation of miRNA-disease association prediction models

Ngan Thi Dong *, Megha Khosla

L3S Research Center, Leibniz University Hannover, Hannover, 30167, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A variety of machine learning based approaches have been applied to predicting miRNA-disease association. Although promising, the evaluation set up to measure prediction performance is inconsistent making it difficult to assess the actual progress. A more acute problem is that most of the models overlook the problem of data leakage due to the use of precomputed miRNA and disease similarity features.

Results: We unearth a crucial problem of data leakage in evaluation of machine learning models for miRNA-disease association prediction. In particular, information from test set, in the form of precomputed input features for miRNA and disease, is used during training of the model. Moreover, we point out problems in the widely used performance metrics used in model evaluation. While resolving the issues of data leakage and model evaluation, we perform an indepth study of 3 recent models along with our proposed 9 variants of these models. Our proposed variants have resulted in improvements in Average Precision scores (as compared to original models) by approximately 287.7% and 36.7% on HMDDv2.0 (AP:0.504) and HMDDv3.0 (AP: 0.216) datasets respectively.

Availability and Implementation: We release a unified evaluation framework including all models and datasets at `https://git.l3s.uni-hannover.de/dong/simplifying_mirna_disea.`

1 Introduction

MicroRNA or miRNA is a class of noncoding RNA with a length of approximately 22 nucleotides that is involved in the regulation of gene expression. miRNA can up-regulate or down-regulate genes and thus indirectly affect the transcription process. It has been observed that many of the miRNAs are associated with the development of complex diseases in humans. Identification of potential association between miRNA and disease would therefore help in clinical diagnosis, treatment and in finding potential drug targets.

As the biological methods for detecting these associations are expensive and time-consuming, recent years have seen an upsurge in computational approaches for predicting miRNA-disease associations. These approaches can be broadly grouped into three classes: scoring based methods, network topology based methods and machine learning based methods. Assuming that the miRNA pairs linked to common diseases are functionally more related [17, 15] proposed scoring systems to prioritize miRNA-disease associations. A more sophisticated scoring scheme while integrating information from miRNA and disease similarity networks was proposed in [20]. The second class of methods, i.e the network based algorithms [11, 4] construct miRNAs and/or disease similarity networks and aims at efficient transferring known miRNA-disease association labels between similar miRNAs and/or similar diseases in the network. Chen et al. [2] employ repeated random walks with restarts over the miRNA functional similarity network and prioritize candidate miRNA-disease associations using the final stable walk probability. In this work we mainly focus on machine learning based methods which typically build classification or regression models on input features extracted from different classes of heterogeneous information like disease semantic information, miRNA–disease association information and miRNA function information and sometimes also miRNA sequence information. For example, RBMMMDA [3] uses Restricted Boltzmann machine for the prediction task. DRMDA[1] applies a support vector machine model over the miRNA and disease features extracted using stacked autoencoders with unsupervised pre-training. DeepMDA[6] is a deep ensemble model which extracts features using stacked autoencoders and predicts finals scores of associations using a 3-layer neural network. A comparative

The Author 2020.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

study of various computational methods for miRNA-disease association prediction can be found in [7].

More recent works (published in January 2020) include EPMDA [5] which first extracts edge perturbation based features from the miRNA and disease similarity networks and then train a multilayer perceptron regression model to prioritize miRNA-disease associations. Another work, DBMDA [21] uses autoencoders to extract latent representations of miRNA and disease similarity as well as miRNA sequence similarity to further train a rotation forest classifier to predict prospective associations. NIMGCN [12] uses the graph convolution networks (GCNs)[9] to extract representations of miRNA and similarity networks and further use them to train a neural matrix completion network. All of these methods integrate different heterogeneous data sources, most notably, miRNA functional similarity, disease similarity extracted from the disease ontology, known miRNA-disease associations and in some cases also miRNA sequence similarity.

Problem of Data Leakage. While information integration from multiple sources has shown to improve performance, in this paper we show that the evaluation schemes in such a multi-source scenario suffers from a data leakage problem. Data leakage in machine learning refers to scenario when information from the test set, that should in principle be inaccessible during training, is used to create the model. To understand the problem of data leakage in miRNA-disease association predictive models, consider the concrete example of using precomputed miRNA functional similarities during model training. miRNA functional similarity (for details see Section 2) is computed using the known miRNA-disease associations and disease semantic information. More often, the precomputed miRNA functional similarity from MISIM database is used in model building without giving proper attention to the actual train/test split giving rise to data leakage. This implies that some of the associations which are to be tested by the model might be present in the association network that were used to compute the similarity features for example miRNA functional similarity. This leads to erroneous evaluation of models and overestimation of actual learning/prediction capabilities of the models. As an example, [7] also observes a fall in the performance of various models when trained on HMDD v2.0 dataset and tested on HMDD v3.0 dataset which the authors attributed to overfitting of models when trained and tested on the same dataset. We clarify this issue by showing that this discrepancy occured because of the data leakage due to the use of precomputed miRNA functional similarities deposited in databases and in some cases the use of the entire association networks to compute for example Gaussian Interaction Profile kernels.

Limitations of the evaluation setup. In addition, there has been a huge diversity in the experimental datasets used and evaluation metrics. Some models only report Area under the Receiver Operating Characteristic Curve (AUROC) scores which is typically more optimistic when the size of positive (actual associations) class is much smaller than the negative (non associations) class. A better performance metric is Area under Precision-Recall curve (AUPRC) which takes into account both precision and recall of the model at different decision thresholds. As computing AUPRC requires interpolating between adjacent supporting points, it too overestimates the actual performance in extreme cases, for example when there are just two pints in precision-recall curve. In this work, we instead use Average Precision (AP) (more details in Section 4.2) which summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. It does not use linear interpolation and provides a more reliable performance estimate than AUPRC.

Our Contributions. In this work, we perform an indepth study of 3 recent machine learning models, namely EPMDA, DBMDA and NIMGCN. In addition to resolving the problems of data leakage and evaluation setup, we propose and study 9 variants of the studied models. Through a comparative analysis of resulting 12 models over two widely used datasets our work offers deeper insights into the workings of very recent machine learning models for the miRNA-disease association prediction task. To summarize, we make the following contributions.

- We unearth a crucial problem of data leakage in evaluation of machine learning models for miRNA-disease association prediction. We develop and release an integrated evaluation framework in which various miRNA and disease similarities are computed for each training set thus disallowing any data leakage from the test set.
- We thoroughly investigate 3 state of the art models including 9 of their simpler variants showing advantages of reduced model complexities. Our proposed simpler variants have resulted in one of the best Average Precision score observed so far on HMDD v2.0 [13] and HMDD v3.0 [8] datasets.
- We compared the effect of using different input miRNA and disease similarity features and find that using Gaussian Interaction Profile (GIP) kernel similarity leads to better performance.

2 Preliminaries

In this paper, we study supervised machine learning methods for the miRNA-disease associations prediction problem. That is to say, from a set of known miRNA-disease associations, learn a model to predict potential miRNA-disease associated pairs. Our studied models treat that problem as either a binary classification task where each pair is represented as a feature vector fed as input to a classifier, or as a matrix completion problem. In order to construct the feature vector or build the matrix, different similarities measures have been used. This section will give a brief overview of all the similarity measures employed by the models studied in this work.

Disease Similarity. Disease Semantic Similarity[17] computes the semantic similarity of two diseases based on their relative positions on the disease directed acyclic graph (DAG) which is derived from the MeSH[1] disease descriptor database. Specifically in the disease DAG the nodes represent diseases and directed links from parent nodes to child nodes encode a 'is-a' relationship. The contribution of disease A to the semantic value of disease t is then defined as:

$$\begin{cases} D_A(A) = 1 \\ D_A(t) = \max\{\Delta \cdot D_A(t') | t' \in \text{children of t}\}, \text{ if } t \neq A \end{cases} \tag{1}$$

where Δ is the decaying factor and is set to 0.5 as in [17]. Then the semantic value for disease A is calculated as $DS(A) = \Sigma_{t \in P_A} D_A(t)$, where P_A is the set of t's ancestors. Finally the **semantic similarity** between two diseases A and B is given by:

$$S(A, B) = \frac{\sum_{t \in P_A \cap P_B} (D_A(t) + D_B(t))}{DS(A) + DS(B)} \tag{2}$$

where P_A and P_B are the set of ancestors for diseases A and B respectively. However, Xuan et al. [18] argue that the more a particular disease appears as an ancestor of other diseases the less specific it is for a particular disease. Given that the likelihood of a disease node t appearing as ancestors of all other diseases, [18] quantifies the information content of t, $IC(t)$, as the negative log of the likelihood, i.e. $IC(t) =$

[1] http://www.ncbi.nlm.nih.gov/mesh

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

$-\log(p(t))$. The phenotype similarity score between two diseases is then given by

$$S'(A,B) = \frac{\sum_{t \in P_A \cap P_B}(IC(t) + IC(t))}{\sum_{t \in P_A} IC(t) + \sum_{t \in P_B} IC(t)} \tag{3}$$

The updated similarity score between two diseases is then calculated as an average of their semantic similarity and phenotype similarity scores.

$$DS(A,B) = \frac{S(A,B) + S'(A,B)}{2} \tag{4}$$

For simplicity, from now on, without explicitly given the formula, by saying **disease semantic similarity**, we mean the disease semantic similarity given in Equation (4).

miRNA functional similarity [17]. The miRNA functional similarity is estimated based on the similarities between the diseases to which they are associated. In particular, first the similarity between disease d_i and a group of disease DG is computed as $DSG_{d_i}(DG) = \max_{B \in DG}(DS(d_i, B))$. Given that $\mathcal{D}_i$ denote the set of disease associated with miRNA m_i, the functional similarity between two miRNAs m_1, m_2 is then calculated as:

$$FS(m_1, m_2) = \frac{\sum_{dt1 \in \mathcal{D}_1} DSG_{dt1}(\mathcal{D}_2) + \sum_{dt2 \in \mathcal{D}_2} DSG_{dt2}(\mathcal{D}_1)}{|\mathcal{D}_1| + |\mathcal{D}_2|}. \tag{5}$$

miRNA sequence similarity [12] The miRNA sequence is first mapped to its 2 dimensional (2D) representation where each nucleotide in its sequence is converted to a 2-dimensional vector. For ith nucleotide in a particular miRNA sequence, its 2D encoding is computed by the recursive relation $nt_i = nt_{i-1}\theta(nt_{i-1} - \lambda_i)$ such that $nt_0 = (0.5, 0.5)$ and the decaying factor $\theta = -0.5$. The parameter λ_i is set as follows

$$\lambda_i = \begin{cases} (0,0), \text{ if } i\text{th nucleotide is A} \\ (0,1), \text{if } i\text{th nucleotide is C} \\ (1,1), \text{if } i\text{th nucleotide is G} \\ (1,0), \text{ if } i\text{th nucleotide is U} \end{cases}.$$

The sequence similarity of two miRNAs m_i and m_j is then defined as the Euclidean distance between their two vectors in the above defined 2D space.

Gaussian Interaction Profile (GIP) Kernel Similarity [16]. Given the miRNA-disease association matrix MD, the GIP kernel similarity between two miRNAs is defined as:

$$GM(m_i, m_j) = \exp(\lambda_m ||IP(m_i) - IP(m_j||^2), \tag{6}$$

where $IP(m_i)$ and $IP(m_j)$ correspond to the ith and jth rows of matrix MD respectively. The parameter λ_m is set to $\frac{n_m}{\Sigma_{i=1}^{n_m}||IP(m_i)||^2}$ where n_m is the total number of miRNAs. The Gaussian Interaction Profile kernel similarity between two diseases is calculated similarly as:

$$GD(d_i, d_j) = \exp(\lambda_d ||IP(d_i) - IP(d_j||^2), \tag{7}$$

where $IP(d_i)$ and $IP(d_j)$ correspond to the ith and jth column of matrix MD accordingly. The parameter λ_d is set to $\frac{n_d}{\Sigma_{i=1}^{n_d}||IP(d_i)||^2}$ where n_d is the total number of diseases in MD.

3 Compared Models

In this section we describe the 3 recent models studied in this work along with our proposed variants. Our proposals for model variants are based on (i) reducing redundant non-linearities hence decreasing the number of model parameters (ii) identifying pitfalls in training data set for example the use of imbalanced training set in EPMDA. In the following we provide details of the studied models and our proposed variants.

EPMDA [5]. The edge perturbation based method for miRNA-disease association prediction (EPMDA) provides a new feature extraction method based on edge perturbation in miRNA-disease heterogenous network.

The heterogenous network G input to EPMDA can be divided into three parts as follows:

- *The miRNA-disease bipartite network* consists of two type of nodes: one for miRNA and one for disease. There is no connection between two nodes of the same type and a link between a miRNA and a disease node indicates a known association.
- *The miRNA similarity network* is a weighted network where each node is a miRNA and the edge weight is equivalent to the `GIP` kernel similarity between two miRNA nodes (as defined in Equation (6)).
- *The disease similarity network* is a weighted network where each node represents one disease and the weight of the edges are the disease `GIP` kernel similarities (as given in Equation (7)).

The weighted adjacency matrix (A) of G is given as

$$A = \begin{bmatrix} GM & MD \\ MD^\top & GD \end{bmatrix}$$

The problem of miRNA-disease association prediction is treated as the problem of predicting links between miRNA nodes and disease nodes on G. Feature extraction corresponding to edges in G is then proposed based on the network topology where each feature f_i reflects the disturbance level resulted from the deletion of that particular edge to the set of cycles (in which it participates) of length i in G. The higher the disturbance, the more important that particular edge is. A 5-layer Multilayer Perceptron Regression model is then trained using the extracted features with regularized square loss to predict whether there is a link between the miRNA-disease pair.

Table 1. EPMDA and its variants

Model	Training dataset		Regression Model	
	IMBALANCED	BALANCED	MLP	LINEAR
EPMDA	✓	x	✓	x
EPMDA1	✓	x	x	✓
EPMDA2	x	✓	✓	x
EPMDA3	x	✓	x	✓

Variants of EPMDA. We investigated a variant of EPMDA where we replaced the MLP regressor with a linear regressor. In addition, as elaborated later in Section 4.3.1, EPMDA uses an imbalanced training set which might downgrade its performance. Therefore, besides the original proposed model, we also propose and investigate three variants of EPMDA as summarized in Table 1. We call a data set that has equal number of positive and negative samples a balanced dataset. EPMDA denotes the original proposed model which uses an imbalanced training dataset and the MLP regressor. EPMDA1 is the model retrieved when we replace the MLP regression model with its linear counterpart. EPMDA2 is obtained when we use a balance training set instead of the imbalanced

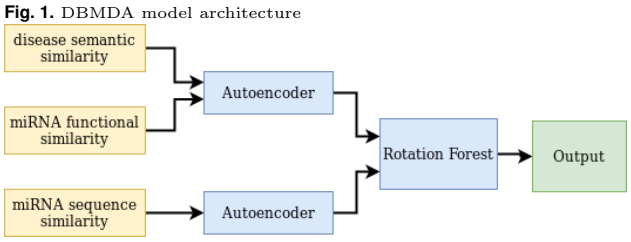
picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

one. Finally, EPMDA3 combines the changes we made in EPMDA1 and EPMDA2. Namely, we both replace the MLP regression model with Linear regression model and use a balance training set to train the model.

DBMDA [21]. The distance-based sequence similarity for miRNA-disease association prediction (DBMDA) combines the information of miRNA sequence, miRNA functional similarity, miRNA-disease associations and disease semantic similarity. DBMDA architecture is illustrated in Figure 1. In particular for a given miRNA-disease pair it first concatenates the miRNA functional similarity and disease similarity features (computed in Section 2) and extracts a joint low dimensional representation using an autoencoder (denoted as Autoencoder 1 in Table 2). Also corresponding to miRNA, the sequence similarity features as described in Section 2 are fed to another autoencoder (denoted as Autoencoder 2 in table 2) to extract latent representation corresponding to miRNA sequence. The two newly extracted representations are concatenated and then fed to a rotation forest classifier to predict potential miRNA-disease associations.

Fig. 1. DBMDA model architecture



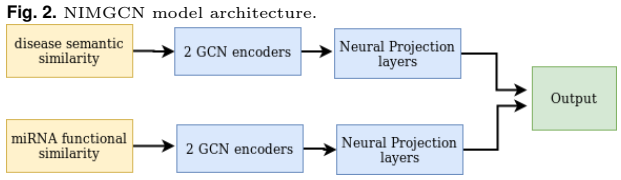
Variants of DBMDA. We want to investigate the effectiveness of miRNA sequence similarity, autoencoders on DBMDA performance. Therefore, in this paper, apart from the original proposed model, we also report results from its three variants in which we either remove the auto-encoder, the miRNA sequence similarity features or both. As shown in Table 2, the DBMDA model corresponds to the original model. We obtain DBMDA1 from the removing the two autoencoders which are used to extract low dimensional representations from miRNA-disease pair and miRNA sequence similarity. Instead, for a particular miRNA-disease pair, we directly concatenate the miRNA functional, disease and miRNA sequence similarity to form a features vector. DBMDA2 corresponds to not using or removing the miRNA sequence similarity features from the input. For DBMDA3 we do not use miRNA sequence similarity features and any of the autoencoders. The concatenated features of miRNA-disease pair are directly used as input to the rotation forest classifier.

Table 2. DBMDA and its variants

Model	Autoencoder 1	Autoencoder 2	miRNA sequence features
DBMDA	✓	✓	✓
DBMDA1	x	x	✓
DBMDA2	✓	✓	x
DBMDA3	x	x	x

NIMGCN [12]. NIMGCN operates on the heterogeneous miRNA-disease association network in which there are two type of nodes: miRNAs and diseases. miRNAs nodes form a miRNA functional similarity network where the edge weights correspond to the miRNA functional similarity scores as given in (5). Disease nodes form a disease semantic network where the edge weights are given by disease semantic similarity score calculated according to (4). A link between a miRNA and a disease node indicate a known association. Node representations are learnt using two Graph Convolution Networks (denoted as GCN encoders in figure 2) each for miRNA functional similarity network and disease similarity network separately. Non linear transformation or neural projections are applied over these learnt representations to extract latent features corresponding to miRNA and disease respectively. The predictions for miRNA-disease association ratings are then computed as the inner products of the latent features corresponding to miRNA and disease. NIMGCN architecture is depicted in Figure 2.

Fig. 2. NIMGCN model architecture.



Variants of NIMGCN. In this paper, apart from the original NIMGCN model, we also construct experiments on three of its variations in which we remove some part of the original model to make a simpler one. Details about those variants are given in table 3. For NIMGCN1, we remove the GCN encoders. miRNA functional and disease similarity scores are fed directly as input to the neural projections to learn hidden representation. For NIMGCN2, we remove the stack of three neural projections. Output from GCN encoders are used directly as miRNA and disease hidden representation. And for NIMGCN3, we use only one GCN encoder and one neural projection to learn the latent features of miRNA/disease.

Table 3. NIMGCN and its variants

Model	Architecture
NIMGCN	original
NIMGCN1	without GCNs
NIMGCN2	2 GCN layers, without neural projection
NIMGCN3	1 GCN layer and 1 neural projection layer

4 Data and Experimental setup

4.1 Data Collection

We use two datasets derived from HMDD 2.0 [13] and HMDD 3.0 [8] as our evaluation datasets. The first dataset, which from now on we denoted as HMDD2 consists of 5430 associations between 495 and 383 diseases. The second dataset, which from now on we denoted as HMDD3, consists of 35362 associations between 1062 and 893 diseases.

Table 4. The evaluated datasets statistics. n_m, n_d, N denote the number of miRNAs, diseases, and known associations respectively. $n_{d \notin MESH}$ refers to the number of diseases that are not found in MESH.

Dataset	n_m	n_d	N	$n_{d \notin MESH}$
HMDD2	495	383	5430	55
HMDD3	1062	893	35362	360

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

The MESH terms for disease semantic similarity calculation are downloaded from the National Library of Medicine[2]. Diseases are matched by names. There are 55 and 360 disease names which are not found in MESH in HMDD2 and HMDD3 respectively. For those diseases, we fill in the similarity matrix with their corresponding Gaussian Interaction Profile (GIP) kernel similarity (calculated from the miRNA-disease association network) as described in Section 2. All miRNA sequence information was retrieved from miRBase [10]. The two datasets statistics are given in Table 4.

4.2 Evaluation metrics

In terms of measurements, most of the paper in the fields can use either top K Predictive Rate, Area under the Receiver Operating Characteristic (AUROC) or Area under the Precision Recall Curve (AUPRC). However, for top K Predictive Rate, different value of K leads to different evaluations and in real applications, it is hard to specify the appropriate value of K. Therefore, we believe that using top K Predictive Rate is not recommendable. For AUROC score, as previously discussed in [19], AUROC score sometimes misleading for highly imbalanced datasets like our network of interest (the rate of positive link : negative link is 5430 : 184155 $\approx$ 0.0295 and 35362:948366 $\approx$ 0.0373 for HMDD2 and HMDD3 respectively). We argue that the use of AUPRC is more appropriate in this case. However, a point to be noted here is the calculation of the area under the precision-recall curve. The sklearn function to calculate the area under the cuve[3] might suffer from the overestimation of the actual area because of linear interpolation. For example if there are just two points of x- and y-axis it computes the area under the line joining these two points which is clearly incorrect. We instead use Average Precision(AP)[4] which summarizes a precision-recall curve as the weighted mean of precision achieved at each threshold (the weight here is the increase in recall from the previous threshold). AP is computed as:

$$AP = \sum_n (R_n - R_{n-1}) \cdot P_n \tag{8}$$

where R_n, P_n are the recall and precision at the nth threshold correspondingly. Since EPMDA and NIMGCN report AUROC as their evaluation metrics, for completeness, we report both AUROC and AP in our results.

4.3 Experimental setup

For all the models on a particular data set, we run 5 times 5-fold CV on the known miRNA-disease association set with different random seeds. Overall each model was trained/tested 25 times corresponding to 25 possible combination of training/testing splits. For simplicity, we call a training/testing split as a data split. More details about the data preparation process is given in Section 4.3.1. For all models, we use the same set of data splits.

In terms of the choice of similarity measure fed as input to the models, we experimented with three different type of similarities (input features): (i) the **GIP**: where we use the GIP kernel similarity for miRNA and disease; (ii) the **Functional+Semantic** where we use the miRNA functional similarity and disease semantic similarity given in Equation (4); and (iii) the **miRNA sequence+Semantic** where we use the miRNA sequence similarity and disease semantic similarity given in Equation (4).

We also construct experiments where we use the disease semantic similarity as stated in (2) instead of the one given in (4). However, from the results retrieved, we saw no significant difference between the two. Because of that reason and also due to space limit, we do not present the results corresponding to this experiment in this paper but details are available upon request.

4.3.1 Training-testing data set up

Balanced training set. To create training splits we first randomly split all the known associations (positive examples) into 5 folds. The 4 folds constitute the training data corresponding to positive examples. To that we add an equal number of random sampled pairs which are not in those 4 folds to form the negative examples of the training set.

Test set preparation To build the test set we use the remaining fold of the positive examples and mark all the others possible miRNA-disease combination which are not in the set of known association as negative examples. In our evaluation set up we use all negative pairs in the test set as opposed to artificially creating a balanced test set by sampling a smaller number of negative examples. This setting is more realistic as in real time the task will be to rank the small number of prospective associations among a large number of possible associations. Such a setting is also emphasized in [19].

A different train/test set up is used in the original papers of NIMGCN and DBMDA which explicitly mark all unknown pairs as negative samples and divide the set of all possible pairs into 5 folds. Each time, a fold is used for testing and the others for training. For example, in case of HMDD2 dataset, the 5430 known miRNA-disease associations are used as positive samples while the other 184155 - 5430 = 183725 pairs are explicitly marked as negative samples. Instead of a complete test set of more than 180k pairs, their test set size is just around 36k pairs. There are two problems with this kind of configuration: (i) their training set is highly biased as the number of negative examples are much more than that of positive examples which is often detrimental to the model behavior, for example training error might be close to zero if the model labels all examples as negative. (ii) the test set contains much smaller amount of negative examples than the actual, hence the model is not tested in a realistic setting where it is required to predict/prioritize a small number of associations from a large number of possible combinations.

For EPMDA, the authors also divide only the positive associations into 5 folds like our setting. However, for each data split, their training data composed of 4 folds of positive samples and all the other possible pairs as negative samples which also results in an imbalanced training set. We test the EPMDA model using their original setting where they use an imbalanced training set (referred to as EPMDA in in Table 1, Figures 3 and 4) and in our setting where a balanced training dataset is used (referred to as EPMDA3 in tables 1, 7 and Figure 4). We see a considerable improvement in the model performance on HMDD2 dataset when a balanced training set was used.

4.3.2 Computation of Similarities

As already pointed out, existing works suffer from a data leakage problem in which associations from the test network were used for generating miRNA functional and GIP kernel similarities. In this work, we resolve the data leakage issues by re-calculating all similarities in accordance with the given data splits. More specifically, for each run, miRNA functional similarity or miRNA/disease GIP kernel similarity scores are computed using only the associations from the training data. That means when we

[2] http://www.nlm.nih.gov/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html##sklearn.metrics.auc

[4] https://en.wikipedia.org/w/index.php?title=Information_retrieval&oldid=793358396##Average_precision

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

do 5-fold cross validation, those similarity scores are calculated 5 times with different input.

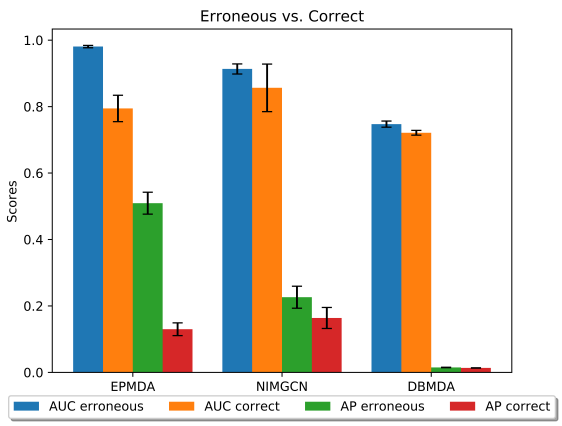
For miRNA sequence similarity, we use the code from DBMDA. To compute GIP kernel similarity, we use the code provided by EPMDA authors'. For the other similarity measures, we cannot get the original code to calculate the disease similarity as well as miRNA functional similarity scores given in MISIM v1.0 database. Therefore, for consistency, we use our own implementation to calculate all the miRNA functional and disease similarities (both from the whole network (as previously used by the studied models) and from only the training data alone). We provide all the code for similarities calculation at `https://git.l3s.uni-hannover.de/dong/simplifying_mirna_disease`.

4.3.3 Hyperparameter settings

For DBMDA, since only the code for miRNA sequence similarity calculation is available, we use a simple autoencoder consisting of one encoder and one decoder. The encoder is a densely connected layer of size 32 (which is the encoded dimension explicitly given in DBMDA paper) with ReLU activation function and L1 regularization. The decoder, which has the same size as the input, is a densely connected layer with sigmoid activation function. We train that autoencoder for 1000 epoch in all experiments. For the Rotation Forest [14] classifier, we use the implementation from `https://github.com/digital-idiot/RotationForest` with all the default parameters. For NIMGCN, we use all the author's parameters like the number of epochs, loss function, number of hidden units, etc. For EPMDA, we use the authors' code and settings for both feature calculation and MLP regression model.

5 Results and discussion

Fig. 3. Erroneous and correct results for three original studied models on HMDD2 dataset

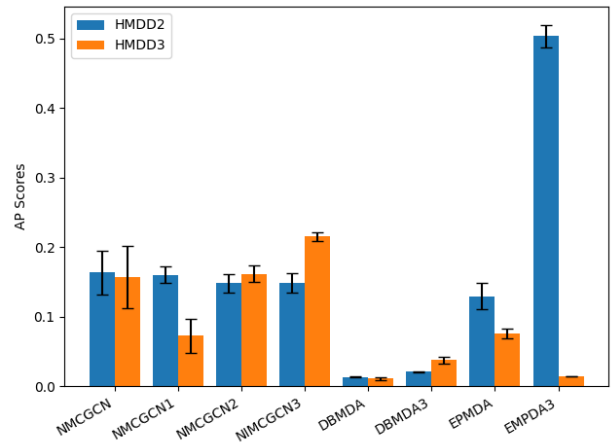


In our experiments, we investigate and answer the following research questions.

- **RQ1:** Which models are most affected by the data leakage problem caused by using precomputed similarity features?
- **RQ2:** How do our proposed model variants compare with the original models?
- **RQ3:** How sensitive are the models to different kinds of miRNA and disease similarity features (GIP, miRNA functional, disease semantic etc.)?

In Section 5.1 we answer RQ1 by comparing originally proposed models with the similarity features computed on the whole association data and using only the training split. The results corresponding to RQ2 are discussed in Section 5.2. In Section 5.3 we compare the difference in performance with respect to input similarity features (c.f. RQ3).

Fig. 4. Mean AP scores (with standard deviation) for different models on HMDD2 and HMDD3 dataset. The worst performing model variants are removed from the plot. The input similarity features correspond to the ones used in the original models.



5.1 Impact of pre-computed similarities

Figure 3 shows the comparison the AUROC and AP scores of the three models on HMDD2 dataset. For all models, the "Erroneous results" denote results retrieved from models which use pre-computed similarities while "Correct results" denote results obtained from models that use similarities calculated at run time using only the given training data.

As expected, we find that the use of pre-calculated similarity does result in a much higher AUROC or AP scores. We observe the highest performance drop in EPMDA because its calculated features are heavily relied on the fed GIP kernel similarities. We believe that the reason behind this observation is that the GIP similarities are computed using only the association information. To represent a disease node, NIMGCN and DBMDA use disease semantic similarity features which (unlike GIP based disease similarity) are computed from the disease ontology and not from the miRNA-disease associations. We, therefore, observe a lower impact of this correction step on DBMDA and NIMGCN. Moreover, DBMDA also employs additional miRNA sequence similarity features which can be computed independently from the associations which makes it relatively more robust to the changes in the miRNA functional similarity features. Therefore, DBMDA is impacted least of all models.

5.2 Impact of model architecture and training/test data setups

In order to answer RQ2, we compared the 3 studied models along with their several variants on HDMM2 and HDMM3 datasets. Figure 4 presents the mean AP scores for these models with standard deviation. We make the following observations.

1. EPMDA3, which corresponds to our proposed variant of EPMDA, outperforms all other models with a large margin(around 287.7% of improvement in AP score) for the HMDD2 dataset. We believe

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Table 5. Mean AUROC and AP scores (with standard deviation) for NIMGCN and its variants on HMDD2 and HMDD3 datasets. The original NIMGCN model uses Functional+Semantic similarity.

	Method	GIP		**Functional+Semantic**		miRNA sequence+Semantic	
		AUROC	AP	AUROC	AP	AUROC	AP
HMDD2	NIMGCN	0.906 ± 0.006	0.152 ± 0.009	0.857 ± 0.072	$\mathbf{0.164 \pm 0.032}$	0.850 ± 0.005	0.056 ± 0.003
	NIMGCN1	0.913 ± 0.004	0.146 ± 0.010	0.889 ± 0.008	0.161 ± 0.011	0.848 ± 0.006	0.056 ± 0.003
	NIMGCN2	0.889 ± 0.017	$\mathbf{0.174 \pm 0.015}$	0.856 ± 0.008	0.148 ± 0.014	0.892 ± 0.008	$\mathbf{0.177 \pm 0.011}$
	NIMGCN3	0.779 ± 0.019	0.086 ± 0.010	0.849 ± 0.007	0.149 ± 0.014	0.883 ± 0.006	0.166 ± 0.012
HMDD3	NIMGCN	0.934 ± 0.007	0.199 ± 0.006	0.887 ± 0.086	0.158 ± 0.045	0.856 ± 0.006	0.036 ± 0.003
	NIMGCN1	0.939 ± 0.002	0.205 ± 0.007	0.858 ± 0.043	0.073 ± 0.025	0.858 ± 0.003	0.038 ± 0.001
	NIMGCN2	0.918 ± 0.024	$\mathbf{0.212 \pm 0.013}$	0.872 ± 0.022	0.162 ± 0.011	0.813 ± 0.067	0.070 ± 0.042
	NIMGCN3	0.913 ± 0.004	0.126 ± 0.006	0.913 ± 0.004	$\mathbf{0.216 \pm 0.006}$	0.909 ± 0.004	$\mathbf{0.171 \pm 0.006}$

Table 6. Mean AUROC and AP scores (with standard deviation) for DBMDA and its variants on HMDD2 and HMDD3 datasets. The original DBMDA model uses Functional+Semantic similarity along with miRNA sequence similarity. Results for DBMDA and DBMDA1 are not presented in the last column as the original models use miRNA sequence similarity. DBMDA1 could not be computed for HMDD3 because of out of memory error.

	Method	GIP		**Functional+Semantic**		miRNA sequence+Semantic	
		AUROC	AP	AUROC	AP	AUROC	AP
HMDD2	DBMDA	0.798 ± 0.016	0.022 ± 0.003	0.721 ± 0.007	0.013 ± 0.001	-	-
	DBMDA1	0.807 ± 0.006	$\mathbf{0.025 \pm 0.001}$	0.785 ± 0.005	$\mathbf{0.021 \pm 0.001}$	-	-
	DBMDA2	0.804 ± 0.009	0.024 ± 0.002	0.723 ± 0.007	0.014 ± 0.000	0.720 ± 0.007	0.013 ± 0.000
	DBMDA3	0.807 ± 0.007	$\mathbf{0.025 \pm 0.001}$	0.786 ± 0.005	$\mathbf{0.021 \pm 0.001}$	0.785 ± 0.007	$\mathbf{0.021 \pm 0.001}$
HMDD3	DBMDA	0.826 ± 0.008	0.018 ± 0.002	0.757 ± 0.029	0.011 ± 0.002	-	-
	DBMDA2	0.837 ± 0.011	0.020 ± 0.002	0.783 ± 0.005	0.012 ± 0.000	0.685 ± 0.029	0.007 ± 0.001
	DBMDA3	0.839 ± 0.005	$\mathbf{0.021 \pm 0.001}$	0.869 ± 0.026	$\mathbf{0.038 \pm 0.005}$	0.801 ± 0.004	$\mathbf{0.014 \pm 0.000}$

Table 7. Mean AUROC and AP scores (with standard deviation) for EPMDA and its variants on HMDD2 and HMDD3 datasets. The original model used GIP kernel similarity. Because of high time complexity, only 5 runs were performed for HMDD3 dataset. For the same reason, the scores corresponding to the last column could not be computed.

	Method	**GIP**		Functional+Semantic		miRNA sequence+Semantic	
		AUROC	AP	AUROC	AP	AUROC	AP
HMDD2	EPMDA	0.794 ± 0.040	0.130 ± 0.019	0.94 ± 0.006	0.382 ± 0.020	0.921 ± 0.004	0.200 ± 0.011
	EPMDA1	0.855 ± 0.008	0.170 ± 0.008	0.944 ± 0.006	$\mathbf{0.459 \pm 0.019}$	0.922 ± 0.005	$\mathbf{0.216 \pm 0.012}$
	EPMDA2	0.498 ± 0.015	0.013 ± 0.001	0.883 ± 0.005	0.144 ± 0.008	0.869 ± 0.004	0.092 ± 0.007
	EPMDA3	0.964 ± 0.004	$\mathbf{0.504 \pm 0.017}$	0.870 ± 0.009	0.152 ± 0.012	0.920 ± 0.005	0.155 ± 0.010
HMDD3	EPMDA	0.798 ± 0.047	$\mathbf{0.076 \pm 0.007}$	0.908 ± 0.007	0.071 ± 0.003	-	-
	EPMDA1	0.870 ± 0.002	0.068 ± 0.002	0.900 ± 0.004	$\mathbf{0.083 \pm 0.002}$	-	-
	EPMDA2	0.493 ± 0.008	0.005 ± 0.000	0.648 ± 0.005	0.009 ± 0.000	-	-
	EPMDA3	0.775 ± 0.003	0.015 ± 0.000	0.799 ± 0.004	0.031 ± 0.001	-	-

that the use of balanced training data as also discussed earlier have contributed to a better trained model.

2. NIMGCN3, which is also one of our proposed variants outperforms other models for HMDD3 dataset. Moreover, it shows a lower standard deviation over the 25 runs as compared to the original NIMGCN model.
3. From the above two observations we conclude that the edge perturbation features based on fixed cycle lengths used by EPMDA are not generalizable for the larger HMDD3 dataset. We believe that its performance can be improved by fine-tuning the hyperparameter corresponding to the length of the longest cycle.
4. Both NIMGCN and EPMDA consider actively the topological structure of the heterogeneous miRNA-disease network whereas DBMDA simply extract low dimensional representations from the similarity features. The poor performance of DBMDA as compared to other models points to the importance of learning from the topological structure of the network.

Detailed results for all models and its variants using different kinds of similarity features (discussed in next section) are presented in tables 5, 6 and 7. We would also like to point out that among all the studied models, EPMDA took the longest time to run. To calculate the features, we let 5 threads running in parallel. However, it still took EPMDA 35 hours to run on HMDD2 (25 runs) and nearly 27 days to run on HMDD3 (only 5 runs) on our servers of 16 cores, 128G memory. The theoretical time complexity of EPMDA is $O(\ell n_m n_d (n_m + n_d)^3)$ where ℓ is the length of longest cycles considered in graph G that are used to calculate the edge disturbance level, n_m is the number of and n_d is the number of diseases. Therefore, for large network, EPMDA feature calculation is extremely expensive. Moreover we also note that EPMDA performance on HMDD3 is lower than NIMGCN and its variants by a large margin.

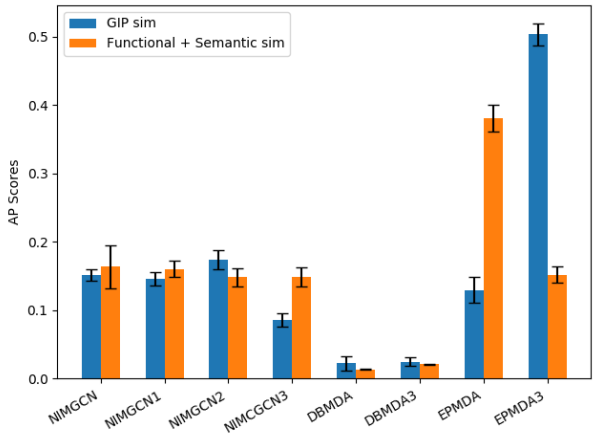
picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

5.2.1 The rate of negative training samples in the training data

As discussed earlier, the use of an imbalance training dataset might downgrade the model performance. To quantify the effect of different training data set up on the model performance, we use both the balance and imbalance training data to train our investigated models[5]. Interestingly, for NIMGCN and its variant, the imbalance training set slightly boost up those models performance but it is not the case for HMDD3. For EPMDA, it works the other way round: balance training dataset results in a much higher AP score on HMDD2 but on HMDD3, it produces lower scores. Since the best AP score on HMDD2 is acquired by EPMDA with balance training set and the best AP score on HMDD3 is achieved by NIMGCN3 with balance training set, we recommend the use of balance training set.

Fig. 5. Mean AP scores (with standard deviation) corresponding to different similarity measures on HMDD2. The worst performing models are not shown in the plot.



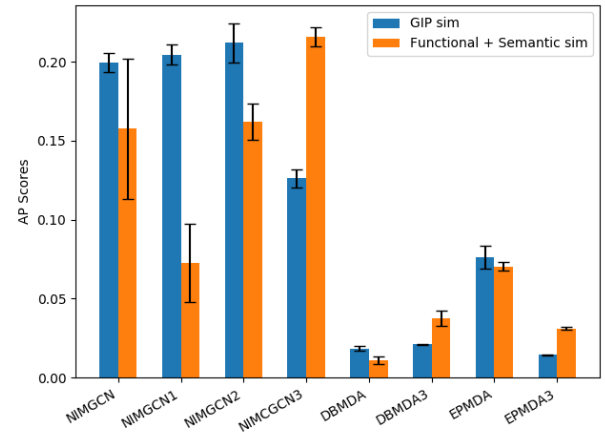
5.3 Impact of input similarity features

As discussed in Section 2, the input features corresponding to miRNA and disease nodes can be computed in several ways. In particular for disease, we can compute disease semantic, phenotype and `GIP` kernel similarities. EPMDA and DBMDA uses disease semantic similarity whereas EPMDA uses `GIP` kernel based similarity. For miRNA, we have miRNA functional similarity (used by DBMDA and NIMGCN), `GIP` kernel based miRNA similarity (used by EPMDA) and miRNA sequence similarity (used by DBMDA). In figures 6 and 5 we report mean AP scores comparing various models when they use (i) `GIP` similarity for disease and miRNA nodes (ii) `miRNA sequence+Semantic` similarity for HMDD2 and HMDD3 datasets respectively. We observe that `GIP` kernel based similarity measure which is directly computable on the association network and requires no other external information is in fact shows better performance in most of the models.

We now discuss the detailed results presented in tables 5, 6 and 7. Note that the column **`GIP`** refers to the case when for both miRNA and disease `GIP` kernel similarity features are used, **`Functional+Semantic`** denotes that miRNA fucntional similarity and disease semantic similarity were used and **`miRNA sequence+Semantic`** corresponds to using miRNA sequence similarity as input miRNA features and disease semantic similarity.

[5] The scores for NIMGCN, DBMDA and their variants are not reported here as significant improvements in performance were not observed

Fig. 6. Mean AP scores (with standard deviation) corresponding to different similarity measures on HMDD3. The worst performing models are not shown in the plot.



miRNA sequence similarity In terms of miRNA sequence similarity, we observe that in Table 6, DBMDA2 (DBMDA without miRNA sequence similarity) performs slightly better than DBMDA and DBMDA1 performs comparable to DBMDA3 (DBMDA1 without miRNA sequence similarity). That means removing miRNA sequence similarity from DBMDA or DBMDA1 does not have any significant impact on those models performance. Therefore, it is obvious that miRNA sequence similarity has very little effect on DBMDA performance.

The results corresponding to the use of miRNA functional and miRNA sequence similarity as the only source of features for miRNA are presented in the `Functional+Semantic` and `miRNA sequence+Semantic` columns of tables 5–7, respectively. Taking a closer look at the given AP scores we can see that overall, for the two experimented data sets, miRNA functional similarity which takes into account the known miRNA-disease associations is more informative than miRNA sequence similarity. However, miRNA sequence similarity still has its benefit for new miRNA where no known miRNA-disease association is available. When miRNA sequence similarity is to be used, NIMGCN2 or NIMGCN3 should be given more preference.

Effect of `GIP` kernel similarity. Figure 5 and 6 present the results for different models with different similarity measures as input. Looking at the plots we can see that on HMDD2, `GIP` similarities, which are based on the Gaussian Interaction Profile kernel, results in better performance of NIMGCN2 (which is built up from neural projections only) but significantly lower performance of NIMGCN3 on both data sets. In addition, EPMDA with `GIP` similarities result in the highest AP scores on HMDD2, which is significantly higher than the other methods. On HMDD3, though the highest AP score is the result of NIMGCN3 and `Functional+Semantic`, AP score of `GIP` similarities in other methods tend to be higher than the other combination.

Additionally, we also experimented with the two different disease similarity scores as described in equations (2) and (4). However, we find that overall, the two similarity measures perform comparable to each other and therefore we present results only corresponding to using disease semantic similarity given in Equation (4).

6 Conclusion and recommendations

In this work, we investigate three state-of-the-art models and several simpler variants of them thereby comparing a total of 12 models

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

on a common experimental framework which we also release for reproducibility of our results and future development. We report AP and AUROC scores by running 5 times 5-Fold cross validation. For each fold, we use a complete test set miRNA-disease associations (consisting of all possible pair combination except the positive links in the training set) and recompute the similarity features each time by using the information only from the training set.

We unearth the problem of data leakage in machine learning models which employ pre-computed miRNA functional similarities or `GIP` kernel similarities computed over the complete association network. As a part or whole of test set associations were used in the feature computation, the reported results overestimate actual performance. In this work, we computed similarity features each time corresponding to the training set and showed that the EPMDA model relying completely on association network for `GIP` kernel based miRNA and disease similarity features show a larger drop in performance when correct feature computation method was used.

In addition we showed that several of our proposed simpler variants of the studied state of the art techniques outperform the actual models in our experimented data sets. Our proposed variants have resulted in improvements in Average Precision scores (as compared to original models) by approximately 287.7% and 36.7% on HMDDv2.0 (AP: 0.504) and HMDDv3.0 (AP: 0.216) datasets respectively. Apart from measuring the importance of various components of the actual models, we also compare the models using different kinds of miRNA and disease similarity features. The results suggest that the use of `GIP` similarities as input often produce higher performance and hence is recommended for use in future models.

Our results indicate that a carefully chosen set of input features works better than adding more complexity/non-linearity to the model. We also find that the methods which take into account of the topological structure of the miRNA-disease network perform better. We discover that one of EPMDA variants which first extracts edge perturbation based features is the best performing method for HMDD2 dataset. For the larger HMDD3 dataset, EPMDA's feature extraction step is very expensive and the hyperparameter corresponding to the length of longest cycle might need to be fine tuned to improve its performance on HMDD3. We therefore recommend the design of deep learning methods based on graph neural networks for dataset dependent automatic feature extraction from the graph topology.

7 Acknowledgement

We would like to thank the authors of EPMDA and NIMGCN for sharing their code, data and spending time in answering our questions.

References

[1]X. Chen, Y. Gong, D.-H. Zhang, Z.-H. You, and Z.-W. Li. Drmda: deep representations-based mirna–disease association prediction. *Journal of cellular and molecular medicine*, 22(1):472–485, 2018.

[2]X. Chen, M.-X. Liu, and G.-Y. Yan. Rwrmda: predicting novel human microrna–disease associations. *Molecular BioSystems*, 8(10):2792–2798, 2012.

[3]X. Chen, C. C. Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang, and Q. Dai. Rbmmmda: predicting multiple types of disease-microrna associations. *Scientific reports*, 5:13877, 2015.

[4]X. Chen, D.-H. Zhang, and Z.-H. You. A heterogeneous label propagation approach to explore the potential associations between mirna and disease. *Journal of translational medicine*, 16(1):348, 2018.

[5]Y. Dong, Y. Sun, C. Qin, and W. Zhu. Epmda: Edge perturbation based method for mirna-disease association prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[6]L. Fu and Q. Peng. A deep ensemble model to predict mirna-disease association. *Scientific reports*, 7(1):1–13, 2017.

[7]Z. Huang, L. Liu, Y. Gao, J. Shi, Q. Cui, J. Li, and Y. Zhou. Benchmark of computational methods for predicting microrna-disease associations. *Genome biology*, 20(1):1–13, 2019.

[8]Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui. Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. *Nucleic acids research*, 47(D1):D1013–D1017, 2019.

[9]T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[10]A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019.

[11]G. Li, J. Luo, Q. Xiao, C. Liang, and P. Ding. Predicting microrna-disease associations using label propagation based on linear neighborhood similarity. *Journal of biomedical informatics*, 82:169–177, 2018.

[12]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*, 2020.

[13]Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui. Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research*, 42(D1):D1070–D1074, 2014.

[14]J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006.

[15]E. M. Small, R. J. Frost, and E. N. Olson. Micrornas add a new dimension to cardiovascular disease. *Circulation*, 121(8):1022–1032, 2010.

[16]T. van Laarhoven, S. B. Nabuurs, and E. Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.

[17]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*, 26(13):1644–1650, 2010.

[18]P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, et al. Prediction of micrornas associated with human diseases based on weighted k most similar neighbors. *PloS one*, 8(8), 2013.

[19]Y. Yang, R. N. Lichtenwalter, and N. V. Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[20]Z. Yang, L. Wu, A. Wang, W. Tang, Y. Zhao, H. Zhao, and A. E. Teschendorff. dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic acids research*, 45(D1):D812–D818, 2017.

[21]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. *Molecular Therapy-Nucleic Acids*, 19:602–611, 2020.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture