# Timely-MDA: A Benchmark for Generalizable MiRNA-Disease Association Prediction

Yi Zhou*
*College of Computer Science*
*Sichuan University*
Chengdu, China
echochou99@163.com

Xian Guan*
*College of Computer Science*
*Sichuan University*
Chengdu, China
guanxian@stu.scu.edu.cn

Meixuan Wu
*College of Computer Science*
*Sichuan University*
Chengdu, China
wumeixuan@stu.scu.edu.cn

Chengzhou Ouyang
*College of Life Science*
*Sichuan University*
Chengdu, China
ouyangchengzhou@gmail.com

Min Zhu†
*College of Computer Science*
*Sichuan University*
Chengdu, China
zhumin@scu.edu.cn

***Abstract*—The identification of miRNA-disease associations (MDAs) holds significant value in the field of disease diagnosis and treatment. Recently, computational prediction methods have been increasingly proposed to detect potential MDAs, so as to assist experimental verifications. Despite the success of deep learning models, studies in the MDA prediction are still limited by the datasets and the evaluation framework employed. Concretely, existing datasets comprise only hundreds of diseases, and the random-split-based evaluation framework provides an overly optimistic estimate of the performance of MDA prediction methods. In this study, we propose a novel benchmark, Timely-MDA, for generalizable MDA prediction. First, we construct a comprehensive dataset comprising a broad scope of miRNA and disease entities and diverse semantic features. Second, four existing MDA prediction methods are implemented, and a new baseline is proposed based on our dataset. Third, the performance of these methods is analyzed using our timely-split evaluation framework. Overall, Timely-MDA provides a robust data foundation for MDA modeling, and enables quantitative estimation of the generalization ability of MDA prediction methods. Data and code are available at https://github.com/EchoChou990919/Timely-MDA.**

***Index Terms*—MDA prediction, Dataset, Evaluation**

## I. Introduction

MicroRNA (miRNA) is a class of non-coding RNA that plays an important regulatory role in the living system by influencing the output of protein-coding genes (PCGs) [1]. To date, a considerable number of miRNA-disease associations (MDAs) have been verified by biological experiments. It is witnessed that miRNAs can serve as the biomarker in diseases, and the identification of MDAs would pave the path for medical diagnosis and treatment.

In recent years, deep learning techniques, particularly graph neural networks, have greatly empowered the MDA prediction models. Nevertheless, there is another persistent question: whether the evaluation results in academics exactly demonstrate the capability of MDA prediction methods in practical applications. Unfortunately, we have observed two pitfalls in MDA prediction studies regarding the dataset and evaluation framework, which result in insufficient exploration of miRNA-disease space and over-optimistic estimation of MDA prediction methods, respectively.

**Dataset - Limited Entity Scope**. Existing datasets mainly utilize version-specific public databases such as HMDD v2.0 [2] and v3.2 [3]. During dataset construction, miRNAs and diseases present in the database were identified as entities. However, miRNAs and diseases without known associations were implicitly excluded, which prevented subsequent prediction models from further exploring unknown regions.

**Evaluation Framework - Inappropriate Data Split**. Existing evaluation frameworks usually split the MDA instances into training and test sets randomly, so as to train the models and calculate evaluation metrics. The random data split assumes MDA instances are independent and identically distributed. However, in reality, there is a sequential logic in the identification of MDAs – Prospective explorations are conducted based on the retrospective analyses. Therefore, the random-split-based evaluation framework would induces over-optimistic evaluation metrics, and fails to provide quantitative estimation of the generalization ability of prediction methods.

To address these two pitfalls, we present a new benchmark called Timely-MDA, including a richer dataset and a novel evaluation framework. On this basis, we implement four existing MDA prediction methods and newly propose a simple yet effective baseline method. Experimental results indicate that Timely-MDA brings challenges and oppotunities to the field of MDA prediction.

## II. Dataset Construction

In this section, we present the process of dataset construction, and provide the statistics of our dataset.

### A. Determining MiRNA and Disease Entities

In this study, miRNA entities are determined based on miRBase [4], an authoritatively acknowledged database that

*The first two authors contributed equally. †Corresponding author.

provides miRNA naming, sequences, and annotations. We collect 1917 terms, which are all miRNAs discovered and rigistered to date. Then, disease entities are determined based on the Medical Subject Headings (MeSH) thesaurus, a hierarchically organized vocabulary. We obtain 5032 diseases with authoritative endorsements.

### B. Processing MDA Instances

Original MDA records are downloaded from HMDD v3.2 [3]/v4.0 [5] and RNADisease [6] in January 2024. We decide to emerge these records to construct a comprehensive dataset that presents all known MDAs as far as possible.

The data quality control is implemented from three perspectives, i.e., entity alignment, evidence confirmation, and duplicate removal. First, both ends of MDA records are aligned to the determined miRNA and disease entities. Considering the names of miRNAs and diseases utilized in the literature could be unnormalized, we should ensure that they indicate certified entities unambiguously. Second, we should confirm that all the MDA instances are supported by academic evidence. The NCBI Entrez API package is employed to retrieve reference information for the evidence articles. Third, we ensure that each MDA instance is unique in the dataset while supported by at least on piece of academic evidence.

Eventually, we obtained 69602 trustworthy MDA instances. To the best of our knowledge, it's an unparalleled scale in MDA prediction studies.

### C. Collection of Semantics Features

For each miRNA, one precursor sequence and two mature sequences are extracted from miRBase. For each disease, we adopt the textual heading and scope note from MeSH, which conclude keywords and typical symptoms of the disease. Given that miRNAs perform their regulatory function by targeting the outputs of protein-coding genes (PCGs), we introduce PCGs as auxiliary entities. By referencing the HGNC [7] database, we identify the PCGs and document their names.

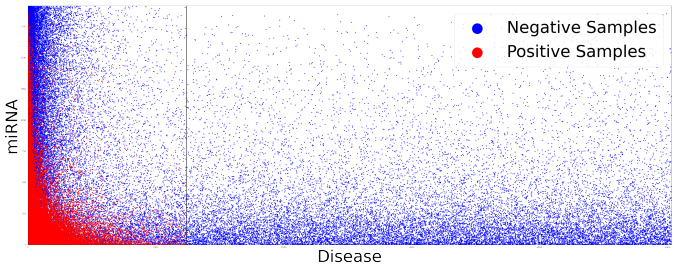
Furthermore, we acquire father-son relationships between diseases by processing the hierarchical organization of MeSH, retrieve homology kinship of miRNAs from miRBase, and get de-duplicated records of PCG-PCG interactions from HumanNet v3 [8]. Besides, miRTarBase v9.0 [9] shows records of miRNA-PCG associations, which are also aligned to entities and de-duplicated. DisGeNet 2020 [10] provides records of PCG-disease associations, which are similarly processed.

### D. Statistics

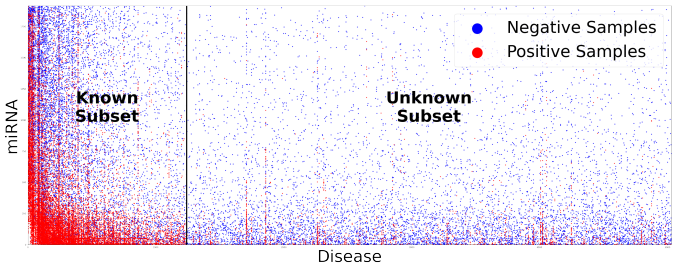
By taking entities as nodes and various relationships as edges, our dataset can be formulated as an attributed miRNA-PCG-disease graph. A conclusion is presented in Table I.

TABLE I: Statistic of the MiRNA-PCG-Disease Graph

| Node / Edge | Number | Node Attribute / Edge Semantics |
|---|---|---|
| **miRNA** | **1917** | pre- and mature sequences |
| **disease** | **5032** | text of MeSH heading and scope note |
| PCG | 19257 | text of name |
| **miRNA-disease** | **69602** | miRNA-disease association |
| miRNA-miRNA | 4513 | miRNA family membership |
| disease-disease | 7855 | disease father-Son relationship |
| PCG-PCG | 972334 | PCG interaction |
| miRNA-PCG | 144625 | miRNA-PCG interaction |
| PCG-disease | 134796 | PCG-disease association |




(a) The Training Set




(b) The Test Set

Fig. 1: Visualization of the Training and Test Sets

## III. Evalution Framework

We propose a novel evaluation framework that takes the time-based distribution shift of MDAs into account and enables the comparison of the generalization ability of MDA prediction methods.

### A. Data Split and Negative Sample Selection

The scientific process is retrospective-to-prospective, using real-life evidence to inform future research. We used this temporal pattern to model the dataset division. In this study, we split the 51321 instances first verified on and before 2020 as the training set (∼73.7%), and the 18281 instances published after 2020 as the test set (∼26.3%). For each MDA instance $(m, d)$ in the training set, we obtain a negative sample by replacing the miRNA or disease end with another unassociated one randomly. Each has a 50% chance of getting the $(m', d)$ or $(m, d')$. Notably, it's carefully constrained that negative samples in the training and test sets do not repeat. In addition, we aim to evaluate whether the MDA prediction can be effectively generalized to novel diseases. Therefore, the test set is split into "known" and "unknown" subsets: the former involves 16563 MDA instances and 10277 negative samples, and the latter contains 1725 positive samples and 8005 negative samples.

### B. Visualization Analysis

Fig.1 visualizes our MDA samples in scatter plots, where positive samples are represented as red dots and negative samples are represented as blue dots. On the x-axis, 5032

diseases are arranged from left to right according to the "degree of known" from high to low, which indicates how many times has the disease present as the tail of positive samples in the training set. A similar observation can be made with regard to the 1917 miRNAs on the y-axis.

By comparing the occupied area of the red dots with the full miRNA-disease space, we can intuitively see that only a small fraction of MDAs ( 0.72%) have been verified, and there are still broad regions awaiting exploration. By comparing the Fig.1a and Fig.1b, we find that the verified MDAs are spread to more disease entities and further blank areas as time goes on (before and after 2020). These observations meet the essential goal of MDA prediction: we should identify potential MDAs for forward-looking scientific research.

### C. Evaluation Metrics

There are six widely used evaluation metrics in binary classification problems, i.e., Accuracy, Precision, Recall, F1-score, AUC, and AUPR. To quantatively assess the overall performance and the generalization ability of MDA prediction methods, we tend to calculate all six metrics on the full test set first, and then prioritize analysis of the AUC and AUPR values from the unknown subset.

## IV. Experiments

In computational experiments, we primarily investigate the following research questions:

- **RQ1**: How well do existing MDA prediction methods perform on our benchmark? Whether they lack generalization ability?
- **RQ2**: Can we propose a new baseline method that achieves superior MDA prediction performance by making better use of our dataset?
- **RQ3**: Compared to the traditional random-split-based evaluation framework, how is the rationality of our timely-split-based evaluation framework?

### A. Performance of Existing Baseline Methods

We implemented four state-of-the-art baseline methods on the Timely-MDA benchmark, namely NIMCGCN [11], DFELMDA [12], AGAEMD [13], MINIMDA [14]. During the reproduction process, we recalculate all relevant similarity features based on our dataset, and the problem of information leakage is carefully prevented.

In Table II, the first four rows present the evaluation results of these baseline methods. It's obvious that MINIMDA outperforms other methods on most metrics. MINIMDA achieves the highest AUC, AUPR, and Precision in terms of the full test set, presents the best AUC on the known subset, and exhibits optimal AUC and AUPR on the unknown subset. Notably, there exists a significant performance gap between known and unknown subsets for all methods. As exemplified by MINIMDA, the AUC is 0.818 on the known subset while only 0.679 on the unknown subset. Since AUC is determined by the true positive rate and the false positive rate with all possible thresholds, a higher AUC is achieved when a greater

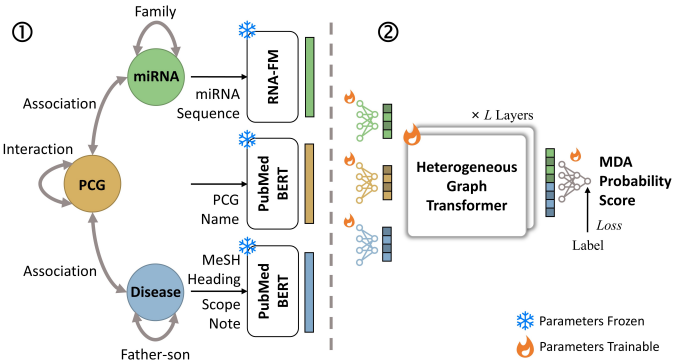



Fig. 2: Architectural Components of PLM-HGNN.

proportion of positive samples are accurately predicted with top-ranked MDA probability scores. Therefore, it is evident that MINIMDA encounters greater challenges when predicting MDAs for unknown diseases. The same holds true for other methods, accurate predictions are biased toward the well-known diseases.

So, here is the answer to **RQ1**: By analyzing the benchmark results, it is witnessed that existing MDA prediction methods are limited in the generalization ability.

### B. PLM-HGNN: A New Baseline Method

Existing MDA prediction methods generally rely on pre-extracted similarity features, which limits their generalization ability. To break through this bottleneck, we propose a simple yet effective method, PLM-HGNN, which can make full use of our dataset.

Fig.2 is an overview of PLM-HGNN. Firstly, Pre-trained Langauge Models (i.e., RNA-FM [14] and PubMed-BERT [15]) can generate informative vector representations for heterogeneous node attributes. Secondly, an MDA prediction model, based on Heterogeneous Graph Neural Networks [16], learns the topology and attribute information in the miRNA-PCG disease graph. Thirdly, the final miRNA-disease embeddings are projected into MDA probability scores.

Table.II shows PLM-HGNN's excellent performance on Timely-MDA, especially in terms of the generalization ability. PLM-HGNN not only advances the AUC, accuracy, recall and F1-score on the full test set, but also achieves the highest AUC of 0.741 and the top AUPR of 0.354 on the unknown subset, which exhibit a 9.13% and 21.65% improvement over MINIMDA. Meanwhile, we highlight that there is scope for enhancement in the performance of PLM-HGNN.

In conclusion, we can answer **RQ2**: Our dataset can be mined for advanced MDA prediction, and PLM-HGNN is freshly proposed as an effective baseline method.

### C. Performance Gap between Traditional and Novel Evaluation Frameworks

As previously stated, we suspect that the random-split-based evaluation framework can result in an over-optimistic estimation of MDA prediction. To ascertain whether this is true, we have conducted experiments under the traditional

TABLE II: Performance of Baseline Methods

| Method | Test Set | | | | | | Known Subset | | Unknown Subset | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | AUPR | Accuracy | Precision | Recall | F1-score | AUC | AUPR | AUC | AUPR |
| NIMCGCN | 0.773 | 0.732 | 0.650 | 0.776 | 0.421 | 0.546 | 0.776 | 0.817 | 0.614 | 0.229 |
| DFELMDA | 0.827 | 0.831 | 0.737 | 0.855 | 0.570 | 0.684 | 0.804 | 0.889 | 0.556 | 0.22 |
| AGAEMD | 0.818 | 0.829 | 0.744 | 0.837 | 0.606 | 0.703 | 0.800 | 0.862 | 0.54 | 0.183 |
| MINIMDA | 0.841 | 0.840 | 0.715 | 0.880 | 0.497 | 0.635 | 0.818 | 0.867 | 0.679 | 0.291 |
| PLM-HGNN | **0.845** $\Delta$ | 0.833 | **0.749** $\Delta$ | 0.831 | **0.627** $\Delta$ | **0.714** $\Delta$ | 0.811 | 0.858 | **0.741** $\Delta$ | **0.354** $\Delta$ |

Note: Best in existing baselines; **Improved by the new baseline** $\Delta$

TABLE III: Performance of Baseline Methods Under Random-Split-Based Evaluation Framework

| Method | AUC | AUPR | ACC | P | R | F1 |
|---|---|---|---|---|---|---|
| NIMCGCN | 0.855 | 0.843 | 0.76 | 0.708 | 0.893 | 0.79 |
| DFELMDA | 0.933 | 0.924 | 0.862 | 0.863 | 0.865 | 0.864 |
| AGAEMD | 0.937 | 0.938 | 0.864 | 0.857 | 0.879 | 0.867 |
| MINIMDA | 0.916 | 0.917 | 0.845 | 0.846 | 0.847 | 0.847 |
| PLM-HGNN | 0.917 | 0.915 | 0.839 | 0.824 | 0.867 | 0.845 |

random data split as well. Comparing to our timely-split-based dataset, a random-split-based dataset constructed here maintains an equal training/test proportion, and employs the same engative sample selection strategy. The four baseline methods and PLM-HGNN are trained and evaluated on this random-split-based dataset, and the six evaluation metrics are calculated as well.

By comparing Table III and Table II, we find that each method demonstrates notable disparities in prediction performance across the traditional and novel evaluation framework. Numerically speaking, the traditional random data split paints a more optimistic picture of all MDA prediction methods, while our timely-split-based evaluation framework makes the prediction more challenging. Focusing on the relative assessment between baseline methods, it is acknowledged that the evaluation results derived from the two frameworks are synchronized to a certain extent. DFELMDA and AGAEMD perform similarly and seem skilled in fitting the distribution of training data, while MINIMDA and PLM-HGNN are competitive with each other and are considered with better generalization ability.

In conclusion, we can answer **RQ3**: Our timely-split-based evaluation framework exhibits sufficient rationality. It provides consistent relative estimations between MDA prediction methods yet presents a less optimistic situation.

## V. Conclusion and Discussion

In this study, we propose Timely-MDA—a benchmark for generalizable MDA prediction. First, we construct a dataset that encompasses a broad scope of authorized entities, multi-source integrated MDA instances, and rich semantics features. Second, we introduce a timely-split-based evaluation framework that simulates the retrospect to-prospect scene of scientific exploration. In experiments, it is demonstrated that Timely-MDA provides challenges and oppotunities in advancing the generalization ability of MDA prediction methods.


## Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos.62172289).


## References


[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," cell, vol. 116, no. 2, pp. 281–297, 2004.

[2] Y. Li et al., "HMDD v2. 0: a database for experimentally supported human microRNA and disease associations," Nucleic acids research, vol. 42, no. D1, pp. D1070–D1074, 2014.

[3] Z. Huang et al., "HMDD v3. 0: a database for experimentally supported human microRNA–disease associations," Nucleic acids research, vol. 47, no. D1, pp. D1013–D1017, 2019.

[4] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," Nucleic acids research, vol. 47, no. D1, pp. D155–D162, 2019.

[5] C. Cui, B. Zhong, R. Fan, and Q. Cui, "HMDD v4. 0: a database for experimentally supported human microRNA-disease associations," Nucleic Acids Research, vol. 52, no. D1, pp. D1327–D1332, 2024.

[6] J. Chen et al., "RNADisease v4. 0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction," Nucleic Acids Research, vol. 51, no. D1, pp. D1397–D1404, 2023.

[7] R. L. Seal et al., "Genenames. org: the HGNC resources in 2023," Nucleic Acids Research, vol. 51, no. D1, pp. D1003–D1009, 2023.

[8] C. Y. Kim et al., "HumanNet v3: an improved database of human gene networks for disease research," Nucleic acids research, vol. 50, no. D1, pp. D632–D639, 2022.

[9] H.-Y. Huang et al., "miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions," Nucleic acids research, vol. 50, no. D1, pp. D222–D230, 2022.

[10] J. Piñero et al., "The DisGeNET knowledge platform for disease genomics: 2019 update," Nucleic acids research, vol. 48, no. D1, pp. D845–D855, 2020.

[11] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction," Bioinformatics, vol. 36, no. 8, pp. 2538–2546, 2020.

[12] W. Liu et al., "Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder," Briefings in Bioinformatics, vol. 23, no. 3, p. bbac104, 2022.

[13] H. Zhang, J. Fang, Y. Sun, G. Xie, Z. Lin, and G. Gu, "Predicting miRNA-disease associations via node-level attention graph auto-encoder," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 2, pp. 1308–1318, 2022.

[14] Z. Lou, Z. Cheng, H. Li, Z. Teng, Y. Liu, and Z. Tian, "Predicting miRNA–disease associations via learning multimodal networks and fusing mixed neighborhood information," Briefings in Bioinformatics, vol. 23, no. 5, p. bbac159, 2022.

[15] J. Chen et al., "Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions," arXiv preprint arXiv:2204.00300, 2022.

[16] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare (HEALTH), vol. 3, no. 1, pp. 1–23, 2021.

[17] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in Proceedings of the web conference 2020, 2020, pp. 2704–2710.