Interval-Shared Information Integration and False-Negative Association Reduction in Multi-Source MiRNA-Disease Association Prediction

Qinghang Cui, Honglie Guo, Yueyi Cai, Yu Fei, and Shunfang Wang, *Member, IEEE*

S1 Multi-Source Similarity Construction

We utilize various biological data sources to comprehensively characterize the similarity view of miRNAs and diseases.

MiRNA sequence similarity. We compute the miRNA sequence similarity using the global matching Needleman-Wunsch algorithm [1]:

$$MSS(m_i, m_j) = \begin{cases} 1 & m_i = m_j \\ \frac{MS(m_i,m_j)-S_{min}}{S_{max}-S_{min}} & m_i \neq m_j \end{cases} \tag{1}$$

where S_{max} and S_{min} represent the maximum and minimum scores in the miRNA sequence similarity matrix, and $MS(m_i, m_j)$ denotes the normalization.

MiRNA functional similarity. We extract a gene-function interaction network from HumanNet [2] to compute the miRNA functional similarity (MFS) using the following equations:

$$S(g_i, g_j) = \begin{cases} 1 & g_i = g_j \\ 0 & e(g_i, g_j) \notin L_{HumanNet} \\ LLS(g_i, g_j) & e(g_i, g_j) \in L_{HumanNet} \end{cases} \tag{2}$$

where $e(g_i, g_j)$ represents the link between g_i and g_j, and $LLS(g_i, g_j)$ is min-max normalization. Subsequently, the similarity between gene g_a and gene set $G = \{g_{a1}, g_{a2}, \cdots, g_{ak}\}$ is computed as follows:

$$S(g_a, G) = \max_{1 \leq i \leq k} (S(g_a, g_{ai})) \tag{3}$$

Finally, we obtain the MFS between m_i and m_j:

$$MFS(m_i, m_j) = \frac{\sum_{g \in G_i} S(g, G_j) + \sum_{g \in G_j} S(g, G_i)}{|G_i| + |G_j|} \tag{4}$$

MiRNA Gaussian interaction profile kernel similarity. To augment the richness of similarity information, we further computed the Gaussian interaction profile (GIP) kernel similarity:

$$\gamma_m = \gamma'_m / (\frac{1}{N_m} \sum_{i=1}^{N_m} \|IP(m_i)\|^2) \tag{5}$$

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

where γ_m represents the parameter governing the kernel bandwidth, and $IP(m_i)$ corresponds to the i-th row in the adjacency matrix A.

Disease semantic similarity. We employ a directed acyclic graph to delineate the hierarchical relationships among diseases from MeSH, which is symbolized as $DAG(d_i) = (d_i, V(d_i), E(d_i))$, where $V(d_i)$ and $E(d_i)$ signify the set of nodes and edges of node d_i. The semantic contribution of disease d_j to d_i can be expressed as follows:

$$D_{d_i}(d_j) = \begin{cases} 1 & d_i = d_j \\ \max\{\Delta \times D_{d_i}(d'_j) \,|\, d'_j \in \text{children of } d_j\} & d_i \neq d_j \end{cases} \tag{7}$$

where Δ signifies the semantic contribution factor. The semantic value $DS(d_i)$ of a disease d_i is defined as the summation of the contributions from all its ancestor nodes:

$$DS(d_i) = \sum_{d_j \in T(d_i)} D_{d_i}(d_j) \tag{8}$$

Disease semantic similarity is defined as follows:

$$DSS(d_i, d_j) = \frac{\sum_{d_t \in T(d_i) \cap T(d_j)} (D_{d_i}(d_t) + D_{d_j}(d_t))}{DS(d_i) + DS(d_j)} \tag{9}$$

Disease functional similarity. Similarly to the computation of miRNA functional similarity, we computed disease functional similarity utilizing disease-gene associations sourced from DisGeNET [3]:

$$DFS(d_i, d_j) = \frac{\sum_{g \in G_i} S(g, G_j) + \sum_{g \in G_j} S(g, G_i)}{|G_i| + |G_j|} \tag{10}$$

Disease Gaussian interaction profile kernel similarity. Similarly, the GIP kernel similarity between disease d_i and d_j is defined as:

$$DGS(d_i, d_j) = \exp(-\gamma_d \|IP(d_i) - IP(d_j)\|^2) \tag{11}$$

$$\gamma_d = \gamma'_d / (\frac{1}{N_d} \sum_{i=1}^{N_d} \|IP(d_i)\|^2) \tag{12}$$

S2 Data Preparation

We obtained miRNA-disease association data from the Human MicroRNA Disease Database (HMDD v3.2) [4]. To ensure consistency, we utilized uniform identifiers for both miRNAs and diseases. MiRNA sequence information was sourced from miRBase [5]. Disease semantic trees from the Medical Subject Headings (MeSH) were used. After merging redundant miRNA transcripts and matching them with MeSH disease descriptors, we identified 12,446 experimentally confirmed MDAs involving 853 miRNAs and 591 diseases. These associations were represented in a binary adjacency matrix $A \in \mathbb{R}^{853\times591}$, where a value of 1 indicates an association between miRNA m_i and disease d_j, and 0 usually indicates no association. In this study, we noted that the value of 0 could indicate a false-negative association, so we focused on addressing this issue.

S3 The Bias Term γ

The bias term γ in Eq (13) is introduced to appropriately weigh the "must-link" constraint loss against the "cannot-link" constraint loss. When $\gamma = 1$, the constraint loss function is optimized using only "must-link" constraints, while when $\gamma = 0$, it is optimized using only "cannot-link" constraints. Table I shows the effect of different values of γ on the prediction performance of ISFNMDA. When we properly balance the "must-link" and "cannot-link" constraints, the performance is better than when only "must-link" or "cannot-link" constraints are used.

$$\mathcal{L}_{mc} = \sum_{n=1}^{k}(\gamma \sum_{SMLM} d^{(n)}(m_i, m_j) - (1-\gamma) \sum_{SCLM} d^{(n)}(m_i, m_j)) \tag{13}$$

S4 The Number of Neighbors K

The choice of the number of neighbors K significantly influences the quality of the initialization and sparsification of the inference view, as shown in Table II. Through experimentation with various values of K, we determined that $K = 30$ is the optimal parameter for our model. As shown in the table, the selection of K aims to strike a balance: choosing a value that is neither too small nor too large. A small K value restricts the number of useful neighbors, while a large K value may introduce noisy connections.

S5 Performance of Different Similarity Fusion

To evaluate the impact of interval-shared information extraction on model performance across different combinations of multi-view similarity attributes, we conducted various similarity fusion experiments. We hypothesized that similarities computed using the same method for both miRNA and disease belong to a single group, and constructed four combinations: ISFNMDA (which combines all similarity attributes), Fusion1 (including miRNA sequence similarity, miRNA functional similarity, disease semantic similarity, and disease functional similarity), Fusion2 (including miRNA sequence similarity, miRNA GIP kernel similarity, disease semantic similarity, and disease GIP kernel similarity), and Fusion3 (including miRNA functional similarity, miRNA GIP kernel similarity, disease functional similarity, and disease GIP kernel similarity). The detailed results are shown in Fig. 1. When only two similarity attributes are used for miRNA and disease, the model performs poorly; however, when all similarity attributes are used, the model performs best. We believe that incorporating multiple attribute types helps the model better capture interactions between attributes, enabling more effective extraction of interval-shared patterns.

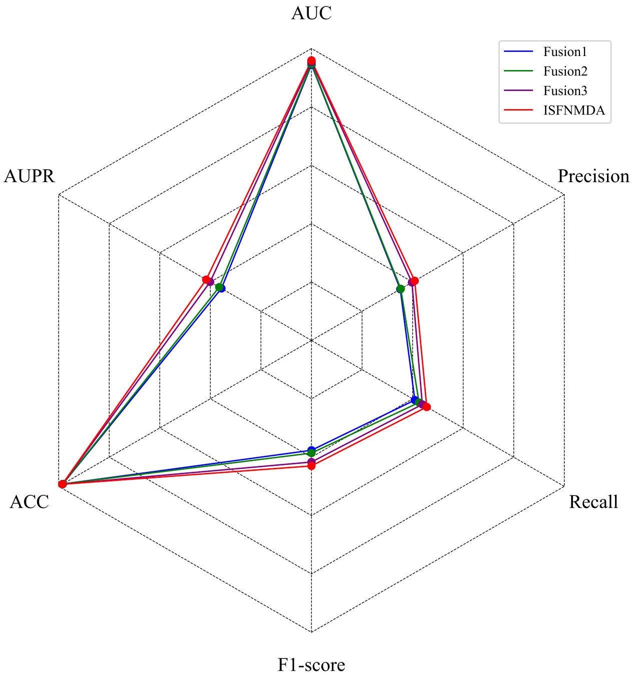


Fig. 1. Performance of different similarity fusion.

S6 Evaluation Metric

AUC: The AUC is a widely accepted metric for evaluating classification performance, particularly for imbalanced datasets. It reflects the model's ability to discriminate between positive and negative classes, providing a global measure of performance that is not affected by class imbalance. Since our model involves predicting rare miRNA-disease associations, AUC is a key indicator of how well the model ranks positive associations against negative ones.

AUPR: The AUPR is especially relevant in scenarios with imbalanced classes, where the positive class is much smaller than the negative class. AUPR focuses on the precision and recall for the positive class, making it a more informative metric when the dataset has a high class imbalance. Since our goal is to predict rare disease associations, AUPR is crucial to assess the model's performance specifically on identifying true positive associations.

Accuracy, F1-Score, Recall, and Precision: These metrics are also commonly used in classification tasks. While accuracy measures the overall correctness of the model, F1-Score provides a balanced assessment between precision and recall,

TABLE I
EFFECT OF DIFFERENT BIAS TERMS γ.

ISFNMDA@γ	γ=0	γ=0.1	γ=0.2	γ=0.3	γ=0.4	γ=0.5	γ=0.6	γ=0.7	γ=0.8	γ=0.9	γ=1
AUC	0.9469	0.9528	0.9539	0.9554	0.9566	**0.9589**	0.9581	0.9574	0.9562	0.9556	0.9507

TABLE II
EFFECT OF DIFFERENT NUMBERS OF NEIGHBORS $\boldsymbol{K}$.

ISFNMDA@K	K=5	K=10	K=15	K=20	K=25	K=30	K=35	K=40	K=45	K=50
AUC	0.9469	0.9528	0.9539	0.9554	0.9566	**0.9589**	0.9581	0.9574	0.9562	0.9556

which is critical when considering both false positives and false negatives. Recall measures the model's ability to identify all true positive associations, while precision evaluates how many of the predicted positive associations are actually correct. These metrics help provide a well-rounded view of the model's performance in both identifying and verifying miRNA-disease associations.

S7 THE EFFECT OF THE PAIRWISE SIMILARITY METHODS

We conducted comprehensive experiments using four metrics: Cosine Similarity, Euclidean Distance, Spearman's Correlation, and Pearson Correlation Coefficient (PCC). As shown in the experimental results 2, the performance differences among these metrics (Cosine Similarity, Euclidean Distance, Spearman's Correlation, and PCC) proved marginal. Nevertheless, we ultimately selected PCC based on the following comprehensive considerations:

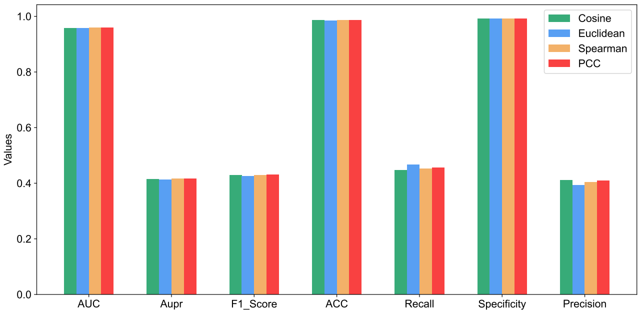


Fig. 2. Performance comparison of different similarity calculation methods.

While all metrics achieved comparable prediction accuracy with minimal average performance variations, PCC demonstrated notably robust performance throughout the five-fold cross-validation cycles. Furthermore, aligning with the fundamental principle that "functionally similar miRNAs are associated with phenotypically similar diseases," PCC exhibits unique biological consistency with miRNA co-regulation dynamics. Specifically, miRNA co-regulation typically manifests linear proportionality at expression levels - a critical characteristic that PCC effectively quantifies through covariance normalization. This essential feature remains unaddressed by rank-based (Spearman) or magnitude-insensitive (Cosine) metrics. Such biological interpretability ensures that the extracted features authentically reflect genuine regulatory relationships.

Based on these experimental observations and bioinformatic considerations, we ultimately selected PCC for computing pairwise node similarities. We fully concur that more advantageous similarity computation methods may emerge in future research, and this will undoubtedly constitute a crucial direction for our subsequent investigations.

S8 THE EFFECT OF THE MOMENTUM CONTRASTIVE LEARNING

To evaluate the performance differences in function encoding accuracy between momentum contrastive learning and general graph contrastive learning methods, we reviewed relevant literature and identified GraphCL [6] — a model with a framework similar to our GraphMoCo module—presented in the NeurIPS 2020 paper "Graph Contrastive Learning with Augmentations".

We conducted a comparative analysis between GraphMoCo and GraphCL, ensuring that GraphCL utilized the same encoder as GraphMoCo for a fair comparison. For GraphCL, we set both the Refinement view and Inference view to share a single encoder and adopted the same loss function as GraphMoCo as the optimization objective. As shown in Fig. 3, the results indicate that GraphMoCo significantly outperforms GraphCL. Specifically, GraphMoCo achieved an AUC of 0.9589, compared to GraphCL's AUC of 0.9484—a 1.05 percentage point improvement. These findings demonstrate that GraphMoCo offers a clear advantage in improving function encoding accuracy and highlight its superiority over traditional general graph contrastive learning approaches.

Although both GraphMoCo and GraphCL maximize mutual information between views as their optimization objective, there is a key difference in their training strategies. GraphMoCo introduces the Exponential Moving Average (EMA) mechanism to update the parameters of the Refinement network, which ensures smoothness and robustness in the training process. In contrast, GraphCL employs a standard contrastive learning update without the additional smoothing mechanism.

The EMA update strategy in GraphMoCo provides more stable guidance signals to the Inference network, alleviating sudden fluctuations in the learned representations. This mechanism not only improves the stability of the training process but also enhances the consistency of the learned representations, allowing the model to better capture the underlying

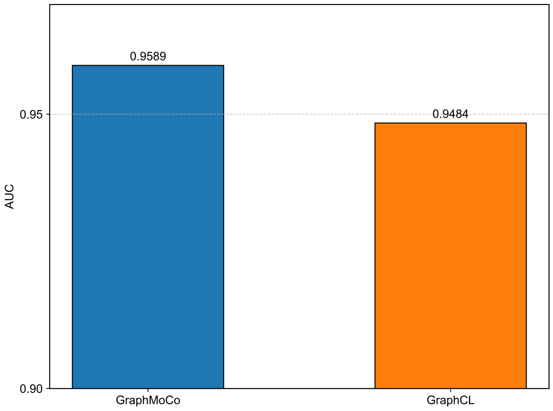


Fig. 3. Performance Comparison between GraphMoCo and GraphCL.

structural relationships in the data. Experimental results further validate this advantage, showing that GraphMoCo outperforms GraphCL on multiple evaluation metrics, proving the effectiveness of the EMA mechanism in graph contrastive learning tasks.

S9 The effect of the StrandCNN encoder

We selected four GNN models, including the Graph Convolutional Network (GCN), GraphSAGE, Graph Isomorphism Network (GIN), and Graph Attention Network (GAT). During the experiments, we replaced the GCN module in the manuscript with GraphSAGE, GIN, and GAT modules to assess their impact on predictive accuracy and robustness.

The ROC curves for the four GNN models are shown in Fig. 4, and the detailed metrics are provided in Table III. Experimental results indicate that the GCN achieved the highest AUC (0.9589) and accuracy (0.9856). These metrics are critical because a high AUC demonstrates that the GCN can effectively distinguish between classes, while high accuracy reflects its overall predictive reliability. Although the Graph Attention Network (GAT) scored highest in AUPR, F1-Score, and precision, and GraphSAGE achieved the highest recall, the GCN maintained strong performance across all metrics. This balanced performance indicates that the GCN is less sensitive to variations in dataset characteristics, such as class imbalance.

[1] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

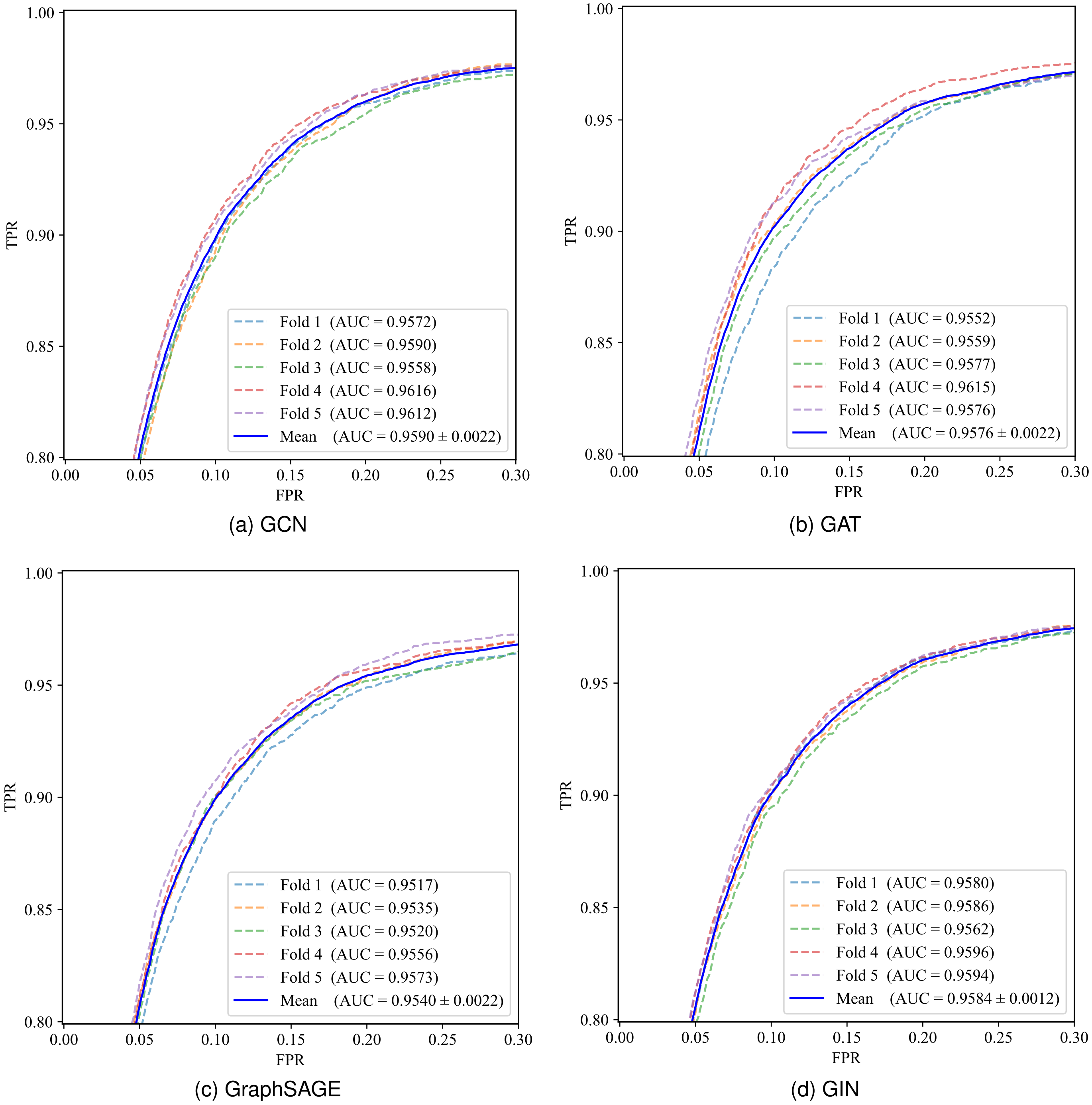
[2] S. Hwang, C. Y. Kim, S. Yang, E. Kim, T. Hart, E. M. Marcotte, and I. Lee, "Humannet v2: human gene networks for disease research," *Nucleic acids research*, vol. 47, no. D1, pp. D573–D580, 2019.

[3] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic acids research*, vol. 48, no. D1, pp. D845–D855, 2020.

[4] 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*, vol. 47, no. D1, pp. D1013–D1017, 2019.

[5] 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.

[6] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020.



(a) GCN

(b) GAT

(c) GraphSAGE

(d) GIN

Fig. 4. ROC of different GNNs.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT GNNS.

GNNS	AUC	AUPR	ACC	F1-Score	Recall	Precision
GCN	**0.9589**	0.4164	**0.9856**	0.4300	0.4561	0.4084
GAT	0.9576	**0.4320**	0.9847	**0.4468**	0.4493	**0.4457**
GraphSAGE	0.9540	0.4093	0.9853	0.4344	**0.4714**	0.4037
GIN	0.9583	0.4143	0.9855	0.4324	0.4622	0.4065

Bold values represent the best result; Underlined values denote the second-best results.