A Message Passing framework with Multiple data integration for miRNA-Disease association prediction

Thi Ngan Dong (dong@l3s.de)
Leibniz University of Hannover
Johanna Schrader
Leibniz University of Hannover
Stefanie Mücke
TRAIN Omics, Translational Alliance in Lower Saxony
Megha Khosla
Delft University of Technology (TU Delft)

Article

Keywords:

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1559305/v1

A Message Passing framework with Multiple data integration for miRNA-Disease association prediction

Thi Ngan Dong[1,*], **Johanna Schrader**[1], **Stefanie Mücke**[2], **and Megha Khosla**[3]

[*]dong@l3s.de
[1]L3S Research Center, Leibniz University of Hannover, Hannover, Germany
[2]TRAIN Omics, Translational Alliance in Lower Saxony, Hannover, Germany
[3]Delft University of Technology (TU Delft), Netherlands

ABSTRACT

Micro RNA or miRNA is a highly conserved class of non-coding RNA that plays an important role in many diseases. Identifying miRNA-disease associations can pave the way for better clinical diagnosis and finding potential drug targets.
We propose a biologically-motivated data-driven approach for the miRNA-disease association prediction, which overcomes the data scarcity problem by exploiting information from multiple data sources. The key idea is to enrich the existing miRNA/disease-protein-coding gene (PCG) associations via a Message Passing framework, followed by the use of disease ontology information for further feature filtering. The enriched and filtered PCG associations are then used to construct the inter-connected miRNA-PCG-disease network to train a structural deep network embedding (SDNE) model. Finally, the pre-trained embedding and the biologically relevant features from the miRNA family and disease semantic similarity are concatenated to form the pair input representation to a Random Forest classifier whose task is to predict the miRNA-disease association probability.
We present large-scale comparative experiments, ablation, and case studies to showcase our approach's superiority. Besides, we make the model prediction results for 1,618 miRNAs and 3,679 diseases, along with all related information publicly available at http://software.mpm.leibniz-ai-lab.de/ to foster assessments and future adoption.

1 Introduction

Micro RNAs or miRNAs are a conserved class of non-coding RNAs that regulate the expression of protein-coding genes (PCGs). MiRNAs affect the translation process of about one-third of PCGs[1] and play key roles in various biological processes and most stages of human development[2–5].

Proteins are responsible for essential biological functions inside living organisms. Disruptions in proteins' expressions are directly associated with various disease conditions[6]. Therefore, miRNAs could serve as potential biomarkers in certain diseases such as cancers or immune-related diseases[7–13]. Identifying potential associations between miRNAs and diseases can further help in clinical diagnosis and finding potential drug targets.

While biological experiments are usually expensive and time-consuming, computational approaches, especially data-driven machine learning (ML) approaches,[14–16] can assist wet-lab experiments by predicting a potential set of associations. Early works[17–22] focus on learning effective miRNA/disease representation from the set of known association data. The feature extraction process usually involves the computation of hand-crafted similarities. Besides the data leakage problem as already discussed in our previous work[23], similarity-based techniques are biased towards the well-studied miRNAs and diseases[16]. Ultimately, they are derived from some hard-coded heuristics and assumptions, which might work effectively on the observed association set but usually do not generalize well to the unseen miRNAs or diseases[16,17]. Besides, the hard-coded heuristics cannot fully exploit the potential of the available information, especially with regard to the association patterns or the motif/frequent subgraphs inside the miRNA-disease bipartite graph constructed from the known association set.

Graph representation learning techniques acquired state-of-the-art performance on many machine learning problems[24]. They have already been applied for the miRNA-disease association prediction problem by recent works[14,19,21,22,25,26]. Nevertheless, a majority of the proposed models operate on the similarity network(s) constructed from hand-crafted similarity measures, not the raw miRNA-disease association data. Therefore, they cannot fully exploit the existing information, especially the structure patterns inside the raw association bipartite graph. A recent work[27] proposes the use of a structural deep network embedding (SDNE) model to mine the network information directly from the miRNA-disease association graph. Nonetheless, new miRNAs or new diseases appear as isolated nodes for which SDNE cannot learn any useful representation. Therefore, the existing model still has limited prediction capability for new miRNAs or new diseases.

Recent works focus more on information integration to overcome the data scarcity problem. NEMII[27] adds miRNA family and disease semantic similarities to enrich the miRNA-disease pair representation. MMGCN[28] proposes a multi-attention

mechanism to combine multiple similarity-based measures. NNMDA[25] proposes a weighted sum over five different miRNA similarities and two disease similarities to build a heterogeneous network for feature learning and association prediction. Ji et al.[26] incorporate information from multiple domains, for example, miRNA-lncRNA and miRNA-PCG interaction, miRNA-drug association, disease-lncRNA, disease-PCG association, disease-drug association, to build a heterogeneous information network for feature extraction. Though promising, with respect to the added side information, current works either employ the whole raw dataset or apply naive filtering steps based on the association confidence score deposited in the databases. Such naive filtering does not ensure the quality of the integrated data. Subsequently, the quality of the trained model suffers.

To this end, we propose a biologically-motivated data-driven approach that aims to counter the above challenges by jointly learning from multiple data sources. We refer to our approach as MPM. A crucial design decision of our approach includes modeling the biological relevance of miRNAs for a particular disease via the associated PCGs. We model each miRNA or disease as a directed network built up from the miRNA-PCG, disease-PCG associations, and PCG-PCG functional interactions. MPM employs a message passing framework operating over the constructed networks to enrich the existing data with potential missing links or indirect connections. To overcome the noisy data problem, we employ a feature selection strategy with a side-supervised task generated from the well-annotated MESH ontology[29]. Feature selection at this stage allows us to reduce the tens of thousands of associated PCGs to only one hundred most important PCGs. This enables us to control the quality and the quantity of the added PCG-related information without introducing any additional parameters. This is extremely important, especially in the context of learning over scarce data when overparameterized models can easily overfit.

Next we encapsulate the enriched and filtered PCGs connections into the existing miRNA-disease bipartite network to overcome the isolated node problem in existing works. Since PCGs are important connections between miRNA and diseases[6], the patterns learned from the miRNA-PCG-disease interconnected networks should be a rich source of information for the miRNA-disease association prediction problem. At the same time, the newly introduced heterogeneous network will include biological connections between new miRNAs or new diseases and their associated PCGs. The learning signals will thus transfer from known miRNAs or known diseases to the new miRNAs or new diseases via the PCGs. MPM employs an SDNE model to extract the patterns (or pre-trained embeddings) from the constructed heterogeneous network. Besides the structural features, the final miRNA-disease pair representation is further augmented with information from the miRNA family and disease semantic similarity and then fed as input to a Random Forest classifier to perform the association prediction task.

In summary, we propose flexible information integration mechanisms at different stages of the model building process to overcome the data scarcity problem. In addition to fusing multiple knowledge sources, we propose a parameter-free mechanism to enrich and control the quality and quantity of the added data. Experimental results on 21 large independent test sets, two case studies indicate that our proposed model significantly outperforms all benchmarked models in both (i) the transductive setting where we test each model performance on the set of partially observed miRNAs and diseases and (ii) the inductive setting where we test the model performance on the set of completely new miRNAs and new diseases. The ablation study results also support our choice of architecture.

We share all the code, pre-processed, and standardized data at https://git.l3s.uni-hannover.de/dong/mpm. In addition, we make the predicted association probabilities for all 1,618 miRNAs and 3,679 diseases publicly available at http://software.mpm.leibniz-ai-lab.de/index.html. To enable a smooth and comprehensive analysis, we also integrate the miRNAs and diseases pathway and functional enrichment analysis results into the website. Section 2.6 and Section 3 in the supplementary file provide more details regarding our website and the integrated information sources.

2 Results

2.1 Compared models

We compare our model with six recently proposed methods: (i) EPMDA[14], DBMDA[15], and NIMGCN[30], which utilize hand-crafted features derived from known miRNA-disease associations, (ii) MUCOMID[16] and DIMIG 2.0[31], which use graph convolution networks (GCNs) for feature extraction from various interaction networks (iii) NEMII[27] which employs hand-crafted features as well as the latent features extracted using a graph embedding method. As an ablation study, we compare MPM with four of its simpler variants as described in Table 1. A detailed description of the compared models is provided in Section 1 in the supplementary file. Details on hyperparameter settings and implementation for all models are provided in Section 2.6 in the supplementary file.

2.2 Evaluation setup

The testing setup. While the K-fold cross-validation technique is widely used among existing works, it is insufficient to evaluate the models' performances on completely new diseases, given the small size of the association datasets. We here propose and employ two realistic testing setups: *transductive* and *inductive* to evaluate and compare models. The transductive testing setup aims at evaluating different models' performances on the set of miRNAs and diseases when some of their associations

Table 1. MPM simpler variants where '✓' and 'X' denote the existence and non-existence of the corresponding components/modules.

Model	Message Passing	Feature Selection	SDNE	Random Forest classifier	PCG associations
MPM-NO-MP	X	✓	✓	✓	✓
MPM-NO-FS	✓	X	✓	✓	✓
MPM-NO-SDNE	✓	✓	X	✓	✓
NEMII[27]	X	X	✓	✓	X
MPM-NO-MPFS	X	X	✓	✓	✓

have already been observed during the training phase. In this setup, we train each model on the HMDD2 dataset[32] and test it on the HELD-OUT1 test set. The inductive testing setup aims at evaluating models' performances on completely new diseases and new miRNAs. In this setup, we conduct large scale experiments on the 20 independent test sets to test each model performance on the (i) dataset with many new miRNAs (the NOVEL-MIRNA test set), (ii) 18 complete test sets for new diseases, and (iii) dataset with many new miRNAs and new diseases (the HELD-OUT2 test set). Details about the data sources, data preprocessing, and how we generate the training and testing data in both testing setups are presented in Section 2 in the supplementary file. All datasets' statistics are presented in Table 2 and Table 3 in the supplementary file.

Table 2. The association data statistics where $|n_{md}|$, $|n_m|$, $|n_d|$ refer to the number of associations, miRNAs and diseases respectively.

DATASET	$\|n_{md}\|$	$\|n_m\|$	$\|n_d\|$
HMDD2	4,592	442	309
HMDD3	10,494	742	545
HMDD2 ∪ HMDD3	10,980	742	591
HELD-OUT1	4,311	382	226
HELD-OUT2	6,388	697	509
NOVEL-MIRNA	4,734	638	227

Evaluation Metrics. By looking at the generated results, we notice that the prediction probability ranges generated by different models are different. For that reason, employing threshold-based evaluation metrics like Precision, Recall, or F1 might result in unfair comparison. Therefore, we report only the non-parametric metrics that are the Area under the Receiver Operating Characteristic (AUC), the Average Precision (AP) (which summarizes the Precision-Recall curve). We report the AP instead of the AUPR score because AP provides a better performance estimate than the AUPR, as discussed in our previous work in[23]. AP is calculated as the discrete sum of the changes in the recall at different thresholds instead of linear interpolation as that of AUPR, which can be too optimistic in cases where the number of thresholds (unique prediction values) is limited[33,34]. For all tables, bold font is used to highlight the best scores.

2.3 MPM vs. existing works (SOTA)

Tables 3 and 4 present the average performance scores for all benchmarked models on our 21 large test sets. In Table 3, we report the average AP and AUC scores corresponding to different positive:negative testing sample rates. We do not have the results for EPMDA on the 18 test sets for new diseases because all pair representations are zeros since new diseases appear as isolated nodes in the network for topology-based feature extraction.

In general, MPM outperforms all benchmarked models (SOTA) on the HELD-OUT1 (transductive setting), NOVEL-MIRNA (with many new miRNAs), and HELD-OUT2 (with new miRNAs and new diseases) test sets with a gain of up to 11.5% in AP score. The gains are more significant when more negative samples are added to the testing data. On the complete test sets for new diseases, MPM gains the highest AP score in 17 out of 18 datasets.

In both transductive and inductive testing setups, we observe similar trends with large performance gaps among the state-of-the-art methods. Generally, DIMIG 2.0 always performs the worst, followed by NIMGCN, then DBMDA, EPMDA, MUCOMID, and then NEMII. DIMIG 2.0 is a recently proposed model that formulates the miRNA-disease association prediction problem as a semi-supervised node classification task with diseases as labels. The model can integrate information from four additional knowledge sources (miRNA-PCG, disease-PCG associations, PCG-PCG interactions, and disease ontology) but only performs training using the the known disease-PCG associations set. Though DIMIG 2.0 can generate predictions for new miRNAs and new diseases, the large and sparse label set, along with the weak training signals leads to its limited predictive performance.

Table 3. Results for all models on the three large independent test sets. $nr = 1$, $nr = 5$, $nr = 10$ indicate that we test all models with the positive:negative rates of 1:1, 1:5, 1:10, respectively. Bold font is used to highlight the best scores.

Method	HELD-OUT1						NOVEL-MIRNA						HELD-OUT2					
	$nr = 1$		$nr = 5$		$nr = 10$		$nr = 1$		$nr = 5$		$nr = 10$		$nr = 1$		$nr = 5$		$nr = 10$	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
NIMGCN	0.542	0.554	0.541	0.207	0.542	0.118	0.532	0.549	0.53	0.202	0.53	0.115	0.513	0.517	0.513	0.176	0.512	0.097
DBMDA	0.657	0.622	0.656	0.256	0.656	0.149	0.644	0.621	0.645	0.261	0.645	0.153	0.638	0.617	0.638	0.257	0.638	0.15
EPMDA	0.698	0.624	0.698	0.256	0.698	0.148	0.716	0.643	0.718	0.281	0.719	0.167	0.704	0.648	0.703	0.291	0.704	0.176
NEMII	0.838	0.831	0.838	0.542	0.838	0.395	0.865	0.857	0.866	0.597	0.866	0.452	0.859	0.853	0.859	0.581	0.858	0.435
MUCOMID	0.832	0.826	0.832	0.534	0.832	0.385	0.827	0.819	0.827	0.519	0.827	0.37	0.811	0.812	0.812	0.514	0.811	0.368
DIMIG 2.0	0.499	0.5	0.499	0.167	0.499	0.091	0.499	0.5	0.499	0.167	0.499	0.091	0.499	0.5	0.499	0.167	0.499	0.091
SOTA Improvement	1.2%	1.6%	1.2%	5.7%	1.2%	8.6%	0.5%	1.1%	0.5%	3.4%	0.5%	6.0%	0.5%	1.4%	0.5%	6.9%	0.5%	11.5%
MPM-NO-MP	0.846	0.84	0.846	0.564	0.847	0.418	0.866	0.859	0.866	0.602	0.867	0.46	0.859	0.86	0.859	0.607	0.859	0.468
MPM-NO-FS	0.814	0.809	0.814	0.503	0.814	0.357	0.823	0.818	0.823	0.519	0.823	0.373	0.814	0.819	0.814	0.533	0.814	0.391
MPM-NO-MPFS	0.824	0.816	0.824	0.516	0.824	0.369	0.836	0.828	0.836	0.538	0.836	0.391	0.831	0.832	0.831	0.554	0.831	0.411
MPM-NO-SDNE	0.837	0.83	0.837	0.546	0.837	0.401	0.842	0.834	0.842	0.552	0.843	0.408	0.846	0.847	0.846	0.581	0.846	0.439
MPM (ours)	**0.848**	**0.844**	**0.848**	**0.573**	**0.848**	**0.429**	**0.869**	**0.866**	**0.87**	**0.62**	**0.87**	**0.479**	**0.863**	**0.865**	**0.863**	**0.621**	**0.862**	**0.485**

Table 4. The AP scores corresponding to the 18 complete test sets for new diseases average over 20 experimental runs.

Disease	MPM	NIMGCN	DBMDA	NEMII	MUCOMID	DIMIG 2.0	MPM-NO -MP	MPM-NO -FS	MPM-NO -MPFS	MPM-NO -SDNE
D001749	**0.782**	0.12	0.274	0.77	0.494	0.119	0.77	0.567	0.589	0.58
D001943	0.82	0.2	0.556	**0.825**	0.432	0.119	0.811	0.679	0.693	0.654
D002289	**0.803**	0.127	0.336	0.801	0.441	0.114	0.795	0.662	0.678	0.589
D002292	**0.679**	0.105	0.187	0.66	0.32	0.095	0.67	0.51	0.525	0.531
D002294	**0.655**	0.188	0.25	0.612	0.316	0.096	0.646	0.529	0.531	0.493
D003110	**0.664**	0.124	0.241	0.617	0.349	0.089	0.619	0.487	0.54	0.515
D005909	**0.737**	0.156	0.352	0.707	0.44	0.079	0.726	0.597	0.63	0.523
D005910	0.761	0.2	0.287	0.732	0.403	0.082	**0.767**	0.642	0.66	0.626
D006333	0.669	0.075	0.287	0.653	0.311	0.078	**0.671**	0.578	0.602	0.566
D008175	**0.759**	0.147	0.421	0.74	0.409	0.073	0.751	0.615	0.62	0.611
D008545	**0.725**	0.16	0.417	0.706	0.32	0.073	0.715	0.58	0.598	0.558
D010051	**0.796**	0.184	0.37	0.741	0.406	0.064	0.782	0.505	0.654	0.579
D010190	**0.77**	0.133	0.359	0.746	0.438	0.069	0.761	0.589	0.622	0.598
D011471	0.724	0.144	0.369	0.66	0.349	0.061	**0.738**	0.618	0.633	0.569
D012516	**0.705**	0.3	0.274	0.659	0.374	0.057	0.699	0.546	0.585	0.55
D013274	**0.838**	0.218	0.522	0.834	0.345	0.051	0.811	0.657	0.693	0.643
D015179	**0.806**	0.189	0.486	0.797	0.451	0.047	0.785	0.645	0.693	0.614
D015470	**0.657**	0.208	0.253	0.623	0.349	0.041	0.653	0.509	0.513	0.497

NIMGCN performs the worst compared to other supervised baselines because it only relies on the miRNA functional and disease semantic similarities to construct the networks for the feature learning. The miRNA functional similarity is heavily biased towards well-known diseases and cannot generalize well to new diseases[17]. Also, new miRNAs appear as isolated nodes in the network and will get completely random representations. Therefore, NIMGCN's prediction capability is limited for the little known or completely new miRNAs or diseases.

Regarding the input sources, DBMDA improves over NIMGCN by integrating another biologically-related information source: the miRNA sequence similarity. DBMDA gains significantly better performance than NIMGCN but is still much lower than MUCOMID, NEMII, and MPM, suggesting that the miRNA sequence similarity does bring additional benefit but not too significant.

EPMDA proposes a topologically related feature extraction technique for miRNA-disease pair representation. Unlike most existing works, which focus on learning effective representation for miRNA and disease separately, EPMDA learns the miRNA-disease pair representation directly as a property of the miRNA-disease heterogeneous network constructed from the miRNA and disease Gaussian Interaction Profile kernel and the miRNA-disease known associations. Even though EPMDA does not employ any additional information source, its performance is still better than NIMGCN and DBMDA. This suggests that learning the pair representation directly from the heterogeneous network with raw miRNA-disease association is a fruitful

direction. Nonetheless, the edge perturbation score has at least $O(n^3)$ time complexity and cannot scale well to a large network[23]. Besides, fine-tuning the network cycle length parameter is not a trivial task[23].

MUCOMID proposes a multitask learning model that integrates five additional information sources to overcome the data scarcity problem. Though promising, the model applies hard-threshold filtering to filter out redundant information in the additional information sources. The results reported in Tables 3 and 4 corresponds to MUCOMID's performance without any information filtering (since not all part of our data have the interaction/association confidence scores available). The thresholds need to be fine-tuned for each dataset separately. For that reason, it requires considerable time and effort for parameter fine-tuning in order to employ MUCOMID for a completely new dataset. This points to an important aspect of information integration which focuses on effectively controlling/managing the quality and quantity of the added knowledge sources. Nonetheless, the method shows promising performance, which overcomes problems associated with hand-crafted similarity-based methods in all testing setups.

NEMII learns a structural embedding directly from the miRNA-disease bipartite network constructed from the known miRNA-disease association data. Besides, the model is further informed by information from the miRNA family and disease semantic similarity. Though new miRNAs and new diseases get completely random representation from the structural embedding learning module, NEMII's performance on the 20 inductive testing datasets is still one of the highest, thanks to the biological information from the miRNA family and disease semantic similarity features. Overall, the effective feature extraction strategy, combined with the domain knowledge from the added side information sources, helped NEMII gain the highest performance scores among state-of-the-art methods on all testing datasets. These results support the exploitation of structural information from the miRNA-disease association data and the importance of information integration.

MPM improves over state-of-the-art methods with a parameter-free yet effective mechanism to control the quality and quantity of the added information sources. At the same time, it addresses the existing limitation in the NEMII model by integrating additional biological relations to the new miRNAs and new diseases. The learned signals from the well-studied miRNAs/diseases will be transferred to the diseases with available scarce data via their associated PCGs. These improvements prove to be quite effective and help MPM gain state-of-the-art performance on 20 out of the 21 benchmarked datasets in both transductive and inductive testing setup with a gain of up to 11.5% in AP score.

2.4 An ablation study

Here we compare MPM with four of its simpler variants as described in Table 1. MPM-NO-MP is a variant of MPM without the Message Passing layer that takes the raw miRNA-PCG, disease-PCG associations as input to the feature selection and structural embedding learning modules. Similarly, MPM-NO-FS is a variant of MPM without the feature selection module. The structural embedding encapsulates all enriched miRNA-PCG and disease-PCG associations output from the message passing layer into its heterogeneous network for learning node embeddings. MPM-NO-MPFS is a variant of MPM without the message passing and the feature selection modules. The heterogeneous network input to SDNE simply integrate all raw miRNA-PCG, disease-PCG associations retrieved from miRTarBase[35] and DisGeNET[36]. MPM-NO-SDNE is a variant of MPM in which there is no structural embedding learning. Instead, the pair representation for a particular miRNA-disease pair is the concatenation of the enriched and filtered miRNA-PCG, disease-PCG associations, miRNA family, and disease semantic similarity features.

Table 3 presents the results for MPM and its variants on three large independent test sets. Table 4 reports the results for the 18 inductive testing datasets for new diseases. We observe that MPM supersedes all of its simpler variants on the transductive testing set (HELD-OUT1), two inductive testing sets with many new miRNAs, and 15 out of 18 complete test sets for new diseases. The gains are the most significant on the three independent test sets (c.f. Table 3), especially when more negative testing samples are added. These results support the contribution of each added component. At the same time, they validate our choice of the architecture.

Besides, among the simpler variants, we observe considerable performance drops on the variants without the Feature Selection modules (MPM-NO-FS and MPM-NO-MPFS) or on the MPM-NO-SDNE model. Without the feature selection module, the network employed for embedding generation contains too many PCG association connections. As biological data usually contains many false positives, adding all PCG associations introduces additional noise and redundancy. Similarly, without the structural embedding (MPM-NO-SDNE), MPM only relies on the associated PCGs, miRNA, and disease semantic similarity features to generate prediction without information about the miRNA/disease interaction patterns. The drops in performance observed in MPM's simpler variants further emphasize the importance of information filtering as well as the raw association structural patterns.

2.5 Case studies

Let $\mathbf{H} = \text{HMDD2} \cup \text{HMDD3}$ be the set of all known associations retrieved from the HMDD databases. We here present two case studies to showcase the application of MPM in a real scenario.

2.5.1 MPM for a disease with scarce knowledge

Down syndrome or Trisomy 21 is a condition in which a child is born with an extra copy of their 21st chromosome[37]. *Down Syndrome*'s patients usually suffer from mild-to-moderate learning disabilities[37]. According to the data deposited in the HMDD 2.0 and HMDD 3.0 databases and two recent papers[38,39], there are only 10 miRNAs known to be associated with the disease of our interest. We assume that *Down Syndrome* is a completely new disease and take similar steps as those presented in Section 2.4.2 in the supplementary file to construct the training and testing data. In short, our training data consists of all known associations in **H** for diseases other than the *Down Syndrome*. We test MPM on the complete test set consisting of all possible combinations between the *Down Syndrome* and 1,618 miRNAs.

Table 5. MPM's average prediction scores for *Down Syndrome* and all 1,618 miRNAs. The associated miRNAs are marked as blue. The model training data does not contain the association data for *Down Syndrome*.

Rank	miRNA	pred.	Rank	miRNA	pred.
	...		82	hsa-mir-125b-2	0.579253110400618
2	hsa-mir-155	0.963881105523116		...	
3	hsa-mir-146a	0.934014942433006	105	hsa-mir-99a	0.482246263067031
4	hsa-mir-16-1	0.895608127697913		...	
	...		140	hsa-mir-1246	0.404202397336283
33	hsa-mir-27b	0.689694528927961		...	
	...		261	hsa-let-7c	0.244887327696169
38	hsa-mir-27a	0.671913693062923		...	
	...		1576	hsa-mir-802	0.130087980984639

How effective is MPM in restricting and prioritizing the search space for the potentially associated miRNAs? Table 5 presents the average predictions made by MPM after 20 experimental runs. Though we performed the searching on a complete test set of 1,618 testing samples, 3 known-to-associate miRNAs (marked as blue in Table 5 already appear in the top 4 highest predicted results. The other associated miRNAs appeared at 33th, 38th, 82th, 105th, 140th, 261th, and 1576th positions in the prediction list. With 3 appearing in the top 4 and 5 out of 10 known associations appearing in the top 38 of the generated prediction results, our method would significantly help restrict and prioritize the search space for wet-lab experiments.

How effective MPM is with some added domain knowledge? Since *Down Syndrome* relates to a redundant chromosome 21 copy, we retrieve the miRNA location information from miRTarBase[35] and present the MPM's predicted results for all miRNAs located on chromosome 21 in Table 6. Blue is used to mark the associated miRNAs.

Table 6. MPM's prediction results for *Down Syndrome* and the miRNAs that are located on chromosome 21. Blue is used to highlight the associated miRNAs. The model training data does not contain the association data for *Down Syndrome*.

rank	miRNA	pred.	rank	miRNA	pred.
1	hsa-mir-155	0.963881105523116	11	hsa-mir-4760	0.172962854437391
2	hsa-mir-125b-2	0.579253110400618	12	hsa-mir-5692b	0.168364046134056
3	hsa-mir-99a	0.482246263067031	13	hsa-mir-6508	0.163143029370321
4	hsa-let-7c	0.244887327696169	14	hsa-mir-6070	0.16232917173827
5	hsa-mir-548x	0.239129159103197	15	hsa-mir-6815	0.159395572782035
6	hsa-mir-3648-1	0.206785057828119	16	hsa-mir-8069-1	0.155993241075239
7	hsa-mir-4759	0.200771150543586	17	hsa-mir-6724-1	0.153456269809843
8	hsa-mir-3197	0.19795748172893	18	hsa-mir-6501	0.152740622433185
9	hsa-mir-6130	0.194382789321313	19	hsa-mir-6814	0.145666592873055
10	hsa-mir-4327	0.176297567535453	20	hsa-mir-802	0.130087980984639

By restricting the miRNA search space, we have much more promising prediction results, with 4 out of 5 associated miRNAs appearing at the top of the list. Adding more related domain information like chromosomal location, tissue expression profiles, etc., would help restrict the miRNA search space, thus, would further help in generating more meaningful prediction results. Nonetheless, we release predicted association probabilities for all 1,618 miRNAs to encourage field experts' assessments as well as to enable them to customize the search without requiring to retrain/rerun the model.

2.5.2 MPM for a disease with many false positives

Parkinson disease (PD) is the second most common neurodegenerative disease worldwide[40]. Existing human association studies for the *Parkinson* disease resulted in inconsistent findings with many "false positives" as reported in[41]. In this case

study, we take a closer look at the generated predictions from MPM for the *Parkinson* disease. We train MPM with all the available data in **H**. More specifically, besides the data for other diseases, the training data contains 61 known associations for *Parkinson*. Among those, there are 8 true positives (those that are confirmed as positives in[41]) and 26 false positives[41] (those that are marked as positive in **H** but are confirmed as negative in[41]).

Table 7. The predicted association probabilities for the *true positive* (marked as blue) and *true negative* miRNAs[41] corresponding to the *Parkinson* disease.

rank	miRNA	pred.	rank	miRNA	pred.	rank	miRNA	pred.	rank	miRNA	pred.	rank	miRNA	pred.
1	hsa-mir-7-1	0.99	23	hsa-mir-127	0.96	45	hsa-mir-99a	0.92	67	hsa-mir-25	0.85	89	hsa-mir-149	0.62
2	hsa-mir-30d	0.99	24	hsa-mir-145	0.96	46	hsa-mir-19a	0.92	68	hsa-mir-23a	0.85	90	hsa-mir-1264	0.62
3	hsa-mir-19b-1	0.99	25	hsa-mir-195	0.96	47	hsa-mir-29c	0.92	69	hsa-mir-191	0.85	91	hsa-mir-744	0.61
4	hsa-mir-146a	0.99	26	hsa-mir-497	0.96	48	hsa-mir-1301	0.91	70	hsa-mir-140	0.84	92	hsa-mir-301b	0.6
5	hsa-mir-335	0.99	27	hsa-mir-338	0.96	49	hsa-mir-30b	0.91	71	hsa-mir-136	0.83	93	hsa-mir-154	0.59
6	hsa-mir-193a	0.99	28	hsa-mir-222	0.96	50	hsa-mir-152	0.9	72	hsa-mir-16-2	0.82	94	hsa-mir-184	0.55
7	hsa-mir-214	0.98	29	hsa-mir-221	0.96	51	hsa-mir-125b-2	0.9	73	hsa-mir-98	0.82	95	hsa-mir-223	0.54
8	hsa-mir-141	0.98	30	hsa-mir-22	0.96	52	hsa-mir-125a	0.9	74	hsa-mir-27b	0.81	96	hsa-mir-532	0.49
9	hsa-mir-151a	0.98	31	hsa-mir-299	0.96	53	hsa-mir-137	0.9	75	hsa-mir-345	0.81	97	hsa-mir-1296	0.48
10	hsa-mir-126	0.98	32	hsa-mir-424	0.95	54	hsa-mir-204	0.89	76	hsa-mir-142	0.8	98	hsa-mir-873	0.44
11	hsa-mir-7-2	0.98	33	hsa-mir-21	0.95	55	hsa-mir-224	0.89	77	hsa-mir-708	0.8	99	hsa-mir-125b-1	0.42
12	hsa-mir-146b	0.98	34	hsa-mir-17	0.95	56	hsa-mir-148b	0.89	78	hsa-mir-1249	0.78	100	hsa-mir-1298	0.35
13	hsa-mir-29b-2	0.98	35	hsa-mir-148a	0.94	57	hsa-mir-409	0.89	79	hsa-mir-190a	0.78	101	hsa-mir-939	0.34
14	hsa-mir-30a	0.98	36	hsa-mir-143	0.94	58	hsa-mir-504	0.89	80	hsa-mir-129-1	0.77	102	hsa-mir-488	0.29
15	hsa-mir-199b	0.98	37	hsa-mir-28	0.94	59	hsa-mir-186	0.89	81	hsa-mir-331	0.76	103	hsa-mir-330	0.24
16	hsa-mir-34c	0.98	38	hsa-mir-425	0.93	60	hsa-mir-448	0.88	82	hsa-mir-181c	0.75	104	hsa-mir-192	0.2
17	hsa-mir-132	0.98	39	hsa-mir-10b	0.93	61	hsa-mir-769	0.87	83	hsa-mir-150	0.73	105	hsa-mir-626	0.19
18	hsa-mir-451a	0.97	40	hsa-mir-29a	0.93	62	hsa-mir-1248	0.87	84	hsa-mir-489	0.72	106	hsa-mir-26b	0.16
19	hsa-mir-133b	0.97	41	hsa-mir-99b	0.93	63	hsa-mir-92a-2	0.87	85	hsa-mir-505	0.68	107	hsa-mir-577	0.16
20	hsa-mir-10a	0.97	42	hsa-mir-543	0.93	64	hsa-mir-328	0.86	86	hsa-mir-203a	0.67	108	hsa-mir-654	0.15
21	hsa-mir-16-1	0.97	43	hsa-mir-34b	0.93	65	hsa-mir-92a-1	0.86	87	hsa-mir-454	0.65	109	hsa-mir-378a	0.15
22	hsa-mir-30c-2	0.97	44	hsa-mir-431	0.92	66	hsa-mir-20a	0.85	88	hsa-mir-130a	0.64	110	hsa-mir-501	0.12

We present the predicted association probabilities for all 12 *true positive* and 98 *true negative* miRNAs retrieved from the meta analysis[41] corresponding to the *Parkinson* disease in Table 7. Though the training data contains more than three folds of the false-positive associations (26 false positives vs. 8 true positives), we observe that all 12 true positives reported in[41] could be found in the top 50 of the prediction list. Among those, 5 out of 12 appear in the top 8, while 8 out of 12 show up in the top 19 highest predictions. These results support that MPM does a good job in differentiating between the true positive and true negative miRNAs.

2.6 An integrated, easy-to-use website for comprehensive analyses

We provide an easy-to-use website to query the predictions generated by our proposed model on 1,618 miRNAs and 3,679 diseases at http://software.mpm.leibniz-ai-lab.de/index.html. It is important to note that the model is trained only from the data corresponding to only a few hundreds of miRNAs and diseases. We offer a large computational prediction capability for thousands of available diseases and miRNAs through the website. All the results corresponding to the pathway and functional enrichment analysis for all miRNAs and diseases are also generated and integrated to enable comprehensive analyses by the field experts. Besides, users can also (i) search for miRNAs in the same family or related diseases (i.e., parents/children in the disease ontology) through the provided search capabilities, (ii) analyze pathways and GO processes for an input miRNA or disease, or (iii) query the list of miRNAs or diseases associated with a particular pathway. A detailed user guide with screenshots of the website is offered in Section 3 in the supplementary file.

3 Conclusion

We propose a Message Passing framework with multiple data sources integration, MPM, for the problem of predicting miRNA-disease associations. MPM exploits information from multiple data sources to enrich and filter the raw biologically relevant features without introducing additional parameters. MPM can be employed in both transductive and inductive settings. The two case studies' results support that MPM generates reasonable predictions for a disease with scarce knowledge and does a good job in differentiating between the true positive and true negative miRNAs for a disease with many false positives in the training set. Besides the proposed machine learning model, we also make the generated predictions more accessible

to non-expert users by encapsulating all the generated and related domain information into a publicly available website. By releasing such a user-friendly interface, we aim at fostering assessments and future adoption.

4 Methods

MPM treats the miRNA-disease association problem as a binary classification where the label for an input pair (m,d) is 1 if there is a known association between miRNA m and disease d and 0 otherwise. A schematic diagram of MPM with its main components is presented in Figure 1. We use gray for the model's components/modules, blue and violet for miRNA and disease-related components, respectively.

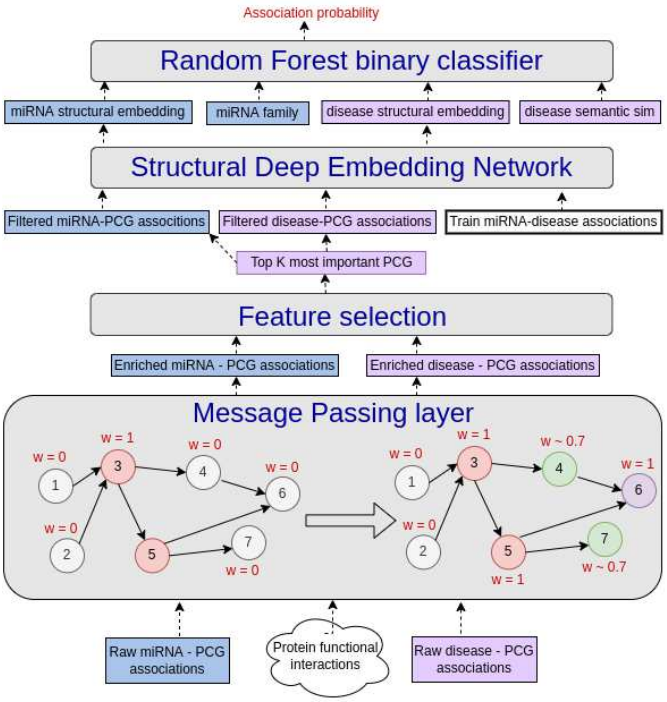

Figure 1. MPM's architecture. MPM consists of a Message Passing layer (section 4.1) , a Feature Selection with a side supervised task (section 4.2), a Structural Deep Embedding network (section 4.3), and a binary classifier (section 4.4).

4.1 The Message Passing framework/module

4.1.1 The data sources

Table 8 shows the statistics regarding our employed data sources. In the following, we describe each source in detail and present the information corresponding to how we utilize it.

The protein functional interaction network. Protein coding genes (PCGs) are essential connections between miRNAs and diseases[6]. For example, miRNAs can affect the translation of protein-coding genes resulting in reduced or increased gene

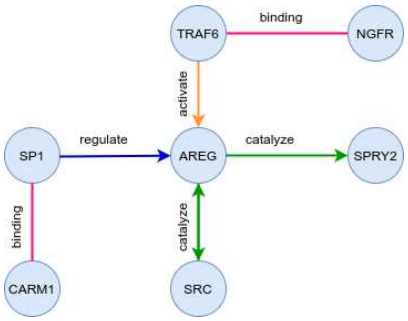

Figure 2. An example of the protein functional interaction network with relation types highlighted by different colors.

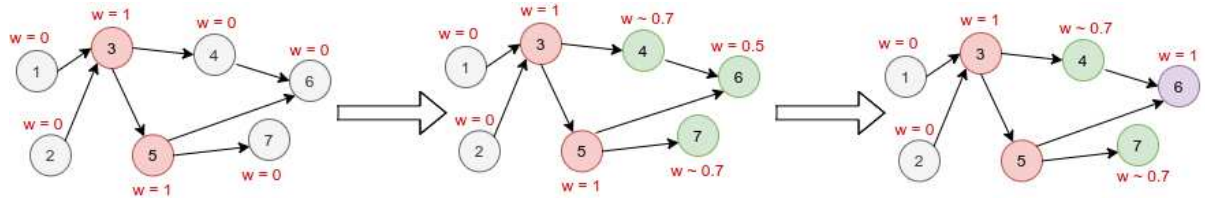

Figure 3. An example of how a message passing framework functions. The numbers inside the circles indicate nodes' IDs. 'w' indicates the node feature weight (as described in section 4.1.1). In the first iteration new weights for nodes $4, 6, 7$ are calculated according to equation (1). Only the weight for node 6 gets updated during the second iteration.

expression, which can then affect the pathology of a disease. Therefore, besides knowledge about the protein-protein interaction as already exploited in[16], knowledge about whether a particular protein regulates/inhibits/catalyzes/activates another protein is also very important for the miRNA-disease association task. We refer to the multi-relational protein-protein interaction network, where an edge corresponds to a protein functional relation as *protein functional interaction network*.

A pictorial example of the protein functional interaction network is presented in Figure 2. Different relations are depicted using different colors. Since regulation, inhibition, catalyze, and activation are one-way relations, we model the protein functional interaction network as a directed graph. We retrieve the protein functional interaction network from[42] (version 2020). We generate a directed graph from the given data as follows. Each PCG is represented as a node; a protein-protein binding interaction is modeled as two directed edges. Each relation, i.e., inhibits, activates, regulates, and catalyzes, is represented by a directed edge between the corresponding nodes. Overall our protein functional interaction network consists of 423,672 directed links between 23,611 PCGs. Some PCG nodes might be isolated in the generated network because we only include experimentally verified interactions.

Modelling miRNAs using protein functional interaction networks. We obtain the experimentally validated miRNA-PCG interactions from the miRTarBase database[35] (release 8.0). We then model each miRNA as a network of PCGs built up from the protein functional interaction network. There is a directed link between two nodes if there is a directed link between the corresponding nodes in the functional interaction network. Each PCG node in the network has a feature vector of one dimension. The feature value of a PCG node is set to 1 if there is a known interaction between it and the current miRNA, and 0 otherwise.

Modelling diseases using protein functional interaction network. We obtain the disease-PCG associations from DisGeNET[36] database, which contains one of the largest publicly available collections of genes associated with human diseases. As above, we then model each disease as a network that contains all PCGs from the protein functional interaction network. There is a directed link between two nodes if there is a directed link between the corresponding nodes in the functional interaction network. Each PCG in the network has a feature vector of one dimension. The feature value of a PCG node is set to be the *normalized confidence score* of the corresponding association between the PCG and the current disease if there exists one, and 0 otherwise.

Table 8. The side data sources statistics. $|E|$ denotes the number of interactions/associations. $|V_m|, |V_d|, |V_p|$ represent the number of miRNAs, diseases, and PCGs, respectively.

NETWORK	$\|E\|$	$\|V_m\|$	$\|V_d\|$	$\|V_p\|$
MIRNA-PCG	345,357	1,618	-	23,611
DISEASE-PCG	510,782	-	3,679	23,611
PROTEIN FUNCTIONAL INTERACTIONS	423,672			23,611

4.1.2 The message passing framework for features enrichment

The message passing module is responsible for further enriching the input representations via a simple message passing technique. It takes as input the miRNAs and diseases modeled using protein functional interaction networks with corresponding node features as described in the previous section.

MiRNA-target or disease-PCG association data might be incomplete due to lack of biological experiments or other technical limitations. Moreover, the data acquisition methods might fail to detect *indirect* PCG associations. Our message passing strategy allows us to infer such indirect or missing miRNA-PCG and disease-PCG connections. In particular, at each iteration, a message passing step is performed in which only weights of nodes with unknown interactions (i.e., nodes with initial 0 weights) with miRNAs/diseases are updated. Formally, the inferred weights for a particular node i whose original weight is 0 at iteration

t is calculated in accordance with its parents and their degrees as follows:

$$\mathbf{w}_t(i) = \frac{1}{\sqrt{\mathbf{d}_{in}(i)}} \sum_{j \in Par(i)} \frac{\mathbf{w}_{t-1}(j)}{\sqrt{\mathbf{d}_{out}(j)}} \tag{1}$$

where $Par(i)$ denotes the set of parent nodes of node i, $\mathbf{w}_{t-1}(j)$ is the weight of node j calculated at iteration $t-1$, $\mathbf{d}_{in}(i)$ and $\mathbf{d}_{out}(j)$ denote in-degree and out-degree of nodes i and j respectively. We provide an example of how the proposed message passing layer/framework works in Figure 3. The results presented in Section 2 correspond to the output from the message passing framework after one iteration as it acquires the best performance on all inductive testing datasets.

4.2 The Feature Selection module

4.2.1 The disease category

The MESH ontology[29] is a well-organized vocabulary produced by the National Library of Medicine, where diseases are classified into different categories. MESH ontology can be visualized as a tree where each layer in the tree represents one level of category. The uppermost level represents the most general category of disease. We obtain the disease category information from the MESH database. We assign a label to each disease that corresponds to its second-level category for "*Infection*" related diseases and its first-level category for the rest. We group categories which have less than ten members into one common "Others" category to make the representation space less sparse. In the end, each disease is assigned one of the 28 categories.

4.2.2 Feature selection with a side-supervised task

To remove redundant and noise in the miRNA/disease-PCG associations, we employ another source of information (the disease categories as described in Section 4.2.1) as input to our feature selection module. The rationale driving the feature selection step is that PCGs that are important for differentiating between diseases of different classes should also be indicative of the disease conditions and should, therefore, be important factors for the miRNA-disease association prediction problem.

Formally, we are given the set of diseases $\mathbf{D}$, their associated categories $\mathbf{C}$, and their inferred (up to t hops) PCGs association profiles $\mathbf{DP}_t$. We are interested in finding the top K most important PCG features predictive of the disease category.

As suggested in[43,44], ReliefF[45,46] is a competitive Feature Selection method for biological datasets. For that reason, we employ ReliefF to select the K most important PCGs. ReliefF estimates each feature's importance according to the relationship of n random samples to their nearest neighbor(s). For a given sample, the algorithm selects k nearest samples (hits) from the same class and k nearest samples (misses) from each of the other classes. The feature importance is then quantified as to how well it can differentiate between the misses and hits samples. The results presented in Section 2 correspond to $K = 100$ as it acquires the best performance on all inductive testing datasets.

4.3 The structural embedding learning

Network construction. Let $\mathbf{P}_K$ be the set of K most informative PCGs for the disease category prediction task obtained as output from the feature selection module. Let $\mathbf{A}_p$ be the adjacency matrix generated from the subset of PCG-PCG interactions for all PCGs in $\mathbf{P}_K$. Similarly, let $\mathbf{A}_{mp}$ be the adjacency matrix generated from the subset of miRNA-PCG associations for all PCGs in $\mathbf{P}_K$. $\mathbf{A}_{dp}$ be the adjacency matrix generated from the subset of disease-PCG associations for all PCGs in $\mathbf{P}_K$. Let $\mathbf{A}_{md}$ be the adjacency matrix constructed from the known miRNA-disease association. We construct an undirected network $\mathscr{G}_{mdp}$ from the training miRNA-disease associations and the filtered sets of miRNA-PCG, disease-PCG associations, and PCG-PCG interactions. The adjacency matrix for $\mathscr{G}_{mdp}$ is then given as follows:

$$\mathbf{A}_{mdp} = \begin{bmatrix} \mathbf{Z}_m & \mathbf{A}_{md} & \mathbf{A}_{mp} \\ \mathbf{A}_{md}^T & \mathbf{Z}_d & \mathbf{A}_{dp} \\ \mathbf{A}_{mp}^T & \mathbf{A}_{dp}^T & \mathbf{A}_p \end{bmatrix}$$

where $\mathbf{Z}_m \in \mathbf{R}^{n_m \times n_m}$ and $\mathbf{Z}_d \in \mathbf{R}^{n_d \times n_d}$ are the matrice of all zeros; n_m and n_d are the number of miRNAs and diseases, respectively.

Structural Deep Network embedding. The Structural Deep Network embedding[47] is a node representation learning method that can capture both the network global and local structure efficiently by employing a deep autoencoder with multi-layer architecture and multiple non-linear functions. The model is claimed to be able to learn highly non-linear network structures while being robust to the network sparsity[47]. SDNE enforces the first-order similarity constraint, which basically implies that two vertices in a network are similar if they are linked by an observed edge, as a supervised signal to learn the local network structure. The second-order proximity, which assumes that two vertices sharing many common neighbors are similar, is also incorporated into the model to capture the global network structure. A comparative study presented in[27] indicates that

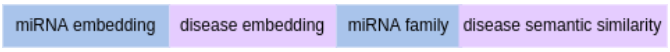

Figure 4. The final miRNA-disease input pair representation.

SDNE acquires the best performance compared with other structural embedding methods for the miRNA-disease association prediction problem. For that reason, we adapt SDNE to learn the structural embedding for miRNAs and diseases from the $\mathcal{G}_{mdp}$ network. We use the SDNE implementation shared by[27] to learn the pre-trained embedding for miRNAs and diseases from the inter-connected miRNA-PCG-disease network. The results presented in Section 2 correspond to the SDNE with 2 encoder layers of size [1000, 128], 1 decoder layer, and the output embedding of 128 dimensions as suggested in[27].

4.4 The classification module

4.4.1 The features

The miRNA family features. MiRNAs belonging to the same family usually share a common ancestor in the phylogenetic tree. They are also believed to share similar secondary structures and have similar biological functions[48]. For those reasons, miRNA family information is highly relevant to the miRNA-disease association prediction task. We retrieve miRNA family information from mirBase database[49]. In the end, each miRNA is assigned to one of the 1,375 families. We model each miRNA's family features as the one-hot encoding of its family.

The disease semantic similarity features. The disease semantic similarity[17,50] quantifies how similar two particular diseases are based on their relative positions on the disease MESH ontology[29]. We use the code and setup in[23] to compute a disease semantic similarity matrix for our 3,679 diseases set. Each entry (i,j) in the matrix indicates how similar disease i is to disease j. We model each disease's semantic similarity features as the corresponding row entry in the similarity matrix.

4.4.2 The classifier

The final classifier module takes the input representation for miRNA-disease pairs and outputs for each pair an association probability in the [0,1] range. The higher the probability, the more likely the input pair is associated. For a particular (m,d) input pair, we construct the input feature vector as the concatenation of their corresponding structural embeddings, the miRNA family and disease semantic similarity features. More specifically, $\mathbf{X}_{md} = [\mathbf{E}_m, \mathbf{E}_d, \mathbf{F}_m, \mathbf{S}_d]$, where X_{md} denotes the input feature vector corresponding to (m,d); $\mathbf{E}_m, \mathbf{E}_d$ represent the pre-trained embedding output from SDNE; while $\mathbf{F}_m$ refers to the miRNA family feature for miRNA m; $\mathbf{S}_d$ corresponds to the disease semantic similarity for disease d. A pictorial illustration of the final miRNA-disease pair representation is given in Figure 4. We train a Random Forest classifier[51,52] with 350 estimators to do the association prediction task.

References

1. Saliminejad, K., Khorshid, H. R. K., Fard, S. S. & Ghaffari, S. H. An overview of micrornas: Biology, functions, therapeutics, and analysis methods. *J. Cell. Physiol.* **234**, 5451–5465, DOI: 10.1002/jcp.27486 (2019).
2. Mattick, J. S. & Makunin, I. V. Small regulatory rnas in mammals. *Hum. molecular genetics* **14**, R121–R132 (2005).
3. Kim, V. N. & Nam, J.-W. Genomics of microrna. *TRENDS Genet.* **22**, 165–173 (2006).
4. Saini, H. K., Griffiths-Jones, S. & Enright, A. J. Genomic analysis of human microrna transcripts. *Proc. Natl. Acad. Sci.* **104**, 17719–17724 (2007).
5. Fu, G., Brkić, J., Hayder, H. & Peng, C. Micrornas in human placental development and pregnancy complications. *Int. journal molecular sciences* **14**, 5519–5544 (2013).
6. Mørk, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J. & Jensen, L. J. Protein-driven inference of mirna–disease associations. *Bioinformatics* **30**, 392–397 (2014).
7. de Ronde, M. W., Ruijter, J. M., Moerland, P. D., Creemers, E. E. & Pinto-Sietsma, S.-J. Study design and qpcr data analysis guidelines for reliable circulating mirna biomarker experiments: a review. *Clin. chemistry* **64**, 1308–1318 (2018).
8. Usuba, W. *et al.* Circulating mirna panels for specific and early detection in bladder cancer. *Cancer science* **110**, 408–419 (2019).
9. Jin, F. *et al.* Serum microrna profiles serve as novel biomarkers for autoimmune diseases. *Front. immunology* **9**, 2381 (2018).
10. Keller, A. *et al.* Toward the blood-borne mirnome of human diseases. *Nat. methods* **8**, 841–843 (2011).

11. Schickel, R., Boyerinas, B., Park, S. & Peter, M. Micrornas: key players in the immune system, differentiation, tumorigenesis and cell death. *Oncogene* **27**, 5959–5974 (2008).

12. Zhang, W., Dahlberg, J. E. & Tam, W. Micrornas in tumorigenesis: a primer. *The Am. journal pathology* **171**, 728–738 (2007).

13. Lin, Y. *et al.* Characterization of microrna expression profiles and the discovery of novel micrornas involved in cancer during human embryonic development. *PloS one* **8**, e69230 (2013).

14. Dong, Y., Sun, Y., Qin, C. & Zhu, W. Epmda: Edge perturbation based method for mirna-disease association prediction. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* (2019).

15. Zheng, K. *et al.* Dbmda: A unified embedding for sequence-based mirna similarity measure with applications to predict and validate mirna-disease associations. *Mol. Ther. Acids* **19**, 602–611 (2020).

16. Dong, T. N. & Khosla, M. Mucomid: A multitask convolutional learning framework for mirna-disease association prediction. *arXiv preprint arXiv:2108.04820* (2021).

17. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* **26**, 1644–1650 (2010).

18. Small, E. M., Frost, R. J. & Olson, E. N. Micrornas add a new dimension to cardiovascular disease. *Circulation* **121**, 1022–1032 (2010).

19. Chen, X., Liu, M.-X. & Yan, G.-Y. Rwrmda: predicting novel human microrna–disease associations. *Mol. BioSystems* **8**, 2792–2798 (2012).

20. Yang, Z. *et al.* dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic acids research* **45**, D812–D818 (2017).

21. Li, G., Luo, J., Xiao, Q., Liang, C. & Ding, P. Predicting microrna-disease associations using label propagation based on linear neighborhood similarity. *J. biomedical informatics* **82**, 169–177 (2018).

22. Chen, X., Zhang, D.-H. & You, Z.-H. A heterogeneous label propagation approach to explore the potential associations between mirna and disease. *J. translational medicine* **16**, 348 (2018).

23. Dong, T. N. & Khosla, M. Towards a consistent evaluation of mirna-disease association prediction models. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1835–1842 (IEEE, 2020).

24. Wu, Z. *et al.* A comprehensive survey on graph neural networks. *IEEE transactions on neural networks learning systems* **32**, 4–24 (2020).

25. Zeng, X., Wang, W., Deng, G., Bing, J. & Zou, Q. Prediction of potential disease-associated micrornas by using neural networks. *Mol. Ther. Acids* **16**, 566–575 (2019).

26. Ji, B.-Y. *et al.* Predicting mirna-disease association from heterogeneous information network with grarep embedding model. *Sci. Reports* **10**, 1–12 (2020).

27. Gong, Y., Niu, Y., Zhang, W. & Li, X. A network embedding-based multiple information integration method for the mirna-disease association prediction. *BMC bioinformatics* **20**, 1–13 (2019).

28. Tang, X., Luo, J., Shen, C. & Lai, Z. Multi-view multichannel attention graph convolutional network for mirna–disease association prediction. *Briefings Bioinforma.* (2021).

29. Bhattacharya, S., Ha-Thuc, V. & Srinivasan, P. Mesh: a window into full text for document summarization. *Bioinformatics* **27**, i120–i128 (2011).

30. Li, J. *et al.* Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. *Bioinformatics* (2020).

31. Pan, X. & Shen, H.-B. Scoring disease-microrna associations by integrating disease hierarchy into graph convolutional networks. *Pattern Recognit.* 107385 (2020).

32. Li, Y. *et al.* Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research* **42**, D1070–D1074 (2014).

33. Wikipedia entry for the average precision. https://en.wikipedia.org/w/index.php?title=Information_retrieval&oldid=793358396#Average_precision.

34. Scikit-learn average precision score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html.

35. Huang, H.-Y. *et al.* mirtarbase 2020: updates to the experimentally validated microrna–target interaction database. *Nucleic acids research* **48**, D148–D154 (2020).

36. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research* **48**, D845–D855 (2020).

37. Roizen, N. J. & Patterson, D. Down's syndrome. *The Lancet* **361**, 1281–1289 (2003).

38. Salvi, A. *et al.* Analysis of a nanoparticle-enriched fraction of plasma reveals mirna candidates for down syndrome pathogenesis. *Int. journal molecular medicine* **43**, 2303–2318 (2019).

39. Elton, T. S., Sansom, S. E. & Martin, M. M. Trisomy-21 gene dosage over-expression of mirnas results in the haploinsufficiency of specific target proteins. *RNA biology* **7**, 540–547 (2010).

40. Kuo, M.-C., Liu, S. C.-H., Hsu, Y.-F. & Wu, R.-M. The role of noncoding rnas in parkinson's disease: biomarkers and associations with pathogenic pathways. *J. biomedical science* **28**, 1–28 (2021).

41. Schulz, J. *et al.* Meta-analyses identify differentially expressed micrornas in parkinson's disease. *Annals neurology* **85**, 835–851 (2019).

42. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome biology* **11**, 1–23 (2010).

43. Dong, N. T. & Khosla, M. Revisiting feature selection with data complexity. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 211–216 (IEEE, 2020).

44. Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S. & Moore, J. H. Relief-based feature selection: Introduction and review. *J. biomedical informatics* **85**, 189–203 (2018).

45. Kononenko, I. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, 171–182 (1994).

46. Olson, R. S. Relieff 0.1.2. https://pypi.org/project/ReliefF/. Released: Mar 20, 2016.

47. Wang, D., Cui, P. & Zhu, W. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 1225–1234 (2016).

48. Kaczkowski, B. *et al.* Structural profiles of human mirna families from pairwise clustering. *Bioinformatics* **25**, 291–294 (2009).

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

50. Xuan, P. *et al.* Prediction of micrornas associated with human diseases based on weighted k most similar neighbors. *PloS one* **8** (2013).

51. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

52. Scikit-learn random forest classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Author contributions statement

T.N.D designed the study, collected the data, implemented the models, and analyzed the results. J.S retrieved part of the baselines' results, developed, and maintained the Web application. S.M. qualitatively validated the design and results of the case studies. M.K. designed and supervised the study as well as analyzed the results. All authors wrote the manuscript.

Funding

The work is partially supported by the PRESENt (grant no. 11-76251-99-3/19 (ZN3434)) and the "Understanding Cochlear Implant Outcome Variability using Big Data and Machine Learning Approaches" (grant no. ZN3429) projects funded by Volkswagenstiftung and the Ministry for Science and Culture of Lower Saxony, Germany (MWK: Ministerium für Wissenschaft und Kultur), the Federal Ministry of Education and Research (BMBF), Germany, under the LeibnizKILabor project (grant no. 01DD20003), and the strategic funds of the Translational Alliance in Lower Saxony.

Competing interests

The authors declare no competing interests.

Additional information

All the code and data is publicly available at https://git.l3s.uni-hannover.de/dong/mpm. All generated predictions and related domain information can be found at http://software.mpm.leibniz-ai-lab.de/.
Correspondence and requests for materials should be addressed to T.N.D or J.S.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf